THE UNIVERSITY OF CHICAGO

DATA-DRIVEN APPROACHES FOR MOLECULAR DESIGN AND SIMULATION: FROM SELF-ASSEMBLING PEPTIDES TO ENHANCED SAMPLING TECHNIQUES AND ATOMISTIC STRUCTURE GENERATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE PRITZKER SCHOOL OF MOLECULAR ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
KIRILL SHMILOVICH

CHICAGO, ILLINOIS
AUGUST 2023

Copyright © 2023 by Kirill Shmilovich

All Rights Reserved

To my parents and gradparents

TABLE OF CONTENTS

LIST OF FIGURES viii

LIST OF TABLES xxiv

ACKNOWLEDGMENTS xxv

ABSTRACT xxvii

1 INTRODUCTION 1
1.1 Self-assembling π-conjugated peptides 2
1.2 Active learning data-driven design 6
1.3 Enhanced sampling in molecular simulations 9
1.4 Backmapping for restoring atomistic detail into coarse-grained molecules . . 11
1.5 Publications 14

2 DISCOVERY OF SELF-ASSEMBLING π-CONJUGATED PEPTIDES BY ACTIVE LEARNING-DIRECTED COARSE-GRAINED MOLECULAR SIMULATION . . 16
2.1 Abstract 16
2.2 Introduction 17
2.3 Methods 21
2.3.1 Molecular dynamics simulation 21
2.3.2 Active learning peptide discovery 23
2.3.3 Nonlinear manifold learning of assembly pathways 40
2.4 Results and Discussion 42
2.4.1 Active learning identification of optimal candidates 42
2.4.2 Manifold learning of assembly pathways 47
2.4.3 Unsupervised spectral clustering into assembly classes 49
2.5 Conclusions 52

3 HYBRID COMPUTATIONAL-EXPERIMENTAL DATA-DRIVEN DESIGN OF SELF-ASSEMBLING π-CONJUGATED PEPTIDES 56
3.1 Abstract 56
3.2 Introduction 57
3.3 Methods 61
3.3.1 Overview 61
3.3.2 All-atom molecular dynamics simulations 63
3.3.3 Chemical space embedding 64
3.3.4 Computational active learning 66
3.3.5 Experimental active learning 67
3.3.6 Nonlinear manifold learning of low-dimensional assembly pathways . 69
3.3.7 Oligopeptide synthesis 70
3.3.8 UV-visible spectroscopy 73

3.4 Results and Discussion . . . 73
3.4.1 Hybrid computational/experimental active learning discovers novel high-performing oligopeptides . . . 73
3.4.2 Molecular design rules . . . 77
3.4.3 Molecular assembly pathways . . . 81
3.5 Conclusions . . . 85

4 DATA-DRIVEN DISCOVERY OF CARDIOLIPIN-SELECTIVE SMALL MOLECULES BY COMPUTATIONAL ACTIVE LEARNING . . . 89
4.1 Abstract . . . 89
4.2 Introduction . . . 90
4.3 Computational Methods . . . 93
4.3.1 Overview . . . 93
4.3.2 Coarse-grained molecular design space . . . 94
4.3.3 Coarse-grained molecular dynamics (CGMD) simulations . . . 96
4.3.4 Alchemical free energy calculations . . . 97
4.3.5 Chemical space embedding . . . 99
4.3.6 Active learning . . . 100
4.3.7 Inference of design rules . . . 102
4.3.8 Functional group analysis . . . 103
4.4 Experimental Methods . . . 104
4.4.1 Fluorescence anisotropy . . . 104
4.5 Results and Discussion . . . 105
4.5.1 Active learning identifies highly cardiolipin-selective compounds . . . 105
4.5.2 Data-driven discovery of chemical design rules . . . 108
4.5.3 All-atom backmapping . . . 111
4.5.4 Validating simulations against experimental measurements . . . 115
4.6 Conclusions . . . 119

5 COMPUTATIONAL DISCOVERY OF HIGH CHARGE MOBILITY SELF-ASSEMBLING π-CONJUGATED PEPTIDES . . . 122
5.1 Abstract . . . 122
5.2 Introduction . . . 123
5.3 Methods . . . 127
5.3.1 Overview of the computational pipeline . . . 127
5.3.2 MD calculations . . . 129
5.3.3 DFT calculations . . . 131
5.3.4 Chemical Stability . . . 132
5.3.5 Charge Mobility . . . 133
5.4 Results and Discussion . . . 137
5.4.1 Chemical Stability . . . 137
5.4.2 Reorganization Energy . . . 139
5.4.3 Transfer Integral . . . 140
5.4.4 Charge Mobility . . . 143

5.5 Conclusions . . . 145

6 TEMPORALLY COHERENT BACKMAPPING OF MOLECULAR TRAJECTORIES FROM COARSE-GRAINED TO ATOMISTIC RESOLUTION . . . 147
6.1 Abstract . . . 147
6.2 Introduction . . . 148
6.3 Methods . . . 151
6.3.1 Representing molecules as spatially voxelized particle densities . . . 151
6.3.2 Conditional Variational Autoencoder . . . 154
6.3.3 Training routines . . . 157
6.3.4 Inference . . . 159
6.3.5 Data curation . . . 161
6.4 Results . . . 164
6.4.1 Alanine dipeptide . . . 165
6.4.2 Chignolin . . . 173
6.5 Discussion & Conclusions . . . 179

7 GIRSANOV REWEIGHTING ENHANCED SAMPLING TECHNIQUE (GREST): ON-THE-FLY DATA-DRIVEN DISCOVERY OF AND ENHANCED SAMPLING IN SLOW COLLECTIVE VARIABLES . . . 184
7.1 Abstract . . . 184
7.2 Introduction . . . 185
7.3 Methods . . . 191
7.3.1 Transfer operator theory . . . 191
7.3.2 Approximating slow modes from data . . . 193
7.3.3 State-free Reversible VAMPnets (SRVs) . . . 196
7.3.4 Dynamical reweighting . . . 199
7.3.5 GREST adaptive sampling workflow . . . 205
7.3.6 Simulation settings and parameters . . . 209
7.4 Results . . . 210
7.4.1 One-dimensional four-well potential . . . 210
7.4.2 Alanine dipeptide . . . 215
7.4.3 WLALL pentapeptide . . . 222
7.5 Discussion & Conclusions . . . 228

8 DIAMONDBACK: DIFFUSION-DENOISING AUTOREGRESSIVE MODEL FOR NON-DETERMINISTIC BACKMAPPING OF Cα PROTEIN TRACES . . . 232
8.1 Abstract . . . 232
8.2 Introduction . . . 233
8.3 Methods . . . 237
8.3.1 Denoising diffusion probabilistic model (DDPM) formalism . . . 238
8.3.2 U-net neural network architecture . . . 241
8.3.3 Autoregressive backmapping of Cα traces . . . 242
8.3.4 Terminal residues . . . 247

8.3.5 Data Curation . . . 247
8.3.6 Evaluation metrics . . . 251
8.4 Results and Discussion . . . 253
8.4.1 PED training and evaluation . . . 254
8.4.2 PDB training with MD evaluation . . . 255
8.4.3 Analysis of side chain dihedral angles in generated atomistic structures 260
8.4.4 Fine-tuning impact on MD trajectories . . . 262
8.4.5 Analysis of residue-wise performance . . . 264
8.4.6 Application to bespoke coarse-grained trajectories . . . 265
8.5 Conclusions . . . 268

9 CONCLUSIONS AND FUTURE WORK . . . 270
9.1 Conclusions . . . 271
9.2 Future directions . . . 276

REFERENCES . . . 281

LIST OF FIGURES

2.1 The DXXX-OPV3-XXXD system. (a) Chemical structure of the prototypical DXXX-OPV3-XXXD peptide monomer. The oligophenylenevinylene π core (OPV3) is flanked by oligopeptide wings (DXXX) that are mirror symmetric such that the identity of the amino acids is inverted and the molecule possesses two C-termini. The X residues are selected from the 20 natural amino acids such that the family comprises 20^3 = 8,000 distinct molecules. (b) Molecular simulation snapshot of a self-assembled pseudo-1D nanoaggregate spontaneously formed by the spontaneous association of DVAA-OPV3-VAAD peptide monomers into a linear stack. Good stacking between the π-cores (colored blue) favors π orbital overlap, electronic delocalization along the backbone of the nanoaggragate, and the emergence of optical and electronic functionality. (c) Experimental transmission electron microscopy (TEM) image of self-assembled fibrils formed by DFFG-OPV3-GFFD peptides in an acidic environment. Reprinted with permission from Ref. [185] . Copyright (2014) American Chemical Society. (d) Illustration of the mapping from the DGAG-OPV3-GAGD all-atom structure to the coarse-grained representation at which the simulations in this work are conducted. The coarse-grained beads corresponding to groupings of neighboring atoms are labeled according to the Martini model employed in this work [193, 194, 195]. 19

2.2 Active learning cycle for the data-driven discovery of optimally self-assembling DXXX-OPV3-XXXD peptides. The cycle contains four components. (1) Coarse-grained molecular simulations are performed on selected DXXX-OPV3-XXXD candidates and the quality of the self-assembled aggregates formed by molecule i measured according to a scalar fitness f_i. (2) The DXXX-OPV3-XXXD family is projected from the high-dimensional chemical space of molecular structures into a low-dimensional latent space embedding $E : \text{DXXX-OPV3-XXXD}_i \rightarrow \mathbf{z}_i \in \mathbb{R}^d$ using a variational autoencoder (VAE). The dimensionality of the latent space is optimized during each cycle. (3) A Gaussian process regression (GPR) model is constructed over the VAE latent space linking the latent space coordinates of each DXXX-OPV3-XXXD family member to the scalar fitness function measuring the quality of their self-assembled aggregates $f : \hat{f}_i = f(\mathbf{z}_i) = (f \circ E)(\text{DXXX-OPV3-XXXD}_i)$. The GPR mapping f is retrained each cycle over all DXXX-OPV3-XXXD candidates that have been simulated to date and for which measures of the fitness function $f_i,\ i \in \textit{sampled}$ is available, and is then used to predict the fitness of unsimulated candidates $\hat{f}_j,\ j \in \textit{unsampled}$. (4) The predicted means and uncertainties for $\hat{f}_j,\ j \in \textit{unsampled}$ furnished by the GPR surrogate model are combined within an active learning acquisition function to identify the "next-best" candidates for which to perform coarse-grained molecular simulations to drive sampling towards the most promising candidates. The loop is cycled until the GPR surrogate model no longer changes with additional data collection and can then be used to reliably identify the top candidates for computational validation. 25

2.3 Dynamical evolution of the self-assembled structures of DDAI-OPV3-IADD over the course of a 3 μs coarse-grained molecular simulation. Snapshots of the molecular simulation show molecules in which the OPV3 π cores are colored blue, the peptide wings faded grey, and water is removed for clarity. The interaction graph corresponding to each snapshot is shown directly below each image. The vertices V corresponding to each peptide are colored red, and edges E between peptide pairs defined by $d_{ab}^{\text{optical}} < r_{\text{cut}} = 0.7$ (Eqn. 2.1) and colored grey. The average degree $\kappa = \frac{2|E|}{|V|}$ is reported at the bottom of each panel. At $t = 0$ ns, the 96 randomly placed peptides form essentially a monomeric dispersion, with the exception of five dimer pairs, and the system possesses a correspondingly low $\kappa = 0.13$. As the simulation progresses, the peptides self assemble under the influence of hydrogen bonding, π-π stacking, and hydrophobic interactions into small ($t = 50$ ns, $\kappa = 3.69$) and then larger ($t = 500$ ns, $\kappa = 5.67$; $t = 3{,}000$ ns, $\kappa = 6.08$) aggregates with a commensurate increase in the mean degree κ. In this figure, and throughout the paper, molecular renderings are generated using VMD [211], and interaction graphs produced using NetworkX [212]. 28

2.4 Schematic of the representation of each XXX tripeptide to the VAE. The Martini representation of each tripeptide i, in this example Trp-Val-Tyr (WVY), is converted into an adjacency matrix $\mathbf{A}_i \in \mathbb{R}^{15\times15}$ specifying the connectivity of beads within the tripeptide and a one-hot composition vector $\mathbf{T}_i \in \mathbb{R}^{75}$ specifying the identity of the beads. The tuple $(\mathbf{A}_i, \mathbf{T}_i)$ defines the input provided to the VAE. A pre-defined sequential numbering is employed for the beads in each amino acid. The adjacency matrix is padded with rows and columns of zeros to represent tripeptides containing fewer than the maximum number of 15 beads. The colored blocks in the adjacency matrix and composition vector correspond to the colors of the amino acids in the Martini molecule. 29

2.5 Architecture of the variational autoencoder (VAE) used to generate the DXXX-OPV3-XXXD latent space embedding. The VAE accepts as inputs adjacency matrix and composition vector tuples $(\mathbf{A}_i, \mathbf{T}_i)$ and employs two parallel encoders to perform feature extraction and learn the mean $\boldsymbol{\mu}_i$ and standard deviation $\boldsymbol{\sigma}_i$ of a Gaussian distributed latent space embedding $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \in \mathbb{R}^d$. The decoder generates approximate reconstructions $(\hat{\mathbf{A}}_i, \hat{\mathbf{T}}_i)$ of the inputs from the latent space representation. The network is trained by minimizing a loss function balancing reconstruction accuracy and a regularization term constraining the latent space to follow a multidimensional Gaussian distribution (Eqn. 8.1). The dimensionality d of the latent space is treated as a hyperparameter that is optimized during each cycle of active learning. 31

2.6 Illustrative visualization of $d = 3$ VAE latent space embedding of the DXXX-OPV3-XXXD family. The embedded molecules are colored according to the number of beads in the XXX tripeptide and selected molecules are visualized. The first dimensional of the latent space z_1 is correlated with molecular size $\rho(z_1, \text{size}) = 0.858$ (p-value $< 1\times10^{-15}$) providing a visual illustration that similar molecules are embedded close together in the VAE latent space. 33

2.7 Active learning candidate selection. The expected improvement (EI) acquisition function (Eqn. 2.12) is evaluated over for all unsampled members of the DXXX-OPV3-XXXD family at values of the exploit-explore hyperparameter ξ over the range $\log_{10}\xi \in [-4, 0.4]$. The blue points in the graph indicate which DXXX candidate maximizes the EI at each value of ξ. In this illustrative example there are four candidates – DGGG, DEGA, DGLI, and DALW – that maximize the EI over the range of ξ considered. The vertical line shows the recommended value of $\xi = 0.01$ suggested in literature [55, 67], which would result in the selection of only DEGA as the next molecule to simulate. In our approach, we select all four of the molecules DGGG, DEGA, DGLI, and DALW that maximize EI over the entire range of ξ considered as the next best candidates to simulate in parallel in the next round of coarse-grained molecular simulations. 38
2.8 Tracking of active learning stop criteria. (a) The performance difference (PD) method 5-fold cross validation R^2 score stabilizes at $R^2 \sim 78\%$ by Round 18. (b) The stabilizing predictions (SP) method Bhattacharyya distance between successive GPR posteriors plateaus close to zero at $D_B \sim 2.5$ after Round 13. Values are reported using a moving average with a window size of three, as recommended in Ref. [72] 44
2.9 Diffusion map embeddings into $\psi_2 - \psi_4$ of the $N = 186$ DXXX-OPV3-XXXD molecular simulation trajectories. (a) Composite embedding of all 558,000 simulation snapshots. Each point represents a snapshot from one of the simulation trajectories and points are colored by the total number of edges in the corresponding molecular interaction graph (Section 2.3.2). Temporal assembly courses of selected molecules over the 2D manifold where points are colored by simulation time: (b) DVAA-OPV3-AAVD (good assembler: $\overline{\kappa} = 5.85 \pm 0.03$, rank = 4/186), (c) DGEG-OPV3-GEGD (intermediate assembler: $\overline{\kappa} = 5.02 \pm 0.02$, rank = 89/186), (d) DWWI-OPV3-IWWD (poor assembler: $\overline{\kappa} = 3.74 \pm 0.01$, rank = 175/186). 49
2.10 Spectral clustering of the N=186 sampled molecules into four clusters which are projected into the top three non-trivial diffusion map eigenvectors $\{\psi_i\}_{i=2}^4$. Points are clustered based on agglomerative hierarchical clustering and cutting the resulting dendrogram at the level of three clusters. Grey points represent instantaneous snapshots harvested from all molecular simulations projected into the ψ_2-ψ_4 plane. Colored points represent the average coordinates over the terminal 50 ns of simulation for each of the $N = 186$ molecules. Points are color coded by cluster: green (good assemblers), red (intermediate assemblers), and orange (poor assemblers) 51
2.11 Residue enrichment analysis of each position X_i in $DX_1X_2X_3$-OPV3-$X_3X_2X_1D$ within molecules classified as good assemblers relative to those classified as intermediate or poor assemblers. Good assemblers are enriched in amino acids residing below the dashed black line and depleted in those residing above it. Dashed green and red lines show boundaries for 1.5× and 2× differential enrichment and depletion. 53

3.1 Schematic of the hybrid computational/experimental active learning workflow for the discovery of self-assembling X_n-4T-X_n π-conjugated peptides. We perform separate computational and experimental active learning loops within a shared low-dimensional latent space embedding of the molecular design space learned using regularized autoencoders (RAE). Each active learning loop then consists of three parts. (i) Evaluating the quality of a given molecule k in the X_n-4T-X_n design space by either performing high-throughput all-atom molecular dynamics simulations to measure the average number of contacts per molecule $\kappa^{(k)}$ and radius of gyration ${R_g}^{(k)}$ of the self-assembled nanoaggregate or low-throughput experimental synthesis and measurement of the blue-shift $\lambda^{(k)}$ in the mode of the UV-vis spectrum. (ii) Fitting surrogate models using Gaussian process regression (GPR) to predict the performance of untested candidates given the accumulated simulation and experimental data collected to date. Two separate GPRs are maintained for the two computational objectives GPR_κ and GPR_{R_g}. We build a multi-fidelity GPR (mfGPR) as our experimental surrogate model GPR_λ that also incorporates data from the computational GPRs to improve prediction accuracy beyond what would be possible from the limited experimental data alone. (iii) Employing Bayesian optimization (BO) to interrogate the GPR model and select the next most promising molecular candidates for computation as those lying on the κ-R_g Pareto frontier and for experimentation as those with large values of λ. The molecular renderings in this figure, and throughout the paper, are generated using the Visual Molecular Dynamics (VMD) software [252]. 62

3.2 Progress and convergence of the hybrid computational/experimental active learning screen for high-performing X_n-4T-X_n molecules within the design space of 694,982 candidates. (a) A total of 38 computational rounds of screening are performed and interleaved with three rounds of experimental screening. The experimental screens are conducted at computational Rounds, 0, 14, and 22. Convergence of the computational screen is tracked by $d^{(i)}_{\text{Pareto}}$, the mean distance from the origin of all molecules cumulatively simulated over the first i rounds within the computational κ-R_g Pareto plot (see panel b), as a measure of the advancement of the optimal frontier (red circles). Convergence of the experimental screen is tracked by the measured spectral blue shift λ quantifying the degree of H-type co-facial π-stacking within the self-assembled nanoaggregates (blue violin plots). Violin plots enclose the range of λ values measured each experimental round and the central white tick denotes the mean λ value for the round. The gray shaded area represents the computational rounds used to verify convergence of the active learning but not used to inform any additional rounds of experimental design. (b) Round-by-round advancement of the computational κ-R_g Pareto frontier over the course of the 38 screening rounds. Within each frame, the points corresponding to X_n-4T-X_n candidates collected within that round are shown in red, those points collected in previous rounds are shown in gray, the Pareto frontier defined by the Pareto optimal points is shown as a black solid line, and the shaded green area indicates the advancement of the Pareto frontier relative to Round 0. 75

3.3 Embedding of the computational and experimental molecules sampled in the active learning screen into the κ-R_g objective function space. (a) Embedding of the 1181 X_n-4T-X_n molecules explored in the computational screen. Points shown in the same color were sampled during the same computational active learning round. The labels associated with each point corresponds to the X_n peptide wing sequence. (b) Highlighting the 28 experimentally tested X_n-4T-X_n molecules (colored points) superposed onto all 1181 computationally simulated points (grey points). The color indicates the round of experimental active learning within which the molecule was tested. Encircling the plot are snapshots from our molecular simulation trajectories showing the terminal self-assembled nanoaggregates. The heavy atoms constituting 4T π-cores are rendered as gold space-filling spheres and the X_n peptide wings as semi-transparent ball-and-stick representations. Molecules producing aggregates with large values of both κ and R_g tend to self-assemble into pseudo-1D structures with good stacking of the 4T π-cores. For the experimentally tested candidates we also report the measured values of the spectral blue shift λ. A full accounting of the computed κ and R_g values and measured λ values for all molecules considered in our screen have been made available as detailed in the Data Availability statement [251]. 78

3.4 Two-dimensional diffusion map embedding of 118,100 simulation snapshots harvested from the molecular simulations of 1181 X_n-4T-X_n molecules collected over the course of the computational active learning screen. Each point represents one simulation snapshot projected into a 2D low-dimensional manifold spanned by the leading collective variables $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ learned by diffusion maps. Coloring the points by (a) the log radius of gyration $\log(R_g)$ of the self-assembled nanoaggregate and (b) average number of contacts per molecule κ within the aggregate exposes the strong correlation of the two learned collective variables with these two physical observables ($\rho(\boldsymbol{\psi}_2, R_g) = 0.93$, $\rho(\boldsymbol{\psi}_3, \kappa) = 0.69$). All assembly trajectories commence in a monomeric dispersion contained at the rightmost edge of the manifold. Progression from right-to-left corresponds to a reduction in R_g as the system self-assembles, and progression from bottom-to-top to the formation of more molecular contacts. The temporal progression of the self-assembly pathways over the manifold for molecules (c) EGG-4T-GGE, (d) DVAG-4T-GAVD, (e) DLAG-4T-GALD, and (f) EYIQG-4T-GQIYE, in which points are colored temporally. Grey points in panels c-f represent the embedding of 118,100 simulation snapshots shaded in panels c and e by $\log R_g$, and panels d and f by κ. Insets show representative molecular renderings throughout the trajectory. . . . 83

3.5 Terminal locations of the X_n-4T-X_n self-assembly trajectories over the 2D diffusion map manifold. End points of the 1181 molecules considered in our computational screen colored by the (a) log of the radius of gyration $\log(R_g)$ and (b) average number of contacts per molecule κ computed over the terminal 50 ns of the trajectory. The 118,100 simulation snapshots used to construct the diffusion map embedding are shown in grey. (c) Terminal locations of the 28 experimentally tested molecules colored by their measured spectral blue shift λ and annotated with the sequence of the peptide wing. The purple box bounds a cluster of high-performing experimental candidates residing in the upper-left region of the manifold possessing high values of λ. 86

4.1 (a) Schematic representation of a mitochondrion with lipid bilayer system representing the inner mitochondrial membrane (IMM). (b) Chemical structures of (b) Cardiolipin (CL) and (c) Phosphatidylglycerol (PG) phospholipids. The CL molecule is essentially a dimer of PG. CL contains two phosphatidyl moieties and one hydroxyl group in its headgroup, in comparison one phosphatidyl group and two hydroxyl groups in the PG headgroup. (d) The cardiolipin probe 10-N-nonyl acridine orange (NAO). All structures were drawn with ChemSketch.[306] 92

4.2 Active learning cycle for the automated discovery of small ($\leq$500 Da) molecules with high CL selectivity. (a) Coarse-grained molecular dynamics (CGMD) simulations are performed to calculate the partitioning free energy of a molecule into PG and CL membranes used to evaluate the CL selectivity $\Delta\Delta G$. (b) Deep representational learning using a regularized autoencoder (RAE) is used to construct a fixed-size latent space embedding of the discrete molecular design space encompassing all coarse-grained molecular candidates. Gaussian process regression (GPR) surrogate models are fit using all accumulated simulation data to predict the CL selectivity $\Delta\Delta G$ of all untested compounds within the design space. (c) These surrogate model predictions are then interfaced with a Bayesian optimization platform to select the next most promising compounds for computational simulation. This process iteratively continues until multiple consecutive active learning rounds fail to identify new top-performing compounds. (d) Using our accumulated simulation data we extract design rules linking the inclusion/omission of chemical functional groups to the degree of CL selectivity $\Delta\Delta G$ by building of interpretable linear models using graph representational learning. (e) The design rules are used to select candidate compounds that are subjected to fluorescence anisotropy measurements as well as CG free energy calculations to validate our findings. Structures in this panel are drawn with ChemSketch,[306] the icons were obtained from Flaticon.com. 95

4.3 **(a)** Distribution of calculated thermodynamic selectivities for CL relative to PG membranes $\Delta\Delta G$ for CG compounds simulated in each active learning round. Each active learning cycle employs Bayesian optimization to select 60 candidate molecules for which we calculate $\Delta\Delta G$ using alchemical free energy calculations and which are used to iteratively retrain and update the model for subsequent rounds of selection. **(b)** Latent space embeddings of simulated compounds visualized in a 2D principal component analysis projection of the 16-dimensional RAE latent space. Points are colored by their calculated $\Delta\Delta G$. Dark blue points represent compounds that fail to spontaneously insert into the membrane bilayer and are deemed not interfacial and hence are terminated early within the workflow without finalizing the $\Delta\Delta G$ calculation. Model predictions become more accurate round by round as data is accumulated enabling the acquisition function to progressively better localize sampling in regions of latent space rich in high-performing compounds. 107

4.4 **(a)** Best sampled $\Delta\Delta G$ for each active learning round alongside the cumulatively best sampled molecule in each round. The top axis denotes the number of compounds simulated after each round and in parentheses as a percent of the 124,327 molecule design space. Errors represent uncertainly in the measured $\Delta\Delta G$ for the associated compound. The best candidate was selected after round four, with the final three rounds failing to identify any new better performing candidates. **(b)** The 12 best coarse-grained molecules identified over all seven active learning rounds sorted by increasing $\Delta\Delta G$. Large negative values of $\Delta\Delta G$ correspond to high thermodynamic selectivity for CL over PG membranes. 109

4.5 Rank-ordering of learned coefficient weights θ_k corresponding to CG motifs for predicting measured $\Delta\Delta G$. Each compound n is featurized by calculating the normalized frequency of all k topologically unique subgraphs consisting of 1-5 CG beads, denoted $F_{n,k}$. Using this featurization we perform LASSO regression to predict the calculated $\Delta\Delta G$ obtained from the alchemical free energy calculations using Eqn. 4.2: $\Delta\Delta G^{(n)}_{\text{predicted}} = \Delta\Delta G_{\text{mean}} + \sum_{k=1}^{k=1608} \theta_k F_{n,k}$. We then extract the associated largest and smallest learned coefficient weights θ_k from the reduced feature set retained by the LASSO model. The most negative coefficient weights correspond to chemical motifs indicative of small $\Delta\Delta G$ and good CL selectivity, while the most positive weights correspond to motifs predicative of large $\Delta\Delta G$ and poor CL selectivity. As the weights θ_k have units of kcal/mol these weights can be interpreted as the magnitude of the influence of the corresponding subgraph motif on CL selectivity. 112
4.6 Five functional groups most frequently mapped to the uncharged beads of the *5+1* reduced Martini model.[314, 85] Dark colors represent higher mapping probabilities. 113
4.7 Predominant functional groups mapped to the charged bead type Q0± within our *5+1* reduced Martini model.[314, 85] (a) Functional groups with ap$K$a ≤ 7 are most likely to be negatively charged under physiological conditions (pH ≈ 7.3) and are mapped to the Q0- bead. (b) Those with bp$K$a ≥ 7 are most likely to be positively charged and are mapped to the Q0+ bead. 114
4.8 Backmapping of a coarse-grained T3-T5 dimer. The ensemble of 55639 GDB-derived small molecules consistent with any coarse-grained dimer are projected as grey points into the 2D plane spanned by the two order parameters of molecular weight and octane to water transfer free energy $\Delta G_{O\to W}$. We highlight in red the 2157 GDB-derived small molecules whose coarse-grained mapping corresponds to a T3-T5 dimer. The chemical structures of seven molecules are visualized. A large number of diverse chemical structures map to a particular CG representation. 116
4.9 Comparison of experimental measurements and computational predictions of the selectivity of two candidate molecules quinaldine red and benzothiazolium selected according to our learned design rules within three model membranes consisting of 100% PC, 90% PC / 10% PG, and 90% PC / 10% CL. (a) Experimental measurements of the differential fluorescence anisotropy ΔA. The negative control confirms negligible differences between the different membranes in the measured value of ΔA. Measurements of ΔA for NOA were not reported due to technical issues. (b) Computational predictions of the transfer free energies $\Delta G_{\text{W}\to\text{I}}$ from water to the membrane interface and from octane (representative of the membrane midplane) to the interface $\Delta G_{\text{O}\to\text{I}}$ calculated using the *5+1* CG model and alchemical transformations. Calculated uncertainties of the computational results are smaller than the symbol size. (c) Computational predictions of the relative partitioning free energy $\Delta\Delta G = \Delta G^{\text{CL}}_{\text{W}\to\text{I}} - \Delta G^{\text{PG}}_{\text{W}\to\text{I}}$. 118

5.1 Chemical structure of the different π-conjugated peptide molecules built from six unique aromatic cores that we screen in this work. The mirror symmetric oligopeptide wings VEVAG-Π-GAVEV are highlighted in blue and which flank one of six different aromatic cores $\Pi \in \{\text{PTP, DFT, QTP, NDI, PDI, BDI}\}$ separated from the wings by alkyl spacers of length $n \in \{0, 1, 2, 3\}$. By considering all combinations of π-cores and alkyl chain lengths, our screening library comprises a total of 6×4=24 candidate π-conjugated peptide molecules. 126
5.2 Schematic illustration of our computational pipeline for measuring π-conjugated peptide charge mobility. **(i)** All-atom classical molecular dynamics (MD) simulations are conducted to predict the morphology of the self-assembled aggregates formed by each π-conjugated peptide. Electronically interacting dimer pairs are harvested and passed to density functional theory (DFT) calculations. DFT calculations are performed to compute **(ii)** the chemical stability of molecules supporting a hole or electron within an aqueous environment, **(iii)** the reorganization energy $\lambda_{\pm}$, and **(iv)** the transfer integral $V_{\pm}$. **(v)** Marcus theory is then used to predict the intermolecular drift mobility $\mu_{\pm}$ of holes/electrons within the nanoaggregates as a figure of merit for the potential optoelectronic functionality of the candidate π-conjugated peptide. Molecular visualizations were constructed using Visual Molecular Dynamics (VMD) [352] and molecular orbitals visualized using VESTA [353]. 128
5.3 The chemical stability of the cationic and anionic states of each oligopeptide in water. (a) Calculated ionization potentials relative to the water oxidation potential of +1.23 V (SHE). Molecules residing above the dashed horizontal line are predicted to support stable cationic states and serve as putative p-type semiconductors. (b) Calculated electron affinities relative to the water reduction potential of −0.83 V (SHE). Molecules residing below the dashed line are predicted to support stable anionic states and serve as putative n-type semiconductors. 138
5.4 Reorganization energies of the chemically stable aromatic core libraries. Smaller reorganization energies are associated with higher charge mobility. (a) Reorganization energies of the stable cationic molecules associated with hole transport in p-type semiconductors. (b) Reorganization energies of the stable anionic molecules associated with electron transport in n-type semiconductors. Error bars represent the standard error in the mean. 141
5.5 Transfer integrals for the chemically stable aromatic core libraries. Larger transfer integrals are associated with higher charge mobility. (a) Transfer integrals between stable cationic molecules associated with hole transport in p-type semiconductors. (b) Transfer integrals between stable anionic molecules associated with electron transport in n-type semiconductors. Error bars represent the standard error in the mean. 142

5.6 Charge mobility predicted from the calculated reorganization energies and transfer integrals using Marcus theory. (a) Predicted hole mobilities in stable cationic molecules forming putative p-type semiconductors. (b) Predicted electron mobilities in stable anionic molecules forming putative n-type semiconductors. Error bars represent the standard error in the mean. 144

6.1 Illustration of training and testing setups. 152
6.2 Schematic illustration of our model training and inference setups. 155
6.3 Comparison of the internal potential energy distributions for atomistic and backmapped ADP trajectories. **(a)** Internal energies for the held out test set trajectory and the atomistic reconstructed trajectory generated by backmapping a manual coarse graining of the original atomistic trajectory. **(b)** Generalization test comparing reference atomistic trajectories obtained from Ref. [412] to our model-generated backmappings of a real CG simulation conducted using CGSchNet [152]. 166
6.4 Comparison of atomistic and backmapped MSM-reweighted Free Energy Surfaces (FES) for ADP. Ramanchandran plots for the in distribution test set that includes **(a)** a held-out atomistic trajectory and **(b)** our model generated backmapping of the manually coarse grained atomistic trajectory. For a more challenging test of generalizability we compare the Ramachandran plots for **(c)** a reference atomistic trajectory taken from Ref. [412] and **(d)** a backmapping of a CG simulation performed using CGSchNet [152]. Labeled in **(c)** are phase space locations for the five meta-stable states of ADP. **(e)** Seven superimposed configurations near each of these five major meta-stable states from **(c)** the reference atomistic, **(d)** the CG reconstructed and **(b)** the test set reconstructed trajectories. 168
6.5 Implied timescales of atomistic and backmapped CG trajectories for ADP. **(a)** Comparisonon of timescales recovered for the test set atomistic and test set reconstructed trajectories. **(b)** Implied timescales calculated for the reference atomistic data taken from Ref. [412] compared to the backmapped CG simulation performed with CGSchNet [152] along with the original CG simulation. Timescales for the backampped and CG simulated data are normalized such that the dominant processes is the same as the reference atomistic trajectory. Errors are estimated with Bayesian sampling and represent a 95% confidence interval of the timescales of the sampled MSMs. 170
6.6 Comparison of atomistic and backmapped root mean square velocity distributions for ADP. Velocities are calculated atom-wise between sequential frames using a finite forward difference method and aggregated into a distribution over all atomic velocities. Velocities for the CG reconstructed data corresponding to the generalization set are rescaled by a constant factor ($\sim$3.25) such that the mean of the CG reconstructed velocities matches the mean of the test set atomistic velocities. 172
6.7 Potential energy distributions of atomistic and backmapped CLN trajectories for **(a)** the in distribution test set and **(b)** the generalization set. 174

6.8 Comparison of the atomistic and backmapped MSM-reweighted Free Energy Surface (FES) for CLN. The FES for CLN is represented in the basis of the two leading non-trivial Independent Components (ICs) of a Time-lagged Independent Component Analysis (TICA) model fit to the reference atomic dataset. Shown are FES of the in distribution test set atomistic **(a)** and backmapped **(b)** trajectories, alongside the reference atomistic data **(c)** and the corresponding backmapped CGSchNet simulation **(d)** from the generalization set. **(e)** Visualization for a collection of 25 superposed structures from the three meta-stable states taken from the reference atomistic data **(c)**, the backmapped CG reconstructed data **(d)** and backmapped in distribution test set trajectory **(b)**. 176
6.9 Implied timescales of atomistic and backmapped CG trajectories for CLN. **(a)** Comparison of implied timescales from atomistic and backmapped trajectories for the in distribution test set. **(b)** Implied timescales for the reference atomistic data, the original CGSchNet simulation and the backmapped CGSchNet simulation. Timescales for the original CG simulation and backmapped data are normalized such that the dominant (slowest) processes matches the slowest reference atomistic timescale. 178
6.10 Comparison of atomistic and backmapped root mean square velocity distributions for CLN. Velocities are calculated frame-by-frame using a finite forward difference method and represented in angstrom / 100 ps, as 100 ps is the native temporal spacing between frames in the reference atomistic data [409]. The backmapped generalization set velocities are rescaled by a constant factor ($\sim$17.81) such that the mean of the CG reconstructed velocities equals the mean of the test set atomistic velocities. 179

7.1 Schematic illustration of four-step iterative Girsanov Reweighting Enhanced Sampling Technique (GREST) workflow. **(a)** Initial training data is generated by unbiased or biased simulation data accelerated along intuitive CVs. **(b)** Slow CVs are estimated from the training data using Girsanov path reweighted State-free Reversible VAMPnets (SRVs). **(c)** Enhanced sampling calculations are performed within the learned slow CVs using a CV biasing technique such as metadynamics. **(d)** Convergence of the iterative cycle is assessed by evaluating the stability of the learned CVs in successive iterations. 206

7.2 Application of GREST to the 1D 4-well potential as an illustrative toy example. **(Left column.)** Application of GREST under pre-defined bias. **(a)** The target 4-well potential V_{target} (blue) is shown alongside the 2-well potential under which the biased simulation is performed V_{sim}. The challenge is for path reweighted SRVs to correctly recover the slowly relaxing modes and associated time scales of the 4-well system in simulations conducted under the 2-well potential. **(b-d)** Comparison of the leading three leading slow eigenfunctions $\{\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \phi_3(\mathbf{x})\}$ recovered by path reweighted SRVs from long unbiased reference data V_{target} (green), alongside eigenfunctions learned using biased simulation data collected under V_{sim} to which we apply path weights comprising dynamical and thermodynamic corrections (orange), only thermodynamic corrections (red), and no reweighting corrections (blue). Eigenvectors are defined only up to scaling and sign so are presented under unit normalization and, if necessary, inversion of sign to align them with the unbiased reference data. **(e)** Associated implied time scales $t_i = -\tau/\ln(\lambda_i)$ for the SRV eigenfunctions in panels **(b-e)** with error bars showing the standard deviation over 10 independently trained models with different random seeds. **(Right column.)** Application of GREST under a learned bias. **(f)** A biasing potential U_{bias} is constructed by projecting the unbiased reference data into the leading slow mode $\phi_1(\mathbf{x})$ learned by a path reweighted SRV fit to the unbiased 4-well simulation trajectories. Application of the bias to the target potential results in a simulation potential $V_{sim} = V_{target} - U_{bias}$ that attenuates the highest central potential energy barrier by $\sim$0.5 kJ/mol and the other potential energy barriers by $\sim$0.25 kJ/mol. **(g-i)** Comparison between the reference eigenfunctions and those estimated from biased simulation data under V_{sim} in panel **(f)** to which we apply various path weights. **(j)** Associated implied time scales $t_i = -\tau/\ln(\lambda_i)$ for the SRV eigenfunctions in panels **(g-i)** with error bars showing the standard deviation over 10 independently trained models with different random seeds. 211

7.3 Application of GREST to ADP. **(a-i)** Reference free energy profile constructed from 4 μs of unbiased simulation data projected into the ϕ-ψ backbone angles. The local free energy minima are annotated with their associated macrostate labels: C_5, P_{II}, α_R, α_L, and C_{7ax}. **(a-ii,iii)** Leading two slow CVs $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ identified by application of SRVs to the unbiased reference trajectories. $\phi_1(\mathbf{x})$ characterizes transitions in the ϕ dihedral between $\{C_5, P_{II}, \alpha_R\}$ and $\{\alpha_L, C_{7ax}\}$. $\phi_2(\mathbf{x})$ primarily characterizes transitions in the ψ dihedral between $\{C_5, P_{II}\}$ and α_R. Rows **(b)-(g)** index successive rounds of GREST, commencing with the initial unbiased Round 0 in **(b)** and ending with the terminal adaptive sampling Round 5 in **(g)**. Columns **(i)-(iv)** index various facets of each round of GREST: **(i)** illustrates the unbiased free energy landscape estimated by thermodynamic reweighting of the enhanced sampling simulations performed under the fixed terminal metadynamics bias projected into the ϕ-ψ backbone angles, **(ii)** and **(iii)** illustrate heatmaps of the path reweighted SRV slow modes $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ learned from the enhanced sampling trajectories in the terminal metadynamics bias potential projected into ϕ-ψ space, and **(iv)** illustrates the phase space sampling density projected in ϕ-ψ space from well-tempered metadynamics simulations performed in the two learned slow modes $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ illustrated in the preceding columns. The GREST iterations are terminated when the CVs learned in successive rounds stabilize, indicating no additional enhancement in configurational and dynamical exploration of phase space. Convergence of GREST at Round 4 means that additional enhanced sampling calculations are not conducted in Round 5. 217

7.4 Convergence assessment of GREST applied to ADP by measuring the Pearson correlation coefficient between SRV slow CVs discovered in successive rounds. Stabilization of the CVs to $\rho > 0.8$ indicates that no important new dynamical modes or regions of configurational phase space are being explored under additional rounds of biased sampling. 221

7.5 Unbiased reference simulations of WLALL pentapeptide. **(a)** The free energy surface resulting from projection of 6 μs of unbiased simulation data into the leading two TICA coordinates identifies three predominant metastable minima marked by golden stars and annotated with their corresponding relative free energy. Insets show a collection of representative structures from each minima identifying the *(1)* unfolded state, *(2)* folded state, and *(3)* misfolded state. **(b)** Application of SRVs to the unbiased reference data resolves a leading slow mode $\phi_1(\mathbf{x})$ with an implied time scale of $t_1 = 19.6$ ns characterizing transitions in and out of the misfolded state. **(c)** The next slowest mode $\phi_2(\mathbf{x})$ possesses an implied time scale of $t_2 = 11.3$ ns and corresponds to transitions in and out of the folded state. . . 223

7.6 Application of GREST to WLALL pentapeptide. Rows **(a)-(f)** index successive rounds of GREST, commencing with the initial unbiased Round 0 in **(a)** and ending with the terminal adaptive sampling Round 5 in **(f)**. Columns **(i)-(iv)** index various facets of each round of GREST: **(i)** illustrates the unbiased free energy landscape estimated by thermodynamic reweighting of the enhanced sampling simulations performed under the fixed terminal metadynamics bias data projected into the two leading TICA coordinates TIC_1 and TIC_2, **(ii)** and **(iii)** illustrate heatmaps of the path reweighted SRV slow modes $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ learned from the enhanced sampling trajectories in the terminal metadynamics bias potential projected into TIC_1-TIC_2 space, and **(iv)** illustrates the phase space sampling density projected into TIC_1-TIC_2 space from well-tempered metadynamics simulations performed in the two learned slow modes $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ illustrated in the preceding columns. The GREST iterations are terminated when the CVs learned in successive rounds stabilize, indicating no additional enhancement in configurational and dynamical exploration of phase space. Convergence of GREST at Round 4 means that additional enhanced sampling calculations are not conducted in Round 5. 226

7.7 Convergence assessment of GREST applied to WLALL pentapeptide by measuring the Pearson correlation coefficient between SRV slow CVs discovered in successive rounds. Stabilization of the CVs to $\rho > 0.8$ indicates that no important new dynamical modes or regions of configurational phase space are being explored under additional rounds of biased sampling. 228

8.1 Schematic illustration of DiAMoNDBack (Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping). **(a)** Atomistic detail is restored from Cα traces by backmapping the structure residue-by-residue from the N-terminus to the C-terminus. Each residue prediction task is performed within a canonical reference frame and depends on the local environment around the target Cα (red sphere) comprising the N most spatially proximate residues to the target Cα (green spheres). **(b)** The residue prediction task is performed using a Denoising Diffusion Probabilistic Model (DDPM) that learns to reverse an M-step noising process applied to real samples. The trained DDPM model can then operate on random noise to recover realistic-looking samples. **(c)** Learning the denoising process involves training a U-net model that is designed to predict the noise added to a corrupted sample. The input to the network are Cartesian coordinate representations of the target residue, the local environment, and a one-hot encoding of the residue identity. Conditioning is achieved by only noising and regressing on components of the representation that are allowed to change throughout the diffusion steps in the residue prediction block such as the backbone C, N and O atoms and the side chain atoms. Atoms comprising the partially-decoded local environment and the Cα backbone, along with the one-hot residue type encoding only serve as conditioning information passed to the network, are fixed throughout the diffusion steps, and do not contribute to the loss. 257

8.2 Visualization of protein structures used for training and testing in the PED, PDB, and DES training sets. **(a)** Distribution of sequence lengths. **(b)** Distribution of structural diversity in the space of average percent alpha-helix and beta-sheet content. Disordered/unstructured structures reside close to the origin. Colors correspond to the histograms in panel (a): PED = blue, PDB = green, and DES = orange. **(c)** Test set proteins for each training data set shown in the same space of alpha-helix and beta-sheet content. The size of each point corresponds to the number of residues in the protein. For the PED and DES data, multiple frames are present for each protein and error bars indicate the standard deviation across frames. Selected representative training structures are labeled A-F and the corresponding structures rendered along the bottom of the figure. 258
8.3 Visualization of four protein structures backmapped using our PDB-trained model from the PDB test set (TBM#T0922 and TBM-hard#T0912) and the DES MD test set (NTL9 and UVF). Five samples generated using the Cα backbone as conditioning are shown as translucent structures, with the associated atomistic reference structure shown as a single opaque structure. Insets for each protein zoom into regions highlighting the generative diversity in the side chain placements. 260
8.4 Comparison of side chain dihedral angle distributions in the DES test-set data partitioned by residue identity for generated structures produced by the rules-based approach, our PDB-trained model, and the PDB-trained model fine-tuned on DES data. Jensen-Shannon divergence scores for dihedral distributions across all 17 enumerated residues, with colored stars indicating the select distributions visualized below. Dihedral angles are shown for the C-Cα-Cβ-Cγ atoms enumerated for 17/20 available amino acids containing these atoms, not applicable for GLY and ALA and there are no CYS residues present in the 11 MD sequences. Distribution of dihedral angles for three exemplar residues ASP (top), PRO (middle), and SER (bottom) labeled with Jensen-Shannon divergence scores (JS) between the reference and generated distributions. Errors in the JS score are calculated as standard deviations over generated samples for our model predictions, with no associated error prescribed to the deterministic rules-based approach. 262
8.5 Comparison of clash score and bond quality evaluation metrics over the 11 sequences from the MD trajectory data for the PDB-trained and DES fine-tuned models. **(a)** A split violin plot shows the distribution of clash scores (top) and bond qualities (bottom) between generated samples produced by the PDB-trained and fine-tuned models. **(b)** Scatter plot showing the mean bond quality and clash score for the 24 test-set PDB sequences evaluated with the PDB-trained model (blue points) and the 11 MD trajectory sequences for the PDB-trained model (orange) and DES fine-tuned model (green). Grey dashed lines connect the points corresponding to the mean values associated with each sequence between the PDB-trained and DES fine-tuned models. 263

8.6 Analysis of per-residue bond qualities and clash scores. Scatter plots show the bond quality and clash score for each residue sized by the number of associated side chain atoms and colored by residue-specific properties such as charge, hydrophobicity, polarity, and special (PRO and CYS) with specific more poorly performing residues TRP, LYS, and ARG annotated. Evaluations with the PDB-trained model on the 24 test-set PDB sequences are shown in **(a)**, the same PDB-trained model evaluated on the 11 test-set MD trajectories are shown in **(b)** with the same MD test-set evaluated with the DES fine-tuned model is shown in **(c)**. **(d)** Representation of per-residue training samples in the PDB data training set. 265

LIST OF TABLES

2.1 Top 15 DXXX-OPV3-XXXD molecules identified by the active learning protocol. Additional molecules previously studied in simulation and experiment are also shown for comparison. 46

3.1 Rank-ordered list of the 28 experimentally tested X_n-4T-X_n molecules sampled over the course of the active learning screen. 79

8.1 Comparison of our model trained on the PED dataset, labeled 'DiAMoNDBack (PED)', compared to the work of Yang and Gómez-Bombarelli [532], labeled as 'GenZProt'. Errors for bond and clash scores are reported as standard deviations over five generated samples averaged over frames, while diversity score errors are calculated using Jackknife resampling over generated samples according to Eqn. 8.2. The bolded values indicate superior performance if the results are outside of error. 255

8.2 Comparison of backmapping performance on PDB data and MD trajectories between the rules-based PULCHRA approach [530], GenZProt [532], and DiAMoNDBack trained on the PDB dataset, and the same PDB-trained model fine-tuned on MD trajectory data. The 'PDB overall' and 'DES overall' columns report aggregate metrics averaged over all samples and frames for the two data sets with additional columns reporting metrics on the worst- and best-performing systems for each dataset as determined by clash score. Errors for bond and clash scores are reported as standard deviations over five model-generated samples for our model and GenZProt, with no associated error for the deterministic backmapping of the rules-based approach. Errors in the diversity score are calculated using Jackknife resampling over generated samples according to Eqn. 8.2. 259

8.3 Application of backmapping to bespoke coarse-grained trajectories generated by Majewski et al. [162]. Comparisons are presented for three sequences spanning a range of lengths: 1FME (28 residues), PRB (47 residues), and A3D (73 residues). Errors represent standard deviations over generated samples for the bond and clash scores, with generative diversity measured here as the average pairwise RMSD between generated samples. 267

ACKNOWLEDGMENTS

I would like to express my deepest gratitude and appreciation to all those in my life who have supported and guided me throughout graduate school. Their invaluable contributions and support have made this research possible, and I am deeply indebted to them for their contributions.

First and foremost, I express my sincere appreciation to my supervisor, Prof. Andrew Ferguson, for their unwavering guidance, expertise, and patience. Their insightful feedback and constant encouragement have been pivotal in shaping this research and guiding my path throughout graduate school. I am incredibly thankful to Andy for all the opportunities, memories, and experiences. The time spent at the Institute for Pure and Applied Mathematics (IPAM) during the summer of 2019 was the most enlightening research experience I have had. IPAM provided me with the chance to meet and interact with leaders in the field and engendered many lasting friendships and collaborations that appear throughout this dissertation. I'm appreciative of the opportunity to interact daily with the great group of folks that were my colleagues in Andy's research group. Special thanks to Mike Jones with whom I worked closely on the transferable backmapping project, and who was a delightful and congenial desk mate. I'm also deeply grateful to Hythem Sidky who was a post-doc in the group during my first year and was incredibly helpful, patient, and foundational to my learning and early research in graduate school.

I'm thankful to all the remarkable individuals I have had the pleasure to work with and that have contributed to the research in this dissertation. I'm grateful to Marc Stieffenhofer, Nick Charron, and Moritz Hoffmann for their insights and patience throughout the slow progression of the temporally coherent backmapping work. I'm indebted to Sayak Panda who was our experimental collaborator from Johns Hopkins University on many of the π-conjugated peptide works. Sayak generously devoted his time to helping me better understand the experimental aspects and played an instrumental role in numerous projects

and research directions during my time in graduate school.

Finally, I'd like to thank my family and close friends for their invaluable camaraderie, support, and encouragement. My parents' unwavering support, both in graduate school and in life, helped me to pursue my aspirations of a career in science and provided me with the foundation in life to do so. Their courage in immigrating to America and building a new life from scratch for our family is a testament to their strength and determination. Their sacrifices and hard work have been a constant source of inspiration for me. My closest friends throughout graduate school Noah, Mason, and Sofi helped make the hard times easier and the easier times more fun. My partner, Alyssa, and our cat, Brisket, brought immeasurable joy and love into my life. Alyssa's unwavering belief in me, even during moments of self-doubt, has been a driving force behind my success. Her constant encouragement and understanding have lifted me up during challenging times, and her presence has made every achievement even more meaningful. Brisket, our adorable and mischievous feline companion, has provided endless moments of laughter and comfort throughout this journey. I am truly grateful for the love, support, and laughter my partner and our furry friend have brought into my life.

ABSTRACT

The fusion of artificial intelligence and machine learning techniques with molecular simulation presents a powerful paradigm for understanding and engineering functional molecules and materials. This thesis focuses on data-driven approaches for molecular design and simulation encompassing four themes of study: (i) discovery of self-assembling π-conjugated peptides, (ii) active learning for molecular design, (iii) machine learning-enabled enhanced sampling techniques for molecular simulation, and (iv) methods for generative restoration of atomistic detail into coarse-grained molecules and trajectories. First, the challenge of designing synthetic π-conjugated peptides comprising an aromatic core flanked by oligopeptide wings capable of self-assembling into elongated nanostructures with emergent optoelectronic properties is engaged. High-throughput virtual screening campaigns with experimental validation are performed to discover peptide sequences and π-core chemistries that best facilitate self-assembly and display the most promise as high-charge mobility organic semiconductors. Second, an active learning platform for molecular design is established that seeks to efficiently navigate large chemical spaces to discover high-performing candidates that optimize a target quality metric. This active learning platform is applied to both self-assembling π-conjugated peptides and the discovery of small organic compounds capable of selectively permeating cardiolipin-containing membranes, these molecules could serve as fluorescent dyes and diagnostic tools for quantifying mitochondrial cardiolipin content which is linked to various degenerative diseases. Third, a machine learning-enabled enhanced sampling technique to accelerate barrier crossings and rare event transitions in molecular simulation is developed. This method employs an adaptive sampling workflow that iterates between the data-driven estimation of the slowest evolving dynamical motions from biased simulation data and performing enhanced sampling simulations within these collective variables to promote configurational phase space sampling. Lastly, two deep learning approaches for generatively and non-deterministically reintroducing atomistic detail into coarse-grained

molecular representations are developed. One method is designed to maintain temporal coherence in the reconstructed trajectory using frame-by-frame conditioning, and the other provides an approach transferable to generic Cα protein traces. In conclusion, this dissertation showcases the application of data-driven approaches to various aspects of molecular design and simulation, ranging from the discovery of self-assembling π-conjugated peptides to active learning for molecular design, machine learning-enabled enhanced sampling techniques, and generative restoration of atomistic detail in coarse-grained molecules. These findings contribute to the advancement of computational methods for designing functional molecules and understanding complex molecular systems.

CHAPTER 1

INTRODUCTION

Artificial intelligence and machine learning techniques have revolutionized how people interact with data, facilitating the understanding of vast amounts of information and leading to superior decision-making capabilities. These tools are especially well-suited for tackling the large and complex datasets produced by molecular simulations. By applying these methods to such data sources valuable insights can be gained to deepen understanding of complex molecular systems. The research in this dissertation contributes to the advancement of strategies and methodologies in high-throughput virtual screening for molecular design and improving the efficiency of molecular simulations. Engaging this class of problems presents an opportunity to make a substantial impact across various scientific disciplines and industries. With the ever-increasing volumes of data generated through computation and experiment, traditional methods can struggle to extract meaningful patterns and knowledge. However, the integration of artificial intelligence and machine learning not only allows for navigating this data deluge but also empowers significant breakthroughs in drug discovery, materials science, and other scientific domains. Leveraging these powerful tools can expedite the development of life-saving drugs, uncover new materials with extraordinary properties, and contribute to groundbreaking scientific advancements. The research presented in this dissertation represents progress in harnessing the potential of these methods and has significant implications for the future of drug discovery and molecular design. By continuing to explore and refine these approaches, opportunities can arise with profound impacts on human health and scientific knowledge.

1.1 Self-assembling π-conjugated peptides

Molecules possessing both biocompatibility and electronic functionalities hold immense promise for the advancement of biocompatible electronic devices and interfaces in biological systems. Synthetic π-conjugated peptides are a particularly encouraging class of molecules consisting of oligopeptide wings surrounding a central aromatic π-core. What makes them truly remarkable is their spontaneous assembly into supramolecular nanoaggregates within aqueous and biologically relevant environments, revealing emergent electronic and optical properties [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Hierarchical self-assembly of π-conjugated peptides into elongated nanostructures is driven by a combination of intermolecular π-π stacking, hydrogen bonding, and van der Waals interactions [13, 14, 15, 16]. The versatility and customizability of π-conjugated peptides are evident in their ease of synthesis and their responsiveness to external stimuli such as pH, flow, light, salt concentration, and temperature [16, 17, 18, 19, 20, 21, 22]. Wielding control over the structural and functional properties of self-assembling π-conjugated peptides is also achieved through modulation of the molecular chemistry of the π-core and identity of the amino acids constituting the peptide wing [23, 24, 25, 26, 27]. As a result, they serve as highly adaptable molecular building blocks that facilitate the bottom-up self-assembly of biocompatible supramolecular networks that enable long-range charge transport and hold tremendous potential for applications such as peptide-based field-effect transistors, photovoltaic energy harvesting, and solar cells [4, 28, 29, 30, 24, 31, 32, 33, 34].

The tunability of π-conjugated peptides presents both a blessing and a curse. On one hand, it offers a wide range of molecular chemistries, allowing for diverse material properties and the ability to customize structure and function. On the other hand, this vast space of possibilities poses a challenge in identifying promising candidates. The series of works in Chapters 2, 3, and 5 engages this challenge of efficiently screening and identifying π-conjugated peptide chemistries that self-assemble into nanoaggregates with desirable in-

register stacking and alignment of the π-cores capable of fostering the emergence of optical and electronic properties.

Chapter 2 focuses on the DXXX-OPV3-XXXD class of π-conjugated peptides, where D is an Asp residue, OPV3 is an oligophenylene vinylene oligomer (1,4-distyrylbenzene), and XXX represents tripeptides consisting of the 20 naturally occurring amino acids. Head-to-tail symmetry of the molecules is maintained with the oligopeptide wings designed to be mirror-symmetric in the amino acid composition and the directionality from the N-terminus to the C-terminus such that each molecule has two C-termini. Furthermore, the terminal residues are specifically chosen as Asp, which gives each terminus of the molecule two carboxyl groups. This choice serves as a pH trigger for assembly: at high pH the four carboxyl groups lose their protons resulting in a (-4)e formal charge inducing Coulombic repulsion that discourages large-scale assembly, while at low pH the residues protonate neutralizing the molecule and facilitating self-assembling [35]. The complete DXXX-OPV3-XXXD family comprises 20^3=8000 members constituting a design space too large to engage with Edisonian trial-and-error testing via either simulation or experiment. We use coarse-grained simulations to evaluate the propensity of each molecule to form nanostructures with good in-register π-stacking by tracking early-stage assembly behavior [36, 37, 38]. While simulations provide a means for high-throughput virtual screening, the large design space presents a challenge for efficiently identifying promising candidates. We therefore perform active learning, and, more specifically, Bayesian optimization to navigate the chemical design space and identify the next most promising molecules to simulate based on all previously collected simulation data [39, 40, 41, 42, 43]. Using this active learning approach, we identify and computationally validate top-performing constitutes within the 8000-member DXXX-OPV3-XXXD family after simulating only 2.3% of all possible molecules. A *post-hoc* analysis of the molecular configurations using diffusion maps provides a mechanistic understanding of the assembly pathways and self-assembly behavior exposing practical precepts for molecular design.

In Chapter 3 we approach the same task of identifying π-conjugated with promising self-assembly behavior into pseudo-1D linear aggregates, this time engaging the more diverse X_n-quaterthiophene-X_n (X_n-4T-X_n) family. Here, the X_n peptide wing contains anywhere between $n = 1 - 5$ amino acids surrounding a fixed quaterthiophene π-core selected due to its demonstrated applications in organic electronic field-effect transistors and photovoltaics [44, 45]. Maintaining the mirror-symmetric character of the peptide wings together with the requirement the X_n wing must contain at least one acidic residue (*i.e.,* Asp, Glu) and no basic residues (*i.e.,* Arg, His, Lys) in order to wield control over the pH triggered assembly, this X_n-4T-X_n π-conjugated peptide family consists of 694,982 unique members. To efficiently screen this massive space we apply a hybrid-computational experimental active learning routine where high-throughput all-atom molecular dynamics simulation data is used to better inform the selection of experimental candidates for synthesis and testing. The core idea behind this approach involves on-the-fly training of models to establish sequence-property relationships using computational screening data which are then utilized to direct experimental screening in a virtuous feedback loop. The high-throughput computational screening uses all-atom molecular dynamics simulations to probe the structural morphology of the self-assembled nanoaggregates optimizing for both the presence of π-π stacking along with long fibril formation. In collaboration with our experimental colleagues Prof. John D. Tovar and Dr. Sayak Subhra Panda, the low-throughput experimental screen involves the synthesis and characterization of π-conjugated peptides employing UV-visible spectroscopy to measure the spectral shift between the molecularly dissolved and assembled states to quantify the extent of self-assembly [46, 47, 23]. After ultimately simulating only 1181 molecules (approximately 0.17% of the 694,982 molecules in the design space), and experimentally testing, a mere 28 molecules (approximately 0.004%), we discover and experimentally validate previously unknown high-performing candidate molecules that possess the ability to spontaneously assemble supramolecular aggregates exhibiting a high degree of

H-type character. Furthermore, we extract interpretable design principles linking peptide sequence with emergent supramolecular structure and properties. Lastly, *post-hoc* analysis of the simulation trajectories reveals an additional mechanistic understanding of prototypical assembly pathways in π-conjugated peptides.

Following work in Chapters 2 and 3 optimizing the identity of the peptide wings, in Chapter 5 we explore the influence of π-core chemistry and addition of a variable length alkyl spacer connecting the aromatic core to the peptide wings upon the self-assembled morphologies and emergent charge mobility within the resulting nanoaggregates. In collaboration with Prof. Andre Schleife and Yifan Yao, we utilize Marcus theory [48] to calculate the charge mobility in π-conjugated peptides, which are inherently disordered molecular systems. Since these peptide nanostructures lack a well-defined crystal structure, we employ molecular dynamics simulations to predict their molecular packing behavior into their self-assembled morphologies. Subsequently, density functional theory is used to determine the potential for charge mobility within these nanoaggregates adopting a Marcus hopping model [48], which predicts the transfer rate of electrons/holes between molecules. By combining density functional theory and molecular dynamics simulations in this manner, we are able to investigate the electronic behavior of self-assembled π-conjugated peptide nanostructures and identify promising molecules expected to exhibit favorable charge mobility. The chemical design space we screen is composed of π-conjugated peptides with a fixed VEVAG peptide wing [49, 27, 28] flanking one of six aromatic cores – (i) phenyl-thiophene-phenyl (PTP), (ii) bis(tetrafluorophenyl)thiophene (BFT), (iii) quarterthiophene (QTP), (iv) naphthalene-diimide (NDI), (v) perylene diimide (PDI), and (vi) benzo dithiophene (BDT) [50, 51, 52, 53, 28, 54, 23] – with the oligopeptide wing linked to the core by a variable length alkyl chain containing n=0, 1, 2, or 3 carbon atoms. These 24 candidate molecules are each subjected to our computational workflow to elucidate the combination of molecular building blocks that possess the most promise as high-charge mobility π-conjugated peptide

materials.

1.2 Active learning data-driven design

Active learning is a machine learning technique that aims to optimize the design process by iteratively selecting the most valuable data points for annotation, thereby reducing the overall labeling effort [55, 56, 57, 58]. It involves a feedback loop where the model leverages existing data to identify the next most promising new candidates to query. The newly labeled data is then used to update and improve the model, and the process repeats until some convergence criterion is met or the budget is exhausted. This approach is particularly useful in scenarios where labeling large amounts of data is expensive, time-consuming, or impractical. By actively updating the model iteration-by-iteration with the most informative samples, active learning aims to maximize the efficiency of the design process by accelerating and minimizing the burden of the data collection operation. For this reason, active learning and allied approaches have been swiftly gaining popularity within the materials discovery, engineering, and design sectors. These methodologies have found applications in various domains, for example, the experimental discovery of novel shape memory alloys [59], piezoelectrics [60], high glass transition polymers [39], the computational discovery of drugs [42], and magnetocaloric, superconducting, and thermoelectric materials [40].

Briefly, the active learning procedure employed throughout Chapters 2, 3, and 4 of this thesis can be understood on a high level as the following multi-step process:

> **Step 0:** Annotate an initial set of data. Here a fitness function must be defined that maps members of the design space to a quality metric that the active learning procedure will strive to optimize. This preliminary data set serves to seed the active learning providing a starting point to begin iterating on. Larger initial data sets are beneficial as they provide a more comprehensive representation of the data distribution and help establish a more reliable initial model. Typical methods for collecting this

initial data set include random sampling or employing a data-driven approach like k-means clustering [61, 62].

Step 1: Train a surrogate model using all existing data to predict the fitness of unlabeled candidates in the design space. A popular choice of surrogate model is Gaussian Process Regression (GPR) as this represents a flexible, Bayesian regression approach that comes with built-in uncertainty estimates [63, 64, 58, 55]. To expose the discrete members of the design space to the GPR model an embedding space, or more specifically a distance metric, must be adopted. Using data-driven dimensionality reduction techniques to map individual candidates to a low-dimensional representation where semantically similar members are embedded nearby one another promotes the transfer of information to new out-of-training points based on proximate training examples.

Step 2: Use the surrogate model predictions to select the next most informative candidates to label. This selection is achieved using Bayesian optimization which purposes an acquisition function to measure the potential improvement of a candidate point compared to the current best solution found so far, taking into account the surrogate model predictions and uncertainty estimates [65, 66, 55]. Effective acquisition functions, such as the Expected Improvement acquisition function [55, 67], balance exploration (sampling uncertain regions) and exploitation (focusing on promising areas near the current best solution) to efficiently search the design space and select the most informative new candidates for labeling. Batches of candidates can be selected for higher throughput and to make efficient use of available resources.

Step 3: Iterate through the surrogate model building, new candidate selection, and labeling process in Steps 1 and 2 until some stop criterion is met or resources are exhausted. Aspects that can suggest terminating active learning include stabilization of the GPR surrogate model posterior, converging GPR model performance, and round-

after-round failure to discover new best performers [68, 69, 70, 71, 72, 73, 74].

In Chapters 2 and 3 the active learning workflow is applied for the design π-conjugated peptides that assemble into linear nanoaggregates with good in-register stacking of the π-cores. In both these works, the design space encompasses different arrangements of amino acids forming the peptide wing. In Chapter 2 a single active learning loop is executed using coarse-grained molecular dynamics simulations as the measurement engine to navigate the 20^3=8000 member design space of the DXXX-OPV3-XXXD family of π-conjugated peptides. In Chapter 3, two parallel and asynchronous active learning loops are carried out: one high throughput computational loop using all-atom molecular dynamics simulations, and a lower throughput experimental loop performing actual synthesis and characterization of π-conjugated peptides. A multi-fidelity active learning approach is implemented that uses the voluminous, low-fidelity, computational data to better inform the surrogate model predictions for the scarce, but high-fidelity, experimental data. This process enabled the efficient exploration of peptide wings between $n = 1 - 5$ amino acids in lengths within the X_n-4T-X_n family of π-conjugated peptides comprising 694,982 unique members.

In Chapter 4, the active learning workflow is applied to the discovery of small organic compounds capable of selectively permeating cardiolipin-containing membranes. Mitochondria are organelles found in eukaryotic cells that generate ATP with the inner mitochondrial membrane composed of approximately 20% cardiolipin, a phospholipid comprising four acyl chains that is exclusively found in energy-generating membranes [75, 76]. Abnormalities in the cardiolipin composition of the inner mitochondrial membrane are associated with various diseases such as Barth syndrome, Tangier disease, heart failure, and neurodegeneration [76, 77]. A challenge in experimentally assaying cardiolipin content is the high degree of structural similarity between cardiolipin and other phospholipids, such as phosphatidylglycerol which is a precursor of which cardiolipin is essentially a dimer [78, 79, 80]. To engage this design problem a virtual screening approach is employed based on coarse-grained molecular

dynamics simulations and alchemical free energy calculations [81, 82, 83]. In collaboration with Prof. Tristan Bereau and Dr. Bernadette Mohr, active learning is performed to select molecules that maximize the partitioning free energy difference $\Delta\Delta G$ between cardiolipin and phosphatidylglycerol membrane environments. The design space is composed of coarse-grained molecules with molecular weights below ∼500 Da, enabling the effective representation of $\sim 10^{60}$ small organic compounds using a multiscale-based many-to-one mapping to only 124,327 unique coarse-grained topologies that maintain the pertinent physicochemical properties [84, 85]. After sampling only 0.42% of the design space, 242 compounds with up to 184% superior predicted thermodynamic selectivity for cardiolipin compared to the popular existing cardiolipin probe 10-*N*-nonyl acridine orange [86, 87] are discovered. In addition to generating novel molecular designs, this molecular discovery platform uncovers interpretable design principles that enhance understanding of the intricate physicochemical properties driving high cardiolipin selectivity.

1.3 Enhanced sampling in molecular simulations

The simulation of microscopic phenomena using molecular dynamics is limited by the requirement for short integration time steps in order to maintain numerical stability. As a result, there are practical limitations on the length of simulations that can be achieved. Despite advancements in computing hardware [88, 89] and simulation software [90, 91, 92] that have made trillion-atom simulations possible [93], the feasible time scales are still restricted to approximately microseconds on commodity hardware [94, 95] and milliseconds on specialized hardware [89, 96]. Many interesting processes, such as protein folding [97], protein-protein association [98], ligand binding [99], and molecular self-assembly [100], can extend far beyond these time scales. Consequently, computational strategies have been developed to facilitate the study of slow molecular motions and rare events. Collective variable enhanced sampling techniques are a category of methods designed to facilitate comprehensive exploration of con-

figurational space. These methods achieve this goal by introducing biasing potentials along predefined collective coordinates to facilitate barrier crossing, expedite exploration of phase space, and enhance the sampling of rare events [101]. Proper selection of collective variables is crucial, as they must be correlated with the transitions of interest and nearly orthogonal to the free energy barriers that separate different states (i.e., folded and unfolded states of a protein). Strategic choice of collective variables will therefore enhance the frequency of barrier crossing, enabling rare transitions and improving the overall sampling of configurational phase space by surmounting latent free energy barriers that otherwise frustrate the unbiased dynamics.

In Chapter 7, a new machine learning-enabled collective variable enhanced sampling technique called the Girsanov Reweighting Enhanced Sampling Technique (GREST) is developed. The manual definition of good collective variables is challenging for large and complex molecular systems where intuition is frequently limited. Data-driven approaches that employ machine learning to automatically select good collective variables can help address this problem by exploiting the feature that cooperative couplings between atomic degrees of freedom leads to emergent low-dimensionality in molecular systems. The physical manifestation of this phenomenon is that the long time-scale dynamics can be characterized by the leading slow modes of a system (e.g., large-scale protein folding/unfolding) which renders the fast degrees of freedom (e.g., side chain rotations) as effectively noise orthogonal to the slow modes [102, 103, 104, 105, 106, 107, 104]. The discovery of good collective variables for enhanced sampling can therefore be framed as a task in dimensionality reduction, with the goal of identifying variables maximally correlated with the slowest evolving degrees of freedom. While a variety of methods have been developed for these purposes, the GREST approach makes the important distinction of learning *slow* collective variables rather than *high variance* collective variables. By explicitly treating these slow collective variables as inherently dynamical quantities, GREST makes effective use of both thermodynamic and dynamical

Girsanov reweighting corrections for rigorous estimation of slow collective variables from discontinuous and biased simulation data [108, 109, 110, 111, 112]. By interleaving rounds of data-driven slow collective variable discovery and enhanced sampling in these learned coordinates, GREST effectively drives sampling of all thermally accessible metastable states starting from zero prior knowledge of the system. This approach is demonstrated in Chapter 7 through the application on a toy 1D 4-well potential, a simple biomolecular system alanine dipeptide, and the Trp-Leu-Ala-Leu-Leu (WLALL) pentapeptide.

1.4 Backmapping for restoring atomistic detail into coarse-grained molecules

Similar to how enhanced sampling techniques can help overcome the limitations of molecular dynamics simulations, coarse-graining is a multiscale approach that enables access to spatiotemporal scales entirely out of reach of conventional atomistic molecular dynamics simulations [113, 114, 115, 116, 117, 118]. Coarse-graining involves grouping atoms into "beads" or "superatoms" to retain essential properties from the original atomistic system [119, 120, 121]. This approach allows for studying larger and more intricate systems that would be impractical to analyze in atomistic simulations due to computational constraints. One consequence of this reduction in resolution is the smoothing of the underlying free energy surface, which facilitates the exploration of conformational motions that would otherwise be frustrated in the rugged atomistic landscape [122, 123, 116, 124, 125]. These advantages have led to the increasing popularity and widespread use of coarse-grained models, especially in simulations involving proteins, polymers, molecular self-assembly, membranes, and high-throughput screening [126, 127, 128, 129, 130, 131, 132, 133, 134, 135]. With the primary concession of coarse-graining coming at the cost of losing atomistic detail, backmapping techniques seek to reintroduce atomistic representations conditioned on the coarse-grained structure. While rules-based techniques using heuristics and/or atomistic

fragments have been developed to achieve this task, these are unsatisfying solutions as they frequently rely on expensive *post-hoc* structural refinement and are commonly deterministic in nature failing to capture the multitude of atomic structures associated with a given coarse-grained representation [115, 136, 137, 138, 139]. Data-driven approaches are more appealing backmapping solutions as they can be trained to recapitulate the conformational diversity of atomic structures in a generative fashion. These approaches offer more scalability with higher throughput as they are typically designed to produce well-equilibrated structures and are capable of non-deterministic backmapping, where a variety of feasible and novel atomistic structures can be generated when backmapping any individual coarse-grained configuration [140, 141, 142, 143]. Two backmapping methods are developed in this thesis that leverages state-of-the-art machine learning models treating this as a super-resolution task converting from coarse-grained to atomistic representations.

In Chapter 6, a new temporally coherent backmapping technique is developed that aims at backmapping trajectories generating both well-equilibrated molecular structures for each frame and achieving temporal coherence between frames. Existing data-driven approaches are designed to backmap each frame independently [140, 141, 144, 143, 145, 142], explicitly ignoring temporal information from preceding frames that can be useful for improving reconstruction quality and enabling the recovery of dynamical properties. This task is accomplished by training a conditional variational autoencoder (cVAE) [146] to up-scale coarse-grained configurations by incorporating information from both the current coarse-grained structures and previously decoded atomistic structures in the trajectory. The cVAE model adopts a convolutional neural network architecture, which has proven highly successful in the field of computer vision [147, 148, 149, 150, 151], to process molecules represented as voxelized particle densities. By capturing the diverse range of possible atomistic structures associated with each coarse-grained configuration, the cVAE model enables the generation of reconstructed trajectories that are not mere replicas of the training data. This approach

is applied to two exemplar biomolecular systems alanine dipeptide and the miniprotein chignolin, demonstrating the accurate retrieval of structural, thermodynamic, and kinetic properties. Additionally, the backmapping approach exhibits good generalization capabilities, performing well when applied to coarse-grained data originating from unseen and approximate coarse-grained force fields [152].

In Chapter 8, a chemically transferable backmapping approach capable of restoring atomistic detail into generic Cα protein traces is developed. Most existing data-driven backmapping techniques, including the approach developed in Chapter 6, rely on training bespoke models for each system intended for backmapping [143, 142, 153]. This requirement poses a substantial bottleneck to the broader adoption of these backmapping tools, necessitating practitioners to continuously curate training data for each new application. To achieve transferability, an autoregressive backmapping model is designed that predicts the protein structure residue-by-residue. By stipulating that each residue inference task is dependent on the local environment of the target residue being reconstructed, the intrinsically local nature of this approach suggests its potential for transferability to protein traces of arbitrary length and composition (given adequate training data). The model architecture is based on the recently popularized class of generative models the Denoising Diffusion Probabilistic Model (DDPM) [154, 155] and directly predicts the Cartesian coordinates of the atoms belonging to each residue aligned to a canonical reference frame, avoiding the need for costly data augmentations that would otherwise be required to learn sensitivity to rigid body rotations and translations [153, 156]. After training a generic model on a dataset of 65k+ structures curated from the Protein Data Bank (PDB) [157, 158, 159, 160], the model effectively generalizes to unseen PDB structures, molecular dynamics trajectories of fast-folding miniproteins [161], and bespoke coarse-grained simulation data [162]. Fine-tuning this generic PDB-trained model with atomistic training data also demonstrates improved performance by better calibrating the model to the specific task. By training on a large corpus of PDB

structures with the capability for task-specific fine-tuning, this approach provides a data-driven framework and practical backmapping tool for the realization of diverse atomistic structures from arbitrarily large Cα protein traces.

1.5 Publications

The chapters of this thesis are based on the following peer-reviewed publications (* designates equal contribution):

1. **Shmilovich, K.**, Mansbach, R. A., Sidky, H., Dunne, O. E., Panda, S. S., Tovar, J. D., & Ferguson, A. L. (2020). "Discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation." *The Journal of Physical Chemistry B*, 124(19), 3873-3891. [131]

2. **Shmilovich, K.**, Panda, S. S., Stouffer, A., Tovar, J. D., & Ferguson, A. L. (2022). "Hybrid computational–experimental data-driven design of self-assembling π-conjugated peptides." *Digital Discovery*, 1(4), 448-462. [163]

3. Mohr, B.*, **Shmilovich, K.***, Kleinwächter, I. S., Schneider, D., Ferguson, A. L., & Bereau, T. (2022). "Data-driven discovery of cardiolipin-selective small molecules by computational active learning." *Chemical Science*, 13(16), 4498-4511. [132]

4. **Shmilovich, K.***, Yao, Y.*, Tovar, J. D., Katz, H. E., Schleife, A., & Ferguson, A. L. (2022). "Computational discovery of high charge mobility self-assembling π-conjugated peptides." *Molecular Systems Design & Engineering*, 7(5), 447-459. [164]

5. **Shmilovich, K.**, Stieffenhofer, M., Charron, N. E., & Hoffmann, M. (2022). "Temporally Coherent Backmapping of Molecular Trajectories From Coarse-Grained to Atomistic Resolution." *The Journal of Physical Chemistry A*, 126(48), 9124-9139. [153]

6. **Shmilovich, K.**, & Ferguson, A. L. (2023). "Girsanov Reweighting Enhanced Sampling Technique (GREST): On-the-Fly Data-Driven Discovery of and Enhanced Sampling in Slow Collective Variables." *The Journal of Physical Chemistry A,* 127(15), 3497-3517. [165]

CHAPTER 2

DISCOVERY OF SELF-ASSEMBLING π-CONJUGATED PEPTIDES BY ACTIVE LEARNING-DIRECTED COARSE-GRAINED MOLECULAR SIMULATION

2.1 Abstract

Electronically-active organic molecules have demonstrated great promise as novel soft materials for energy harvesting and transport. Self-assembled nanoaggregates formed from π-conjugated oligopeptides composed of an aromatic core flanked by oligopeptide wings offer emergent optoelectronic properties within a water soluble and biocompatible substrate. Nanoaggregate properties can be controlled by tuning core chemistry and peptide composition, but the sequence-structure-function relations remain poorly characterized. In this work, we employ coarse-grained molecular dynamics simulations within an active learning protocol employing deep representational learning and Bayesian optimization to efficiently identify molecules capable of assembling pseudo-1D nanoaggregates with good stacking of the electronically-active π-cores. We consider the DXXX-OPV3-XXXD oligopeptide family, where D is an Asp residue and OPV3 is an oligophenylene vinylene oligomer (1,4-distyrylbenzene), to identify the top performing XXX tripeptides within all 20^3 = 8,000 possible sequences. By direct simulation of only 2.3% of this space, we identify molecules predicted to exhibit superior assembly relative to those reported in prior work. Spectral clustering of the top candidates reveals new design rules governing assembly. This work establishes new understanding of DXXX-OPV3-XXXD assembly, identifies promising new candidates for experimental testing, and presents a computational design platform that can be generically extended to other peptide-based and peptide-like systems.

This chapter is based on the work reported in: **Shmilovich, K.**, Mansbach, R. A., Sidky, H., Dunne, O. E., Panda, S. S., Tovar, J. D., & Ferguson, A. L. (2020). "Discovery of

self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation." *The Journal of Physical Chemistry B*, 124(19), 3873-3891.

2.2 Introduction

Self-assembling π-conjugated peptides possessing a π-core flanked by peptide wings have emerged as a versatile building block for the bottom-up fabrication of bio-compatible nanoaggregates with engineered optoelectronic properties. Overlaps between π-orbitals in neighboring aromatic cores within supramolecular assemblies lead to the emergence of optical and electronic properties including fluorescence, electron/hole transport, and exciton splitting, and the flanking oligopeptide wings provide the capacity to operate in and interact with biological environments [166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 31, 176]. These peptidic materials have proven readily synthesizable and responsive to external control mediated by pH, flow, light, salt concentration, and temperature [177, 178, 179, 180, 36, 181, 182, 183], and have found a host of potential applications in the context of photovoltaic power generation, energy harvesting, and as organic transistors [184, 172, 185, 174, 186, 187, 188, 189, 35]. The structural and functional properties of the self-assembled nanoaggregates are governed by the molecular chemistry of the π-core and the amino acid sequence of the peptide wings.

The Asp-X-X-X-(oligophenylenevinylene)$_3$-X-X-X-Asp (DXXX-OPV3-XXXD) family represents one class of synthetic π-conjugated peptides possessing an oligophenylenevinylene π core, terminal Asp residues, and amino acid side chains, where X represents one of the 20 natural amino acids (Fig. 2.1a). To assure the molecules are head-to-tail invariant, the oligopeptide wings are constrained to be mirror-symmetric both in the identity of the amino acids and the N-to-C directionality, such that each molecule possesses two C-termini. The terminal residues are constrained to be Asp to endow each terminus of the molecule with two carboxyl groups and provide a pH trigger for assembly: at pH$>$5 the four carboxyls are deprotonated endowing the molecule with a (-4)e formal charge and disfavoring large scale assembly, but

at pH<1 the residues protonate, the molecule becomes neutral, and large-scale aggregation proceeds [35]. The DXXX-OPV3-XXXD family has attracted considerable experimental and computational attention in recent years due to their demonstrated capability to assemble into pseudo-1D optically and electronically active nanoaggregates whose structure and properties can be tuned through selection of the X residues [190, 37, 36, 188, 185, 191]. Assembly in aqueous solvent under acidic conditions is driven by hydrophobic, π-π stacking, and hydrogen bonding interactions [188, 190, 192, 35]. The assembly of elongated peptides into linear aggregates with in-register stacking and alignment of the π-cores favors π orbital overlap, electronic delocalization along the backbone of the nanoaggragate, and the emergence of optical and electronic functionality such as well-defined absorption and emission spectra, HOMO/LUMO gaps, electron/hole conductivity, and exciton splitting capabilities (Fig. 2.1b,c) [190, 185, 25, 1, 2, 27].

The complete DXXX-OPV3-XXXD family comprises 20^3 = 8,000 members corresponding to all possible permutations of the 20 natural amino acids within the unspecified XXX triplet. This vast size of this chemical space is both a blessing – the large palette of molecular chemistries provides enormous versatility in materials properties and the opportunity to tailor structure and function – and a curse – it is a challenge to identify promising candidates within this enormous space. Identifying the candidates capable of self-assembling into well-ordered optoelectronic nanoaggregates and divining the design precepts dictating the mechanism is a key goal in realizing these peptides as novel biocompatible optoelectronic materials.

Edisonian traversal of the large chemical space of DXXX-OPV3-XXXD molecules by trial-and-improvement experimentation is essentially intractable due to the high time and labor costs associated with peptide synthesis and testing. To date, no more than 13 members of the family have been experimentally synthesized and tested [185]. Molecular simulation offers an alternative means to perform high-throughput virtual screening of chemical space

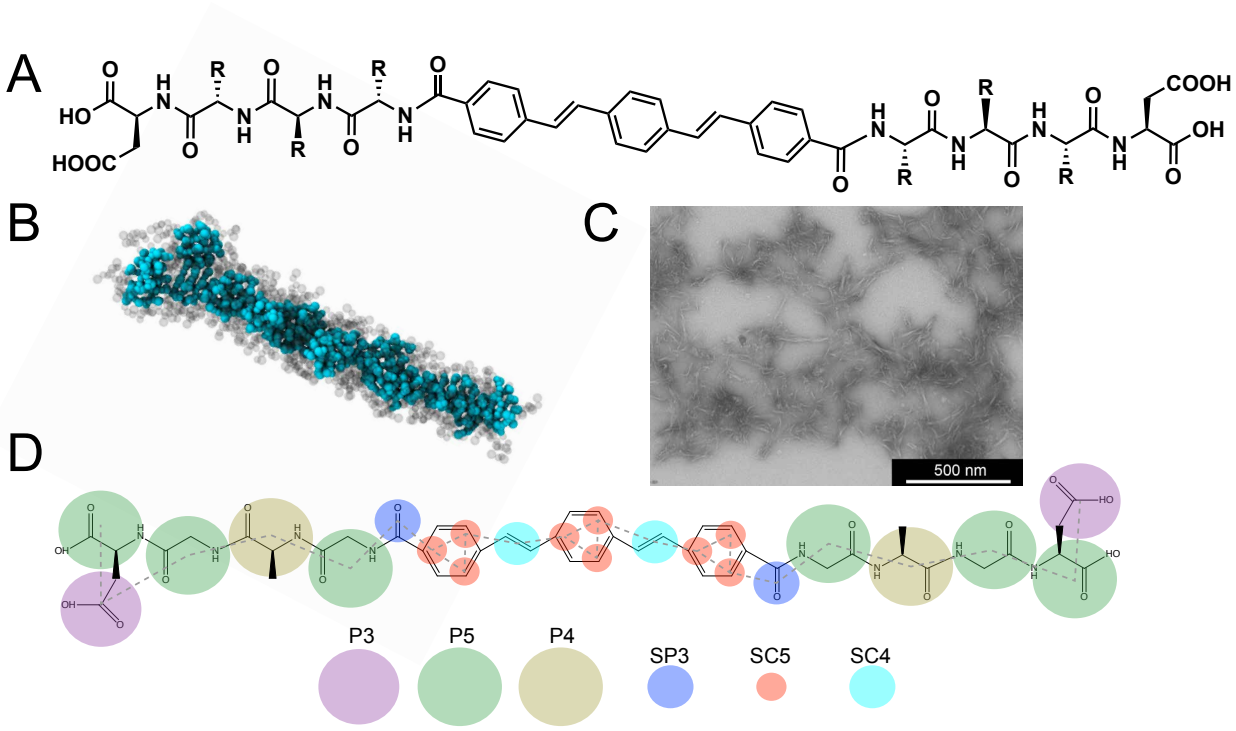


Figure 2.1: The DXXX-OPV3-XXXD system. (a) Chemical structure of the prototypical DXXX-OPV3-XXXD peptide monomer. The oligophenylenevinylene π core (OPV3) is flanked by oligopeptide wings (DXXX) that are mirror symmetric such that the identity of the amino acids is inverted and the molecule possesses two C-termini. The X residues are selected from the 20 natural amino acids such that the family comprises $20^3 = 8{,}000$ distinct molecules. (b) Molecular simulation snapshot of a self-assembled pseudo-1D nanoaggregate spontaneously formed by the spontaneous association of DVAA-OPV3-VAAD peptide monomers into a linear stack. Good stacking between the π-cores (colored blue) favors π orbital overlap, electronic delocalization along the backbone of the nanoaggragate, and the emergence of optical and electronic functionality. (c) Experimental transmission electron microscopy (TEM) image of self-assembled fibrils formed by DFFG-OPV3-GFFD peptides in an acidic environment. Reprinted with permission from Ref. [185] . Copyright (2014) American Chemical Society. (d) Illustration of the mapping from the DGAG-OPV3-GAGD all-atom structure to the coarse-grained representation at which the simulations in this work are conducted. The coarse-grained beads corresponding to groupings of neighboring atoms are labeled according to the Martini model employed in this work [193, 194, 195].

to identify the most promising candidates for experimental testing. Since assembly proceeds on length scales of tens of nanometers and microsecond time scales, this has motivated the development of coarse-grained models explicitly parameterized against all-atom molecular simulations [36, 37, 38] (Fig. 2.1d). These models integrate out the electronic and atomistic

degrees of freedom by lumping together small numbers of atoms into beads in order to furnish a molecular model that offers a judicious compromise between chemical realism and the computational efficiency required to directly simulate peptide assembly [37]. Exhaustive simulation of all 8,000 candidates within the DXXX-OPV3-XXXD family remains, however, computationally expensive. As we shall demonstrate, however, doing so is unnecessary to parameterize a reliable surrogate model of peptide function and identify and validate the most promising candidates within the family.

Chemical intuition is extremely valuable in guiding the computational search through chemical space, but it can perform poorly in the limits of data paucity, where there are too few examples to infer patterns, and data abundance, where there are too many examples to parse effectively. Further, inherent preconceptions and biases may push the search away from potentially profitable regions of chemical space and overlook patterns in the high-dimensional data that may reveal important determinants of molecular performance. Active learning (a.k.a. sequential learning, optimal experimental design), and more specifically, Bayesian optimization, presents a systematic approach to guide traversal of chemical space by using information on all measurements conducted to date to inform the "next-best" measurement to conduct [39, 40, 41, 42, 43]. In this manner, active learning predicts an optimal sequence in which to consider the molecular candidates in order to identify the optimal ones with minimal data collection effort. For this reason, active learning and allied approaches have been rapidly gaining traction in the materials discovery, engineering, and design communities, with these approaches being deployed, for example, in the experimental discovery of novel shape memory alloys [59], piezoelectrics [60], high glass transition polymers [39], the computational discovery of drugs [42], and magnetocaloric, superconducting, and thermoelectric materials [40].

Our primary goal is to efficiently identify members of the DXXX-OPV3-XXXD family that exhibit self-assembly into desired pseudo-1D nanoaggregates with good overlap be-

tween the π-conjugated cores and are thus most promising in displaying emergent optical and electronic functionality. We adopt a coarse-grained bead-level molecular simulation model as the engine for our high-throughput virtual screen and couple this with a deep learning-enabled active learning protocol to guide optimal traversal of chemical space. We identify and computationally validate the top performing constituents of the 8,000-member DXXX-OPV3-XXXD family after simulating only 2.3% of all possible molecules. This represents a massive saving over exhaustive sampling enabled by active learning. The absence of any introduced human bias within the active learning protocol also proved to be valuable in identifying high-performing candidates incorporating methionine residues that were not previously considered. A *post hoc* analysis of the observed assembly pathways provides supporting mechanistic understanding of the self-assembly behavior and exposes practical precepts for molecular design. The rank ordered list of DXXX-OPV3-XXXD molecules produced by our computational analysis provides a useful filtration of the design space with the top-performing candidates offering a massively reduced candidate space for experimental synthesis and testing.

2.3 Methods

2.3.1 Molecular dynamics simulation

The DXXX-OPV3-XXXD peptides were modeled using a previously-developed coarse-grained potential based on the Martini potential [37, 36]. Martini is a popular coarse-grained potential that lumps approximately four heavy atoms into each coarse-grained bead, has demonstrated great successes in modeling peptides, proteins, lipids, and carbohydrates [194, 195, 196, 197, 198, 199, 200], and offers a good compromise between chemical specificity and the computational efficiency necessary to probe the formation of large peptide aggregates. The potential was initially developed for DFAG-OPV3-GAFD by refitting the

native Martini parameters for bonded interactions against all-atom simulation data [37]. This bottom-up reparameterization of the bonded interactions greatly improved agreement between the coarse-grained and all-atom distribution functions, potentials of mean force (PMF) for monomer stretching and dimerization, and time-averaged contact maps [37]. We generalize this model to the complete DXXX-OPV3-XXXD family by maintaining the same parameterization of the bonds, angles, and backbone dihedrals within the OPV3 core and employing default Martini parameters for the amino acid side chains and all non-bonded interactions [193, 195]. An illustration of the all-atom to coarse-grained bead-level mapping for DGAG-OPV3-GAGD is provided in Fig. 2.1d. Calculation and comparison of the translational diffusion constants for the all-atom and coarse-grained models of DFAG-OPV3-GAFD showed these to be in agreement within error bars, indicating no significant discrepancy in the (translational) dynamical time scales between the two models and that no time scale corrections to the coarse grained calculations are required.

Coarse-grained molecular dynamics simulations of peptide assembly were conducted using the Gromacs 2018.6 simulation suite [201]. Initial system configurations for each DXXX-OPV3-XXXD considered were generated by randomly inserting 96 peptides into a $16.2 \times 16.2 \times 16.2$ nm^3 cubic simulation box with 3D periodic boundary conditions, corresponding to a concentration of approximately 35 mM. The amino acid residues are prepared in protonation states corresponding to pH 1 to mimic pH-triggered experimental assembly under acidic conditions. The coarse-grained peptides were then solvated in water to a density of 1.0 g/cm^3 of water using the Martini non-polarizable water model [193]. Steepest descent energy minimization was performed to eliminate high energy overlaps by removing forces greater than 1,000 kJ/mol.nm. Initial particle velocities were assigned from a Maxwell-Boltzmann distribution at 298 K. All simulations were conducted in the NPT ensemble at 298 K and 1 bar using a velocity-rescaling thermostat [202] and Parrinello-Rahman barostat [203]. Equations of motion were numerically integrated using the leap-frog algorithm with a 5 fs time step [204]

and bond lengths fixed using the LINCS algorithm [205]. Lennard Jones interactions were smoothly shifted to zero at 1.1 nm and reaction-field electrostatics were employed using a relative electrostatic screening constant of 15 appropriate for the non-polarizable water model [194]. An initial 100 ps equilibration run was conducted, after which time the temperature, pressure, density, and energy all stabilized. This was followed by a 3 μs production run, after which time the structural evolution of the system as measured by graphical analysis of the self-assembled aggregate (see Section 2.3.2) was stationary in time. Simulation snapshots were harvested for analysis every 50 ps over the course of the production run. Calculations were predominantly conducted on single NVIDIA GeForce RTX 2080 Ti cards and achieved execution speeds of $\sim$1.45 μs/day.

2.3.2 Active learning peptide discovery

An active learning protocol is employed to direct a principled traversal of the DXXX-OPV3-XXXD candidate space and minimize the number of coarse-grained simulations required to discover the highest-performing candidates [39, 40, 41]. The fundamental challenge is that evaluating the quality of each peptide by direct simulation is expensive, so we wish to identify the best peptide candidates in the fewest number of simulations. The procedure we employ is in large part inspired by and adapted from a pioneering deep representational active learning approach for molecular drug discovery developed by Gomez-Bombarelli et al. [42]. Our approach comprises four main steps and is illustrated schematically as an iterative active learning cycle in Fig. 2.2. The coarse-grained molecular simulation engine representing our measurement function within the protocol is described in Section 2.3.1, and we define our fitness function in Section 2.3.2. Appreciating that some of the more technical machine learning concepts may be foreign to some readers in the molecular modeling community, we expose these steps in the protocol in some detail along with their specific adaptations to our molecular system, but those readers familiar with variational autoencoders, Gaussian

process regression, and Bayesian optimization may feel free to skim over Sections 2.3.2-2.3.2. All codes are developed in Python 3 making use of the Scikit-learn [206], NumPy [207], Keras [208], and ORCA [209] libraries. Jupyter notebooks implementing our methods are hosted on GitHub (`https://github.com/KirillShmilovich/ActiveLearningCG`).

Step 1: Definition of fitness function for self-assembled aggregates

To perform active learning discovery in our predefined chemical space we define a scalar-valued fitness function $f : f_i = f(\text{DXXX-OPV3-XXXD}_i)$ that assigns a quality to each peptide in terms of its capacity to self-assemble into pseudo-1D nanoaggregates. Linear aggregates with good overlap between the π-conjugated cores are most promising in displaying emergent optical and electronic functionality and therefore anticipated to possess the most desirable materials properties. We have previously employed DFT calculations to make direct predictions of optoelectronic properties, but the high computational cost of these calculations limit them to aggregates of small numbers of peptides (dimers and trimers) and require omission of the flanking amino acid residues and solvent [190]. As such, these calculations are poorly suited to high-throughput virtual screening for large-scale aggregation behavior. Consequently, we define and optimize a structural measure of assembly quality in our coarse-grained molecular simulations as a proxy for optical and electronic functionality. This simplification massively expedites sampling in the full chemical space and provides a means to coarsely screen chemical space and focus a subsequent experimental search on the most promising candidates. Alternatively, this computational screen can be viewed as a preliminary filtration within the coarsest level of a nested hierarchy of increasingly expensive all-atom and/or electronic structure calculations.

In order to specify f we define a geometric criterion by which a pair of peptides are considered to form part of the same pseudo-1D nanoaggregate. To do so, we adopt a distance metric that we have previously employed to define clustering in DFAG-OPV3-GAFD

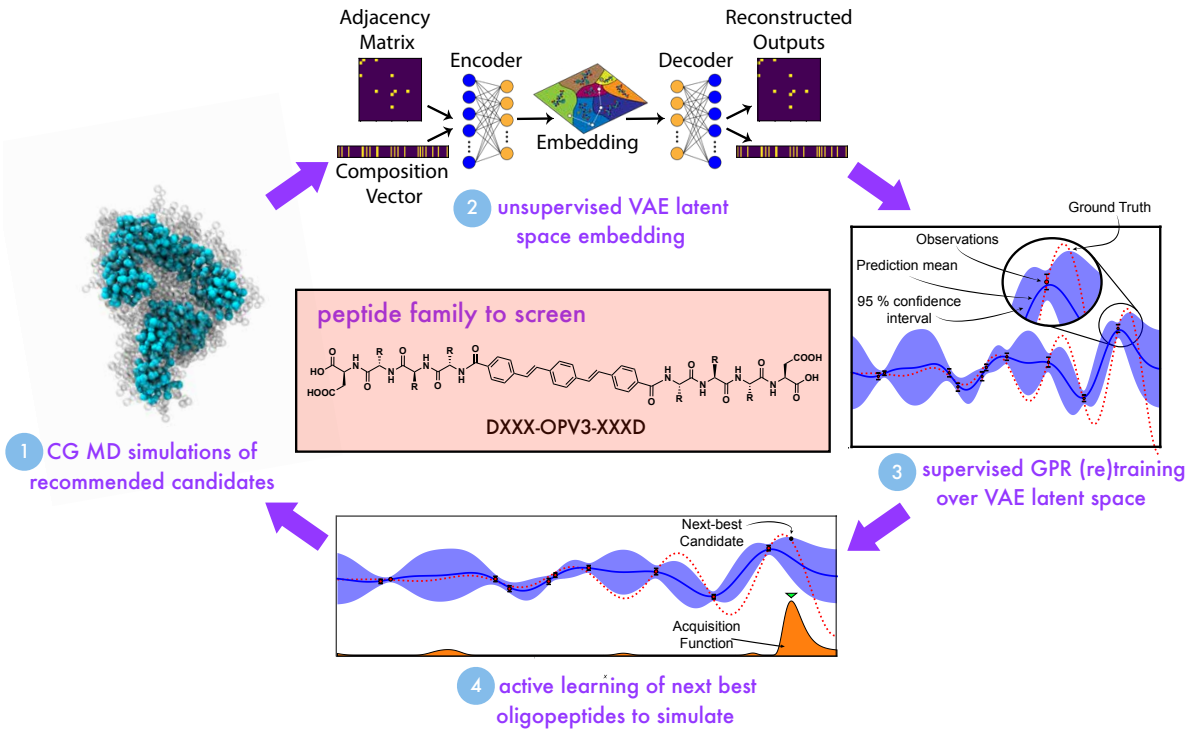


Figure 2.2: Active learning cycle for the data-driven discovery of optimally self-assembling DXXX-OPV3-XXXD peptides. The cycle contains four components. (1) Coarse-grained molecular simulations are performed on selected DXXX-OPV3-XXXD candidates and the quality of the self-assembled aggregates formed by molecule i measured according to a scalar fitness f_i. (2) The DXXX-OPV3-XXXD family is projected from the high-dimensional chemical space of molecular structures into a low-dimensional latent space embedding $E : \text{DXXX-OPV3-XXXD}_i \rightarrow \mathbf{z}_i \in \mathbb{R}^d$ using a variational autoencoder (VAE). The dimensionality of the latent space is optimized during each cycle. (3) A Gaussian process regression (GPR) model is constructed over the VAE latent space linking the latent space coordinates of each DXXX-OPV3-XXXD family member to the scalar fitness function measuring the quality of their self-assembled aggregates $f : \hat{f}_i = f(\mathbf{z}_i) = (f \circ E)(\text{DXXX-OPV3-XXXD}_i)$. The GPR mapping f is retrained each cycle over all DXXX-OPV3-XXXD candidates that have been simulated to date and for which measures of the fitness function $f_i,\ i \in \textit{sampled}$ is available, and is then used to predict the fitness of unsimulated candidates $\hat{f}_j,\ j \in \textit{unsampled}$. (4) The predicted means and uncertainties for $\hat{f}_j,\ j \in \textit{unsampled}$ furnished by the GPR surrogate model are combined within an active learning acquisition function to identify the "next-best" candidates for which to perform coarse-grained molecular simulations to drive sampling towards the most promising candidates. The loop is cycled until the GPR surrogate model no longer changes with additional data collection and can then be used to reliably identify the top candidates for computational validation.

assembly [37, 36] and asphaltene aggregation [210]. This so-called "optical distance" metric is defined as the minimum center of mass distance between aromatic cores in molecule a and b,

$$d_{ab}^{\text{optical}} = \min_{i \in core(a), j \in core(b)} r_{ij}, \tag{2.1}$$

where r_{ij} is the intermolecular center-of-mass distance between the aromatic rings i and j within the OPV3 cores, and the minimization proceeds over the three aromatic rings $i \in core(a)$ in molecule a, and the three aromatic rings $j \in core(b)$ in molecule b. Pairs of molecules a and b which satisfy $d_{ab}^{\text{optical}} < r_{\text{cut}} = 0.7$ nm are considered to reside within the same cluster [37, 36]. The cutoff $r_{\text{cut}} = 0.7$ was tuned to the mean of the distribution of d_{ab}^{optical} collected over DFAG-OPV3-GAFD peptide dimers with good in-register stacking of the OPV3 cores [37]. In contrast with other choices of peptide clustering metrics based, for example, on the overall center-of-mass or proximity of the peptide wings, the optical metric assures close intermolecular proximity of at least one pair of OPV3 aromatic rings in a pair of associated peptides. This close association promotes π electron overlap, electron delocalization, and the emergence of optoelectronic function, and it is for this reason that this metric is termed the optical distance metric [37, 38, 36].

Given this definition, a natural choice for the fitness f_i of molecule i is the number of such optical contacts in a self-assembled aggregate, since maximizing this value will promote electronic delocalization and the emergence of optoelectronic functionality. We evaluate the fitness function by representing the molecular system as a dynamically-evolving interaction graph in which the peptides compose the vertices $V = \{v_1, v_2, \ldots, v_N\}$ and the edges $E = \{e_{1,2}, e_{1,3}, ..., e_{95,96}\}$ are assigned between pairs of vertices v_a and v_b if $d_{ab}^{\text{optical}} < r_{\text{cut}} = 0.7$. An illustration of the evolution of the interaction graph over the course of a 3 μs simulation of DDAI-OPV3-IADD assembly is presented in Fig. 2.3. The number of vertices $|V| = 96$ is fixed by the number of peptides in the system. Maximization of the number of edges $|E(t)|$ at time t is therefore equivalent to maximizing the mean degree of each vertex in the graph

$\kappa(t) = \frac{2|E(t)|}{|V|}$. As such, we adopt as our fitness function,

$$f_i = \overline{\kappa(t; \text{DXXX-OPV3-XXXD}_i)} = \frac{2\overline{|E(t; \text{DXXX-OPV3-XXXD}_i)|}}{|V|}, \tag{2.2}$$

where the time average denoted by the overbar is performed over the terminal 50 ns of the 3 μs production run. Standard errors in the mean are estimated by block averaging the terminal 50 ns in five contiguous 10 ns blocks.

A potential criticism of the fitness function is that κ achieves a maximum for all-to-all connectivity of the graph, and its maximization would therefore appear to not necessarily favor pseudo-1D linear stacks. Mathematically this is true, but there are strong physical limitations on the maximum attainable value of κ since the excluded volume of the π cores allow then to form optical associations with a limited number of partners. The largest value observed in all of our calculations is $\kappa = 6.07$ (cf. Table 2.1), and visual inspection of the terminal aggregates confirms that κ is positively correlated with the formation of elongated pseudo-1D nanoaggregates similar to those illustrated in the $t = 3{,}000$ ns panel of Fig. 2.3.

Step 2: Learning latent space embeddings using variational autoencoders

In Step 3 (Section 2.3.2) we describe our training of a Gaussian process regression (GPR) surrogate model to predict the fitness of candidate molecules that have not been simulated based on those that have. The predictions of this model are then used to perform active learning. We experimented with constructing the GPR directly over the chemical space of DXXX-OPV3-XXXD molecules by measuring pairwise distances between the XXX tripeptides using BLOSUM substitution matrices [213], but following Gomez-Bombarelli et al. [42], we found this approach to yield inferior surrogate models to those constructed over data-driven low-dimensional embeddings of the molecules generated using a variational autoencoder (VAE) [214]. The low-dimensional VAE latent spaces also conveys advantages in

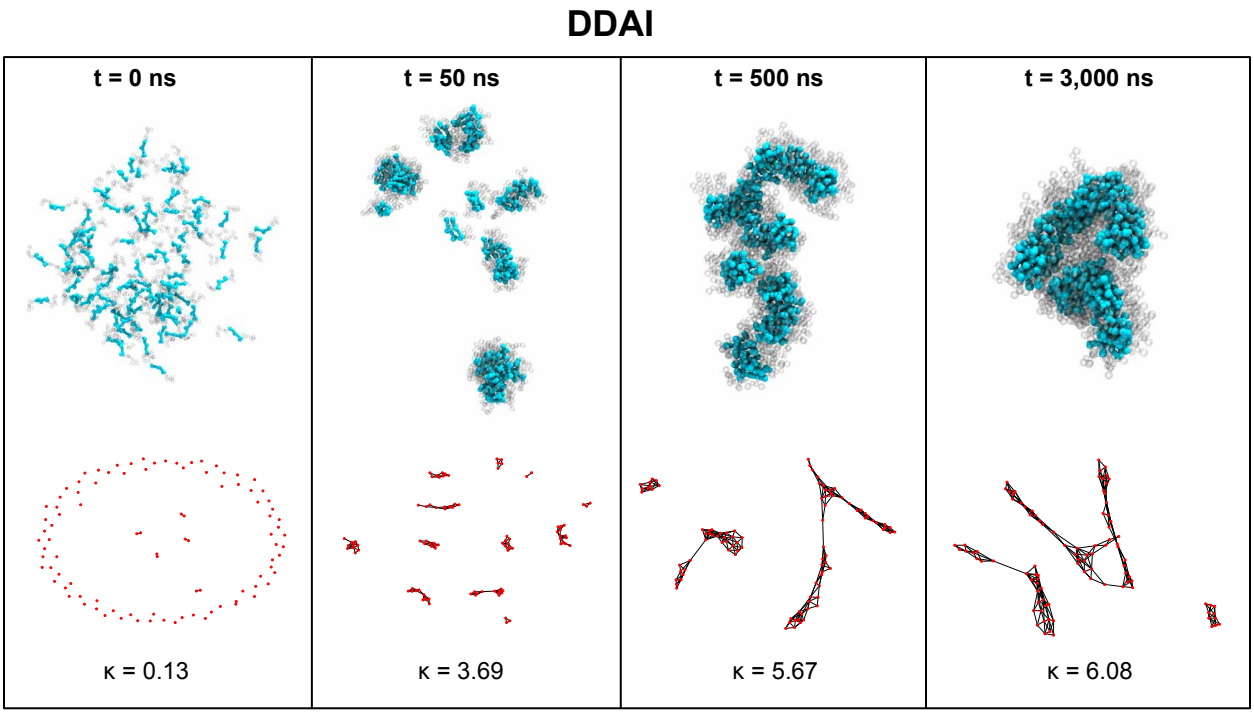


Figure 2.3: Dynamical evolution of the self-assembled structures of DDAI-OPV3-IADD over the course of a 3 μs coarse-grained molecular simulation. Snapshots of the molecular simulation show molecules in which the OPV3 π cores are colored blue, the peptide wings faded grey, and water is removed for clarity. The interaction graph corresponding to each snapshot is shown directly below each image. The vertices V corresponding to each peptide are colored red, and edges E between peptide pairs defined by $d_{ab}^{\text{optical}} < r_{\text{cut}} = 0.7$ (Eqn. 2.1) and colored grey. The average degree $\kappa = \frac{2|E|}{|V|}$ is reported at the bottom of each panel. At $t =$ 0 ns, the 96 randomly placed peptides form essentially a monomeric dispersion, with the exception of five dimer pairs, and the system possesses a correspondingly low $\kappa = 0.13$. As the simulation progresses, the peptides self assemble under the influence of hydrogen bonding, π-π stacking, and hydrophobic interactions into small ($t = 50$ ns, $\kappa = 3.69$) and then larger ($t = 500$ ns, $\kappa = 5.67$; $t = 3{,}000$ ns, $\kappa = 6.08$) aggregates with a commensurate increase in the mean degree κ. In this figure, and throughout the paper, molecular renderings are generated using VMD [211], and interaction graphs produced using NetworkX [212].

that low-dimensional GPRs tend to be more robust, chemically similar molecules tend to be embedded proximately in the latent space providing interpretability of the chemical space through dimensionality reduction, and the continuous and differentiable nature of the latent space makes it well-suited to global optimization [42, 215].

We represent the DXXX-OPV3-XXXD molecules to the VAE only through the identity

of the XXX tripeptide, since this is the only differentiating feature between molecules. We base this representation on the coarse-grained Martini model used to perform our molecular simulations. This representation comprises two components for each molecule i: (i) an adjacency matrix $\mathbf{A}_i$, which captures the connectivity of beads within the tripeptide, and (ii) a one-hot encoded composition vector of bead-types $\mathbf{T}_i$ specifying the identity of the Martini beads (Fig. 2.4). Since peptide sequences may contain varying numbers of coarse-grained beads, we standardize the size of the adjacency matrix $\mathbf{A}_i \in \mathbb{R}^{15\times15}$ to be sufficiently large enough to accommodate the largest tripeptide (Trp-Trp-Trp) and pad the array with zeroes for smaller molecules. A one-hot composition vector of length $\mathbf{T}_i \in \mathbb{R}^{75}$ is sufficient to accommodate all tripeptide compositions considered. For each molecule i, the tuple $(\mathbf{A}_i, \mathbf{T}_i)$ defines the input provided to the VAE.

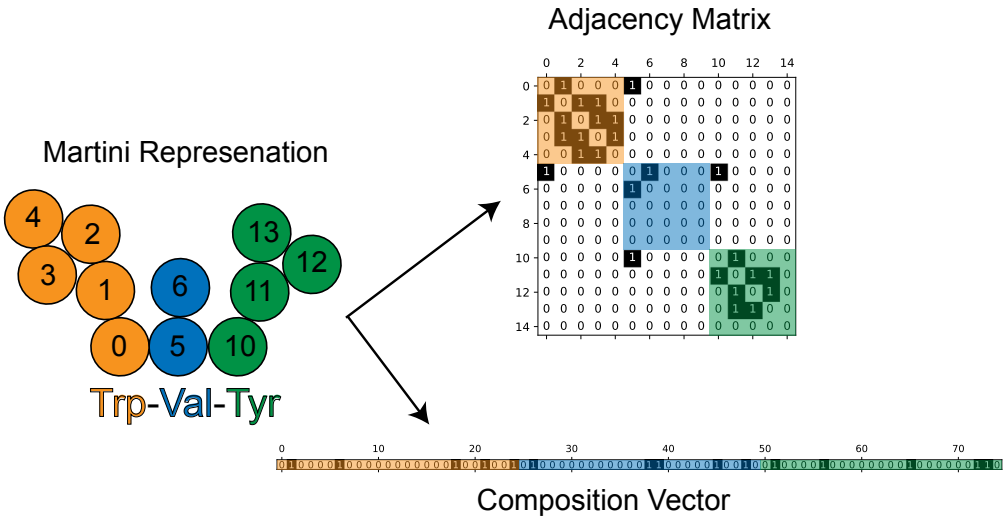


Figure 2.4: Schematic of the representation of each XXX tripeptide to the VAE. The Martini representation of each tripeptide i, in this example Trp-Val-Tyr (WVY), is converted into an adjacency matrix $\mathbf{A}_i \in \mathbb{R}^{15\times15}$ specifying the connectivity of beads within the tripeptide and a one-hot composition vector $\mathbf{T}_i \in \mathbb{R}^{75}$ specifying the identity of the beads. The tuple $(\mathbf{A}_i, \mathbf{T}_i)$ defines the input provided to the VAE. A pre-defined sequential numbering is employed for the beads in each amino acid. The adjacency matrix is padded with rows and columns of zeros to represent tripeptides containing fewer than the maximum number of 15 beads. The colored blocks in the adjacency matrix and composition vector correspond to the colors of the amino acids in the Martini molecule.

The architecture of the VAE is illustrated in Fig. 2.5. Given the two-part input $(\mathbf{A}_i, \mathbf{T}_i)$

for molecule i, the encoder block processes this through two parallel networks to perform feature extraction from each input. The $\mathbf{A}_i$ resemble a small image which motivate using a short series of convolutional layers to treat these inputs, whereas the binary $\mathbf{T}_i$ vectors are passed through a series of fully-connected dense layers. The features extracted by the encoder through the two parallel encoder networks are subsequently concatenated and used to generate the mean $\boldsymbol{\mu}_i$ and standard deviation $\boldsymbol{\sigma}_i$ of a Gaussian distributed latent space embedding $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \in \mathbb{R}^d$. The dimensionality of the latent space is treated as a hyperparameter that is optimized during each cycle of active learning and is found to lie in the range $d \in [4, 10]$. The decoder then attempts to reconstruct $(\mathbf{A}_i, \mathbf{T}_i)$ from the latent encoding $\mathbf{z}_i$ again using two parallel networks. The part of the decoder predicting the reconstruction $\hat{\mathbf{T}}_i$ is identical to the architecture of the encoder, whereas the part predicting the reconstruction $\hat{\mathbf{A}}_i$ is simply another series of fully-connected layers that is reshaped to match the size of the input. The overall action of the VAE is the functional composition of the encoder $E : \mathbf{z}_i = E(\mathbf{A}_i, \mathbf{T}_i)$ and decoder $D : (\hat{\mathbf{A}}_i, \hat{\mathbf{T}}_i) = D(\mathbf{z}_i)$ blocks such that the total effect of the network is $(\hat{\mathbf{A}}_i, \hat{\mathbf{T}}_i) = (D \circ E)(\mathbf{A}_i, \mathbf{T}_i)$.

The VAE is trained by minimizing the VAE loss $\mathcal{L}_{VAE}$ composed of a reconstruction

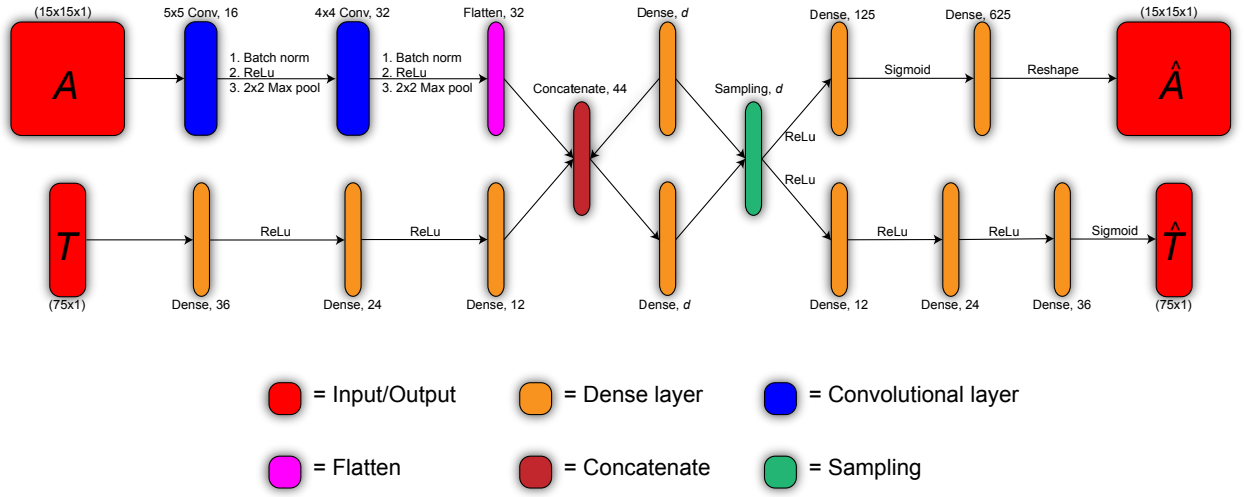


Figure 2.5: Architecture of the variational autoencoder (VAE) used to generate the DXXX-OPV3-XXXD latent space embedding. The VAE accepts as inputs adjacency matrix and composition vector tuples $(\mathbf{A}_i, \mathbf{T}_i)$ and employs two parallel encoders to perform feature extraction and learn the mean $\boldsymbol{\mu}_i$ and standard deviation $\boldsymbol{\sigma}_i$ of a Gaussian distributed latent space embedding $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \in \mathbb{R}^d$. The decoder generates approximate reconstructions $(\hat{\mathbf{A}}_i, \hat{\mathbf{T}}_i)$ of the inputs from the latent space representation. The network is trained by minimizing a loss function balancing reconstruction accuracy and a regularization term constraining the latent space to follow a multidimensional Gaussian distribution (Eqn. 8.1). The dimensionality d of the latent space is treated as a hyperparameter that is optimized during each cycle of active learning.

term $\mathcal{L}_{Rec}$ and a Kullback-Leibler (KL) divergence term $\mathcal{L}_{KL}$ [214, 216],

$$\mathcal{L}_{VAE} = \mathcal{L}_{Rec} + \mathcal{L}_{KL}, \tag{2.3}$$

$$\begin{aligned}\mathcal{L}_{Rec} &= \mathbb{E}_i[\mathrm{BCE}(\hat{\mathbf{A}}_i, \mathbf{A}_i) + \mathrm{BCE}(\hat{\mathbf{T}}_i, \mathbf{T}_i)]) \\ &\approx \sum_{i \in \text{mini-batch}} [\sum_{j=1}^{15}(A_{i,j}\log(\hat{A}_{i,j}) + (1 - A_{i,j})\log(1 - \hat{A}_{i,j})) \\ &\quad + \sum_{j=1}^{75}(T_{i,j}\log(\hat{T}_{i,j}) + (1 - T_{i,j})\log(1 - \hat{T}_{i,j}))],\end{aligned} \tag{2.4}$$

$$\begin{aligned}\mathcal{L}_{KL} &= -D_{KL}(\mathbf{z} = E(\mathbf{A}, \mathbf{T}) \;||\; \mathcal{N}(0, \mathbf{I})) \\ &\approx \sum_{i \in \text{mini-batch}} [-\frac{1}{2}\sum_{j=1}^{d}(1 + \log(\sigma_{i,j}^2) - \sigma_{i,j}^2 - \mu_{i,j}^2)],\end{aligned} \tag{2.5}$$

where BCE($\mathbf{x}, \mathbf{y}$) is the binary cross entropy between the reconstructions $\mathbf{x}$ and ground truth $\mathbf{y}$, $D_{KL}(Q \mid\mid P)$ is the Kullback-Leibler divergence from P to Q, and $\mathbf{I}$ is the n-by-n identity matrix. The reconstruction term $\mathcal{L}_{Rec}$ encourages the VAE to reconstruct the inputs through the low-dimensional latent space information bottleneck. In contrast to a vanilla autoencoder which only aims to minimize $\mathcal{L}_{Rec}$, the KL divergence term $\mathcal{L}_{KL}$ is an effective regularizer which imposes a multivariate Gaussian prior on the latent space and prevents the VAE from essentially "memorizing" the data set and learning a trivial identity mapping through a disconnected latent space [216]. Training is performed by passing tuples $(\mathbf{A}_i, \mathbf{T}_i)$ through the VAE in mini-batches of size 32 and updating the network parameters with mini-batch gradient descent using the Adam optimizer [217]. The VAE loss $\mathcal{L}_{VAE}$ is typically observed to plateau within 4,000 epochs. VAE hyperparamters were selected by exploratory hyperparameter tuning of the batch size over the range [8, 128], the learning rate over the range [0.1, 0.00001], the number of hidden layers in the decoder/encoder over the range [1,10], the dimension of the dense layers over the range [8, 1024], and the size of the convolutional kernel over the range [2, 7]. The final set of hyperparameters including batch size of 32, learning rate of 0.001, and the neural network architecture detailed in Fig. 5 were selected after observing high R^2 scores in downstream property prediction (cf. Sections 2.2.3, 2.2.5, & 3.1). We note that the regularization introduced by the KL divergence term $\mathcal{L}_{KL}$ serves to prevent over-fitting and enables us to train over the full set of molecules to be embedded by the VAE.

We present in Fig. 2.6 an example of a $d = 3$ VAE latent space embedding of the DXXX-OPV3-XXXD family. We color each member of the family by the number of beads in the XXX tripeptide to show that the first dimensional of the latent space z_1 is approximately correlated with molecular size, possessing a Pearson correlation coefficient $\rho(z_1, \text{size}) = 0.858$ (p-value $< 1{\times}10^{-15}$). The other two dimensions in this example are also some functions of molecular composition and topology, but prove more challenging to correlate with phys-

ically interpretable observables. Physical interpretability of the latent space dimensions is a pleasing but not required property of the embedding. The primary purpose of the VAE embedding is to provide a smooth, low-dimensional molecular representations for the GPR surrogate model. We note that the latent space embedding could be shaped and made more interpretable by simultaneous training of a supervised regression model as suggested by Gomez-Bombarelli et al. [42].

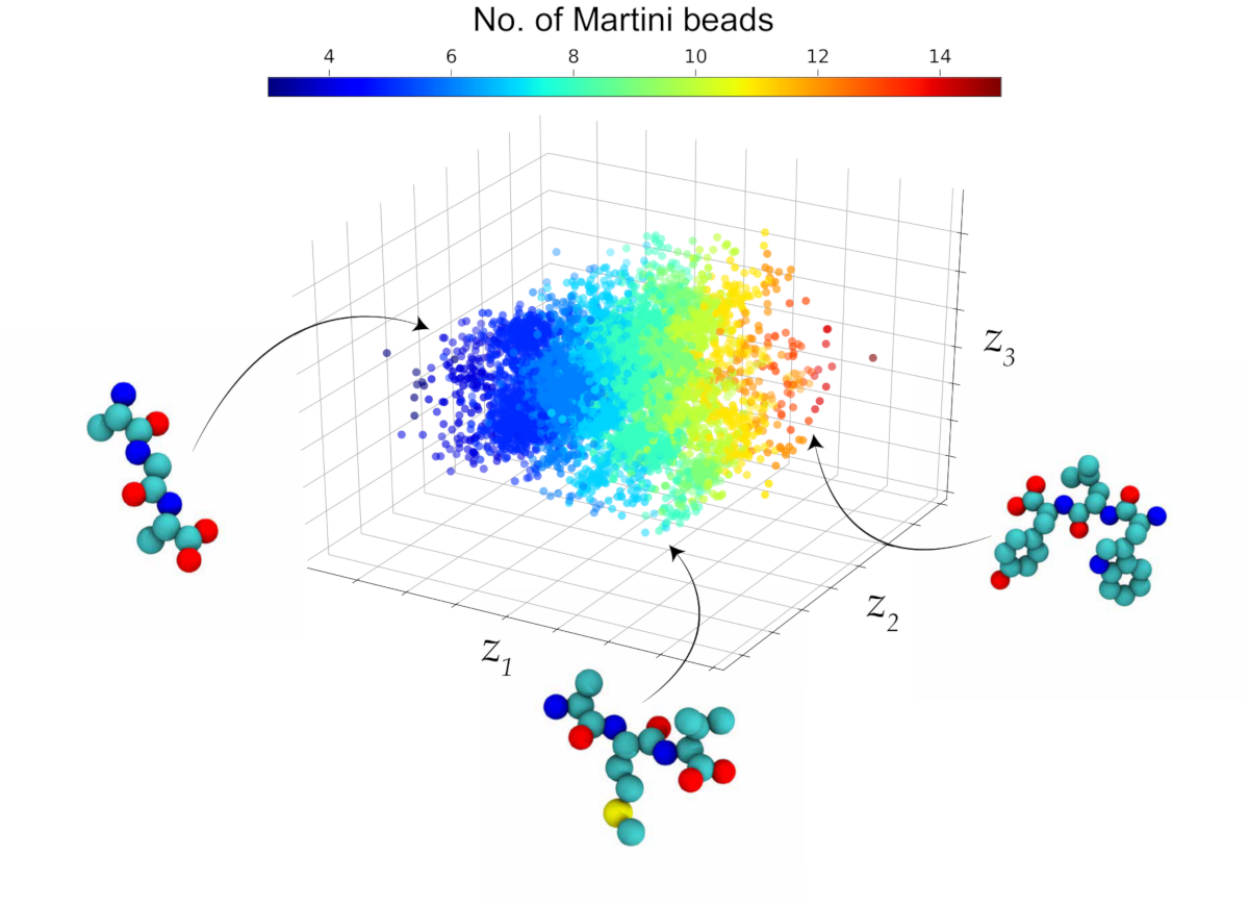


Figure 2.6: Illustrative visualization of $d = 3$ VAE latent space embedding of the DXXX-OPV3-XXXD family. The embedded molecules are colored according to the number of beads in the XXX tripeptide and selected molecules are visualized. The first dimensional of the latent space z_1 is correlated with molecular size $\rho(z_1, \text{size}) = 0.858$ (p-value $< 1{\times}10^{-15}$) providing a visual illustration that similar molecules are embedded close together in the VAE latent space.

Step 3: Gaussian process regression surrogate models

Fitness measurements $f_i,\ i \in \textit{sampled}$ are available for those molecules DXXX-OPV3-XXXD$_i$ for which we have performed coarse-grained molecular simulation. Given these data we wish to predict the fitness of all remaining candidates $\hat{f}_j,\ j \in \textit{unsampled}$. This constitutes a supervised regression task where we wish to train a surrogate model f over a small number of training examples to predict the fitness of out-of-training examples as a function of their location in the VAE latent space: $f : \hat{f}_i = f(\mathbf{z}_i) = (f \circ E)(\mathbf{A}_i, \mathbf{T}_i)$. In this manner, the regression model "short circuits" expensive direct simulation prediction of fitness with a cheap surrogate model, and eliminates the need to perform exhaustive calculations over all molecules in the family. The quality of the model predictions depends on the number and chemical similarity of the training data: the model is expected to perform better with larger training sets and make more accurate predictions for out-of-training examples that are chemically similar to examples in the training set. As such, we expect the model to improve with additional cycles around the active learning loop. For the purposes of active learning (Section 2.3.2), it is also vital to perform uncertainty quantification on the model predictions so that we can both direct sampling towards the most high-performing candidates predicted by the model (exploitation) and towards undersampled areas where the model possesses the highest uncertainties (exploration) [55]. For this reason, we select Gaussian process regression (GPR) to construct our surrogate model $f : \hat{f}_i = f(\mathbf{z}_i)$ as a flexible, non-parametric, Bayesian regression approach that comes with built-in uncertainty estimates [63, 64, 58, 55].

The fundamental principle of a GPR is to employ a Gaussian process to specify a Bayesian prior distribution over regression functions fitting the data, and then to compute the posterior distribution over those functions that are in agreement with the training data [58]. The Gaussian process is fully specified by its mean function, which is typically defined to be zero,

and its covariance function for which we choose the popular squared exponential kernel,

$$k(\mathbf{z}, \mathbf{z}') = \exp(-\frac{1}{2\gamma}||\mathbf{z} - \mathbf{z}'||^2), \tag{2.6}$$

where $\mathbf{z}$ and $\mathbf{z}'$ denote latent space vectors and the bandwidth of the kernel γ is a hyper-parameter defining the characteristic length scale over which latent space vectors "see" one another. Under these choices, the predicted fitness $f^* = f(\mathbf{z}^*)$ for a new point $\mathbf{z}^*$ outside of the training data is a Gaussian distributed random variable with [58, 64],

$$\begin{aligned} f^* &\sim \mathcal{N}(\mu_{f^*}, \sigma_{f^*}), \\ \mu_{f^*} &= K(\mathbf{z}^*, \mathbf{Z}) \left[K(\mathbf{Z}, \mathbf{Z}) + (\boldsymbol{\sigma}_f^2)^T \mathbf{I}\right]^{-1} \mathbf{f}, \\ \sigma_{f^*} &= K(\mathbf{z}^*, \mathbf{z}^*) - K(\mathbf{z}^*, \mathbf{Z}) \left[K(\mathbf{Z}, \mathbf{Z}) + (\boldsymbol{\sigma}_f^2)^T \mathbf{I}\right]^{-1} K(\mathbf{z}^*, \mathbf{Z})^T, \end{aligned} \tag{2.7}$$

where $\mathbf{I}$ is the n-by-n identity matrix, $\mathbf{f} = [f_1, f_2, \ldots, f_n]^T$ is the vector of (noisy) measurements of fitness for the n training points $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$ computed in our coarse-grained molecular simulations, and $\boldsymbol{\sigma}_f^2 = \left[\sigma_{f_1}, \sigma_{f_2}, \ldots, \sigma_{f_n}\right]^T$ are associated variances of assumed i.i.d. Gaussian noise estimated by block averaging (Section 2.3.2), and the K matrices hold the covariances within and between the training data $\mathbf{Z}$ and new point $\mathbf{z}^*$,

$$K(\mathbf{Z}, \mathbf{Z}) = \begin{bmatrix} k(\mathbf{z}_1, \mathbf{z}_1) & k(\mathbf{z}_1, \mathbf{z}_2) & \cdots & k(\mathbf{z}_1, \mathbf{z}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{z}_n, \mathbf{z}_1) & k(\mathbf{z}_n, \mathbf{z}_2) & \cdots & k(\mathbf{z}_n, \mathbf{z}_n) \end{bmatrix}, \tag{2.8}$$

$$K(\mathbf{z}^*, \mathbf{Z}) = \left[k(\mathbf{z}^*, \mathbf{z}_1), k(\mathbf{z}^*, \mathbf{z}_2), \cdots, k(\mathbf{z}^*, \mathbf{z}_n)\right], \tag{2.9}$$

$$K(\mathbf{z}^*, \mathbf{z}^*) = k(\mathbf{z}^*, \mathbf{z}^*). \tag{2.10}$$

The $(\boldsymbol{\sigma}_f^2)^T \mathbf{I}$ terms account for the uncertainty inherent in our measurements of $\mathbf{f}$ through an assumed Gaussian noise model [64]. These terms can also be conceived as a Tikhonov

(a.k.a. ridge or nugget) regularization of the $K(\mathbf{Z}, \mathbf{Z})$ matrix that stabilizes its matrix inverse and is particularly useful when this matrix is ill-conditioned due to the close proximity of two or more training points in $\mathbf{Z}$ [218]. A corollary of this regularization is that the GPR posterior is not a perfect interpolator of the training data due to the presence of measurement noise, and we should anticipate residual discrepancies on the order of $\boldsymbol{\sigma}_f$ between the GPR predictions and the our measurements of $\mathbf{f}$. The predictive accuracy and robustness of the GPR is enhanced by the smooth, continuous, and low-dimensional nature of the VAE latent space, which embeds chemically similar points nearby one another and therefore promotes transfer of information to new out-of-training points based on chemically proximate training examples. The GPR prior and posterior are updated during each cycle of the active learning loop as additional training data are collected.

Step 4: Bayesian optimization

The final step in the cycle is to use the predictions of the surrogate GPR model to identify the next peptide candidates to simulate. We frame this active learning problem as a Bayesian optimization, where we have an expensive, non-differentiable, black-box function with noisy evaluations – the fitness of each molecule evaluated by coarse-grained molecular simulation – that we wish to optimize in the minimum number of evaluations. Bayesian optimization defines an acquisition function u that wraps around the current surrogate model to identify peptides with a high chance of being better than the current leader in the training data. We can represent optimization of the acquisition function as,

$$\mathbf{z}^{\dagger} = \underset{\mathbf{z}}{\operatorname{argmax}}\, u(\mathbf{z}|\mathbf{Z} = \{(\mathbf{z}_1, f_1), (\mathbf{z}_2, f_2)..., (\mathbf{z}_n, f_n)\}), \tag{2.11}$$

where $\mathbf{z}^{\dagger}$ is the VAE latent space coordinates of the DXXX-OPV3-XXXD molecule that maximizes the acquisition function u, and the maximization is conditioned on the n samples

$\{(\mathbf{z}_1, f_1), (\mathbf{z}_2, f_2)..., (\mathbf{z}_n, f_n)\}$ collected to date. The surrogate model f enters the maximization through the choice of acquisition function, for which many choices are available [55]. We employ the popular expected improvement (EI) acquisition function that provides a balanced trade-off between exploitation – selection of points where the surrogate model posterior mean $\mu_f(\mathbf{z})$ is large – and exploration – selection of points where the surrogate model posterior variance $\sigma_f(\mathbf{z})$ is large [65, 66, 55]. Following Lizotte, the EI is defined as [67],

$$u(\mathbf{z}|\mathbf{Z}) = \mathrm{EI}(\mathbf{z}|\mathbf{Z}) = \begin{cases} (\mu_f(\mathbf{z}) - f(\mathbf{z}^+) - \xi)\Phi(Z) + \sigma_f(\mathbf{z})\phi(Z) & \sigma_f(\mathbf{z}) > 0 \\ 0 & \sigma_f(\mathbf{z}) = 0 \end{cases}, \tag{2.12}$$

$$Z = \begin{cases} \frac{\mu_f(\mathbf{z}) - f(\mathbf{x}^+) - \xi}{\sigma_f(\mathbf{z})} & \sigma_f(\mathbf{z}) > 0 \\ 0 & \sigma_f(\mathbf{z}) = 0 \end{cases}, \tag{2.13}$$

where $f(\mathbf{z}^+)$, $\mathbf{z}^+ \in \mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$ is the maximum fitness value among all n sampled candidates to date, Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution, and the hyperparameter ξ controls the exploration-exploitation trade-off. The first term in Eqn. 2.12 promotes exploitation and the second promotes exploration: when ξ is small the EI will favor exploitation and select points with high posterior mean, while if ξ is large exploration is performed selecting points with large posterior uncertainty [55].

Active learning typically proceeds by selecting a fixed value $\xi = 0.01$ [55, 67] of the exploration-exploitation trade-off, identifying the candidate that maximizes the EI, and then performing expensive function evaluation (here a coarse-grained molecular simulation) for that candidate. We employ a slightly modified version of this approach that effectively integrates over ξ and performs active learning in batches, which has the advantages of (i) eliminating the sensitivity in selection to the hyperparameter ξ, (ii) spreading the exploit-explore trade-off, and (iii) making more efficient use of parallel compute resources to conduct

multiple simulations in parallel in the same wall clock time. Specifically, we maximize the EI acquisition function over the range $\log_{10}\xi \in [-4, 0.4]$ and available computational resources enabled us to select up to four candidates over this range as the "next-best" candidates to simulate. Molecules that have already been sampled in preceding rounds are excluded from the pool of available candidates at each round. Where more than four candidates emerge from the EI maximization, we randomly select four members of this set. An example of this selection procedure is presented in Fig. 2.7. Coarse grained molecular simulations of these optimal candidates are then performed to commence another round of the active learning cycle.

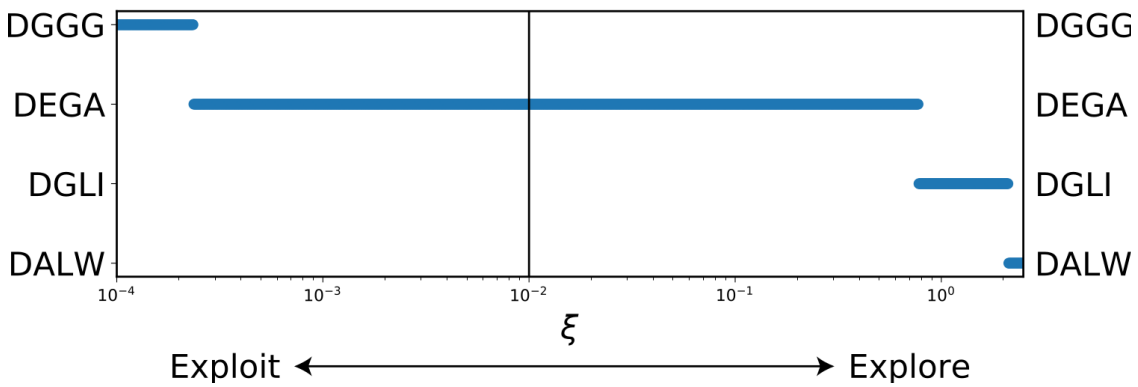


Figure 2.7: Active learning candidate selection. The expected improvement (EI) acquisition function (Eqn. 2.12) is evaluated over for all unsampled members of the DXXX-OPV3-XXXD family at values of the exploit-explore hyperparameter ξ over the range $\log_{10}\xi \in [-4, 0.4]$. The blue points in the graph indicate which DXXX candidate maximizes the EI at each value of ξ. In this illustrative example there are four candidates – DGGG, DEGA, DGLI, and DALW – that maximize the EI over the range of ξ considered. The vertical line shows the recommended value of $\xi = 0.01$ suggested in literature [55, 67], which would result in the selection of only DEGA as the next molecule to simulate. In our approach, we select all four of the molecules DGGG, DEGA, DGLI, and DALW that maximize EI over the entire range of ξ considered as the next best candidates to simulate in parallel in the next round of coarse-grained molecular simulations.

Hyperparameter optimization

The dimensionality d of the VAE latent space embedding and bandwidth γ of the GPR kernel are tunable hyperparameters to be optimized during each cycle of the active learning loop. We perform simultaneous tuning of d and γ during each round by creating 50 embeddings of all 8,000 DXXX-OPV3-XXXD molecules into the VAE latent space $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \in \mathbb{R}^d$, each employing different realization of random numbers to sample from the latent space Gaussian for each point, for $d = [3, 10]$. We then optimize γ for each embedding over the range $\gamma = [0.001, 100]$ using a line search followed by Nelder-Mead optimization to maximize the GPR accuracy under cross-validation. We employ leave-one-out cross-validation (LOO-CV) for the first five cycles of the active learning, and then 5-fold CV for subsequent rounds due to the high cost of LOO-CV for larger quantities of samples. The best performing VAE embedding and associated optimal d and γ are adopted for the remainder of the current active learning cycle.

Stop criteria

We cycle around the active learning loop until the GPR surrogate model no longer improves with the collection of additional training data. A number of stopping criterion for active learning have been proposed [68, 69, 70, 71, 72, 73], but in this work we monitor and define convergence using the stabilizing predictions (SP) method that evaluates performance based on unlabeled data [72] and the performance difference (PD) method that considers the labeled examples [74]. The SP method examines the predictions of consecutive models at each iteration of the active learning procedure on a randomly selected set of 500 points, called the stop set, which is held constant throughout the active learning. We measure the difference in the regression predictions between subsequent rounds using the average Bhattacharyya distance D_B[219] between the posterior of consecutive GPR models over the stop set. Large differences in D_B indicate the model is continuing to update the GPR posterior, whereas

small values indicate that the surrogate model predictions have stabilized.

The PD method is used to evaluate model performance by 5-fold CV of the R^2 score over the accumulated labeled samples collected to date within each round of active learning. A plateau in the R^2 indicates that additional observations result in only marginal improvements to the GPR fit [74]. A caution in assessing convergence using labeled data is that these data may not be representative of the data as a whole [69, 220]. These concerns are mitigated in our application since our initial data set comprises a set of randomly selected peptides to initialize the active learning procedure and we collect up to four new data points each round across the exploit-explore spectrum to assure broad sampling of chemical space.

2.3.3 Nonlinear manifold learning of assembly pathways

We employ diffusion maps as a manifold learning approach to identify the low-dimensional assembly pathways by which the various DXXX-OPV3-XXXD molecules self-assemble into the terminal aggregates. We have previously described the application of diffusion maps to self-assembling systems in Refs. [221, 222, 210, 223]. In brief, we compute a distance metric $d(i, j)$ between each pair of interaction graphs i and j harvested from each frame of each molecular simulation trajectory. A number of graph kernels at varying levels of sophistication and abstraction have been proposed to measure the similarity between pairs of graphs [224, 225, 226, 227, 228]. We follow the approach of Reinhart et al. who employed graphlet decompositions as a diffusion map distance metric to analyze colloidal crystallization [229]. This approach featurizes a graph by enumerating all topologically unique subgraphs ("graphlets") with associated node permutations ("orbits") within the network up to a certain subgraph size (usually up to five vertices), and creating a vector of orbit counts for each vertex in our graph [209, 224, 229]. The vector of orbit counts at each vertex is reweighted to account for over-counting of the smaller graphlets contained in the larger ones (i.e. counts of graphlets comprising two vertices are necessarily contained in counts of graphlets comprising three or

more vertices), averaged over all vertices in the graph, and normalized to unit length. This vector represents a featurization of the graph that is permutationally invariant to vertex labeling, and the L2-norm between pairs of vectors defines the graph kernel $d(i,j)$ used to evaluate pairwise distances between our graphical representations of the configurational state of the molecular system.

Diffusion maps then proceed by applying a Gaussian kernel to construct the convolved similarity matrix,

$$A_{i,j} = \exp\left(\frac{-\left(d(i,j)^{\alpha}\right)^2}{2\epsilon}\right), \tag{2.14}$$

where the kernel bandwidth ϵ controls the hop size of the random walk and can be automatically tuned based on the distribution of the $A_{i,j}$ [230, 210, 222]. The use of the hyperparameter $\alpha \in (0,1]$ was proposed by Wang et al. within a density-adaptive extension of diffusion maps that greatly improves the performance of diffusion maps in applications to systems with large differences in the density of points in the high-dimensional space [231]. For $\alpha = 1$ we recover standard diffusion maps; for $\alpha \to 0$ the pairwise distances become increasingly similar and large fluctuations in the density of points in the high-dimensional space are smoothed out. Adopting the tuning procedure proposed in Ref. [231] we adopt $\alpha = 0.15$.

The $\mathbf{A}$ matrix is row normalized to create the right stochastic Markov transition matrix,

$$\mathbf{M} = \mathbf{D^{-1}A}, \tag{2.15}$$

Where $\mathbf{D}$ is a diagonal matrix of the row sums of $\mathbf{A}$.

$$D_{i,j} = \sum_{j=1}^{N} A_{i,j}. \tag{2.16}$$

The matrix element $M^t_{i,j} = p_t(i,j)$ can interpreted as the probability $p_t(i,j)$ of hopping from

point i to point j in t steps of the discrete random walk [230, 232]. Diagonalization of $\mathbf{M}$ produces an ordered set of eigenvectors and eigenvalues $\{(\psi_1 = \vec{1}, \lambda_1 = 1), (\psi_2, \lambda_2), (\psi_3, \lambda_3), ...\}$ with $\lambda_1 = 1 \geq \lambda_2 \geq \lambda_3 \geq \ldots$. The first pair $(\psi_1 = \vec{1}, \lambda_1 = 1)$ is trivial and associated with the stationary distribution of the random walk [230]. The higher order eigenvectors are associated with a hierarchy of increasingly fast relaxation modes of the random walk. Dimensionality reduction is achieved by identifying a gap in the eigenvalue spectrum after the λ_{k+1} to resolve a subspace of slowly relaxing dynamical modes $\{\psi_2, \psi_3, \ldots, \psi_{k+1}\}$. The diffusion map embedding is the projection of the i^{th} interaction graph into the i^{th} component of the top k non-trivial eigenvectors,

$$i \mapsto (\psi_2(i), \psi_3(i), ..., \psi_{k+1}(i)). \tag{2.17}$$

We implement this formalism using the memory and compute efficient pivot diffusion map approach that reduces the scaling in the number of points N from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \times n)$, where $n << N$ is the number of so-called "pivot points" employed [233]. This approach enables the application of diffusion maps to large data sets by performing on-the-fly definition of the n pivot points defining an approximate spanning tree over the high-dimensional data and which are used to support interpolative embeddings of the remaining points.

2.4 Results and Discussion

2.4.1 Active learning identification of optimal candidates

The complete DXXX-OPV3-XXXD family comprises 20^3 = 8,000 members generated by all permutations of placing each of the 20 natural amino acids within the XXX tripeptide. Prior to conducting active learning we filtered this ensemble to eliminate a subset of candidates containing amino acids known and expected to produce undesired assembly behaviors [38]. Specifically, we reduced our search space to the 11^3 = 1331 candidates in the set defined

by $X \in \{\text{Ala}, \text{Gly}, \text{Glu}, \text{Ile}, \text{Leu}, \text{Met}, \text{Phe}, \text{Trp}, \text{Tyr}, \text{Val}, \text{Asp}\}$ to avoid charged and/or polar amino acids expected to interfere with low-pH triggered assembly [192] and focus on those residues that have expressed good assembly behavior in previous experimental work [185, 47, 234, 235].

We perform active learning over DXXX-OPV3-XXXD sequences following the four-part protocol – molecular simulation, VAE latent space embedding, GPR surrogate model construction, optimal selection of next candidates – described in Section 2.3.2 and illustrated in Fig. 2.2. We seeded the search by conducting coarse-grained molecular dynamics simulations of 90 randomly selected members of the family using the simulation protocol detailed in Section 2.3.1. This initial broad sampling over the candidate space provides the GPR surrogate model with diverse training data that enables it to identify more- and less-promising regions of the latent space prior to making any predictions. We term this initial round of active learning as Round 0. We conduct 25 additional rounds of active learning (Rounds 1-25) selecting up to four additional molecules for simulation during each pass. This resulted in a sampling a total of $N = 186$ molecules (2.3% of the 8,000-member complete family; 14.0% of the 1331-member chemically restricted family) and a cumulative 558 μs of simulation time. The particular candidates selected and sampled in each round are listed in Table S1 in the Supplementary Information.

Sampling was terminated by tracking the performance difference (PD) and stabilizing predictions (SP) methods (Section 2.3.2) [72, 74]. The PD method 5-fold cross validation R^2 score commences at a reasonably high value of $\sim$68% – likely due to the relatively large and diverse $N = 90$ initial candidates considered, and plateaus to a quite high value of $\sim$78% by Round 18 (Fig. 2.8a). The SP method reveals a Bhattacharyya distance between successive GPR posteriors of $D_B > 10$ over the first 13 rounds, indicating that the additional training data incorporated into the GPR surrogate model are substantially altering its predictions. After Round 14, the Bhattacharyya distance plateaus to $D_B \sim 2.5$ indicating that the surro-

gate model has stabilized. Rounds 18-25 are therefore proceeding with a stable GPR model and the exploitation candidates identified by the expected improvement acquisition function (Section 2.3.2) furnish the best predictions of the top performing molecules that did not happen to have already been sampled in previous rounds.

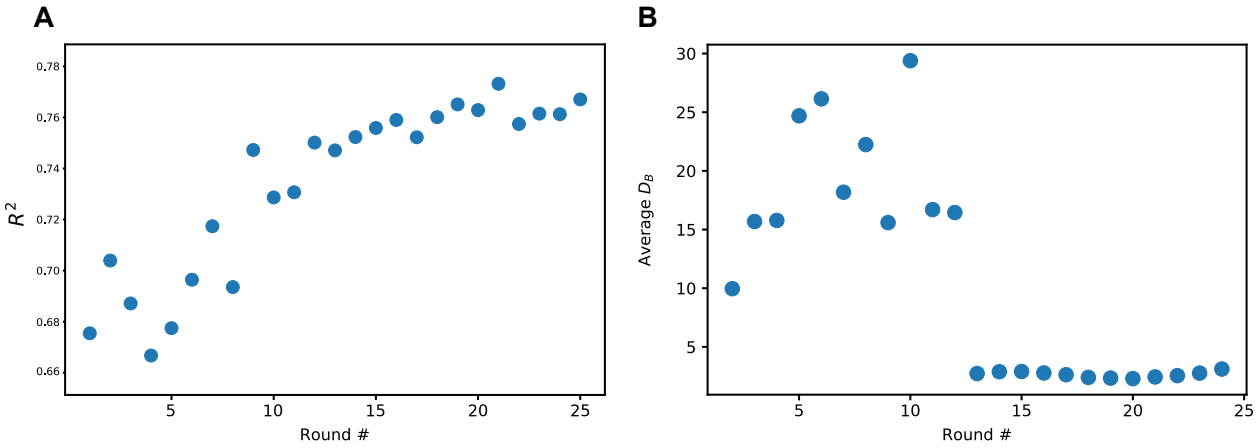


Figure 2.8: Tracking of active learning stop criteria. (a) The performance difference (PD) method 5-fold cross validation R^2 score stabilizes at $R^2 \sim 78\%$ by Round 18. (b) The stabilizing predictions (SP) method Bhattacharyya distance between successive GPR posteriors plateaus close to zero at $D_B \sim 2.5$ after Round 13. Values are reported using a moving average with a window size of three, as recommended in Ref. [72] .

We present in Table 2.1 the top performing molecules among the 186 that were simulated within our active learning protocol. We report their fitness $f_i = \overline{\kappa}_i$ corresponding to the mean number of π-core–π-core contacts per molecule in the terminal self-assembled aggregates (Section 2.3.2), the round of active learning in which they were sampled, and whether they have been previously explored by experiment or simulation. Additional molecules that have been previously identified as high performing by experiment and simulation are also presented for comparison. The list of all 1,331 DXXX-OPV3-XXXD molecules in the family with fitness predictions and rankings assigned by the terminal GPR surrogate model is presented in Table S2. There is very good agreement between the numerical simulation results and the GPR model predictions over the training set of 186 molecules for which measurement data exist: the calculated and predicted values of $f_i = \overline{\kappa}_i$ possess a Pearson correlation

coefficient of $\rho_{\text{Pearson}} = 0.90$ (p-value = 4×10^{-68}) and the calculated and predicted rankings possess a Spearman correlation coefficient of $\rho_{\text{Spearman}} = 0.86$ (p-value = 1×10^{-54}). The agreement is not perfect due to our incorporation of uncertainty estimates in our noisy fitness measurements into GPR training such that the model predictions fall within the error bars of our simulations (Section 2.3.2).

Trends apparent in the active learning-ranking of the tripeptides in terms of amino acid composition and sequence are coincident with aspects of existing understanding, but also suggest new unexplored amino acid sequences as good putative candidates. The bulky aromatic residues F, W, and Y tend to disfavor good assembly behaviors [185], with the large size of these residues impeding good side chain packing and obstructing co-facial stacking of the cores (particularly in the X_3 position of $DX_1X_2X_3$-OPV3-$X_3X_2X_1D$), and their aromatic character disrupting the formation of linear aggregates with in-register stacking of the π-cores by introducing favorable aromatic stacking between the π-cores and peptide wings. These trends are expressed in the low ranking of molecules containing bulky aromatic residues (e.g., DFGG (65), DFAV (85), DFAG (93), DIAG (102), DFAA (111), DFAF(147)) compared to those with smaller hydrophobic side chains (e.g., DVAG (19), DAAG (33), DGAG (45)). The active learning protocol also identifies as highly ranked a number of previously unknown candidates enriched in smaller hydrophobic residues. Interestingly, a number of highly ranked candidates contain an M residue in the X_3 position (e.g., DIAM (3), DGGM (11)). Methionene-containing DXXX-OPV3-XXXD molecules have been completely unexplored due, in part, to the expectation that a thioether group would likely disfavor hydrophobic association. Our calculations predict these candidates to possess excellent assembly behaviors and suggest them as novel molecules for experimental investigation.

Table 2.1: Top 15 DXXX-OPV3-XXXD molecules identified by the active learning protocol. Additional molecules previously studied in simulation and experiment are also shown for comparison.

Rank (out of 186)	Molecule (DXXX)	$\overline{\kappa}$	Discovery round	Previously known?
1	DEAA	6.06 ± 0.02	1	N
2	DDAI	6.03 ± 0.02	0	N
3	DIAM	6.01 ± 0.02	17	N
4	DVAA	5.95 ± 0.03	9	N
5	DAAV	5.92 ± 0.03	19	N
6	DGLG	5.92 ± 0.02	20	N
7	DAEA	5.92 ± 0.02	25	N
8	DAGI	5.90 ± 0.01	21	N
9	DGIG	5.88 ± 0.02	25	N
10	DEAL	5.88 ± 0.01	23	N
11	DGGM	5.87 ± 0.04	0	N
12	DLAV	5.86 ± 0.02	16	N
13	DGDL	5.85 ± 0.03	0	N
14	DGIA	5.80 ± 0.04	15	N
15	DAGL	5.79 ± 0.02	19	N
⋮	⋮	⋮	⋮	⋮
19	DVAG	5.73 ± 0.01	22	Exp [185]
⋮	⋮	⋮	⋮	⋮
33	DAAG	5.62 ± 0.01	2	Exp [185]
⋮	⋮	⋮	⋮	⋮
45	DGAG	5.54 ± 0.03	0	Sim [38]; Exp [185]
⋮	⋮	⋮	⋮	⋮
65	DFGG	5.33 ± 0.03	0	Exp [185]
⋮	⋮	⋮	⋮	⋮
85	DFAV	5.09 ± 0.02	0	Exp [185]
⋮	⋮	⋮	⋮	⋮
93	DFAG	4.98 ± 0.01	0	Sim [37, 36]; Exp [185]
⋮	⋮	⋮	⋮	⋮
102	DIAG	4.86 ± 0.01	2	Exp [185]
⋮	⋮	⋮	⋮	⋮
111	DFAA	4.78 ± 0.01	0	Exp [185]
⋮	⋮	⋮	⋮	⋮
147	DFAF	4.29 ± 0.02	21	Exp [185]

2.4.2 Manifold learning of assembly pathways

The active learning protocol considers only the terminal 50 ns of the 3,000 ns coarse-grained molecular dynamics trajectories to identify DXXX-OPV3-XXXD molecules that form desired pseudo-1D linear aggregates. Having completed the active learning process, we subsequently analyze the ensemble of $N = 186$ molecular simulation trajectories to provide molecular-level understanding of the assembly pathways and mechanisms and furnish design precepts for the observed assembly behaviors as a function of tripeptide sequence.

We hypothesize that the molecular assembly trajectories proceed through configurational phase space over a low-dimensional manifold. We determine this low-dimensional manifold by performing diffusion map manifold learning over the trajectory ensemble [230, 232]. Each frame of each molecular simulation is represented as an interaction graph with vertices V and edges E defined using the optical distance metric (Section 2.3.2). We subsample each trajectory keeping every 20^{th} point and then run diffusion maps on the composite data set of 558,000 graphs as detailed in Section 2.3.3. Diffusion maps then produce a nonlinear projection of this graph ensemble into a low-dimensional space in which graphs sharing a similar structure of edges are embedded close together, and dissimilar graphs embedded far apart. (We emphasize that this low-dimensional embedding represents a nonlinear manifold residing within the *configurational space* of interaction graphs and is completely independent from the VAE latent space embedding of the *chemical space* of XXX tripeptides.) We trace assembly trajectories over this graph embedding to identify DXXX-OPV3-XXXD molecules that follow similar and dissimilar dynamical assembly pathways and terminal states.

The diffusion map eigenvalue spectrum possesses a spectral gap after the third non-trivial eigenvalue, motivating 3D embeddings into the three leading eigenvectors $\{\psi_2, \psi_3, \psi_4\}$. Further, the $\psi_2 - \psi_3$ projection defines a curved, relatively thin manifold indicating that these two embedding dimensions are correlated (Fig. S1) [236]. Accordingly, without too much loss of information we drop ψ_3 and construct visually simpler 2D $\psi_2 - \psi_4$ embeddings

that we present in Fig. 2.9. In Fig. 2.9a we present the composite embedding of all 558,000 simulation snapshots. We find ψ_2 to be moderately strongly correlated with the average number of per molecule π-core–π-core contacts κ ($\rho(\psi_2, \kappa) = 0.78$, p-value $< 1{\times}10^{-15}$), and ψ_4 with the mass averaged cluster size of the system M_z ($\rho(\psi_4, M_z) = 0.62$, p-value $< 1{\times}10^{-15}$) [237].

In Fig. 2.9b-d we highlight the assembly trajectories for three selected molecules: DVAA-OPV3-AAVD as a good assembler with $\overline{\kappa} = (5.85 \pm 0.03)$ and rank = 4/186, DGEG-OPV3-DGEG as an intermediate assembler with $\overline{\kappa} = (5.02 \pm 0.02)$ and rank = 89/186, and DWWI-OPV3-IWWD as a poor assembler with $\overline{\kappa} = (3.74 \pm 0.01)$ and rank = 175/186 (Table S2). These three examples possess assembly pathways over the manifold that are prototypical of three classes of assembly behavior. All pathways commence in the top-left of the manifold at ($\psi_2 \approx -1.0, \psi_4 \approx 1.8$) corresponding to an approximate monomeric dispersion. Good assemblers such as DVAA-OPV3-AAVD follow pathways that travel along the lower perimeter of the manifold and terminate in the top-right corner ($\psi_2 \approx -1.5, \psi_4 \approx 1.8$) $\rightarrow$ ($\psi_2 \approx -1.5, \psi_4 \approx -1.5$) $\rightarrow$ ($\psi_2 \approx 1.5, \psi_4 \approx -1.5$) $\rightarrow$ ($\psi_2 \approx 1.5, \psi_4 \approx 1.5$). The configurations in the top-right corner comprise pseudo-1D aggregates with good in-register stacking between the π-cores and large values of $\overline{\kappa}$. Intermediate assemblers such as DGEG-OPV3-DGEG follow similar pathways that traverse the left and bottom perimeter, but terminate in the bottom-right region of the manifold at ($\psi_2 \approx 1.5, \psi_4 \approx -1.5$). This bottom-right region comprises loosely connected pseudo-1D aggregates which fail to form a globally connected pseudo-1D structure and possess intermediate values of $\overline{\kappa}$. Lastly, poor assemblers such as DWWI-OPV3-IWWD follow pathways that travel along the top of the manifold ($\psi_2 \approx -1.5, \psi_4 \approx 1.0$) $\rightarrow$ ($\psi_2 \approx 0.5, \psi_4 \approx 1.0$) and terminate within the bulk of the manifold ($\psi_2 \approx 0.0, \psi_4 \approx -0.5$) corresponding to disordered aggregates with poor in register stacking and smaller $\overline{\kappa}$.

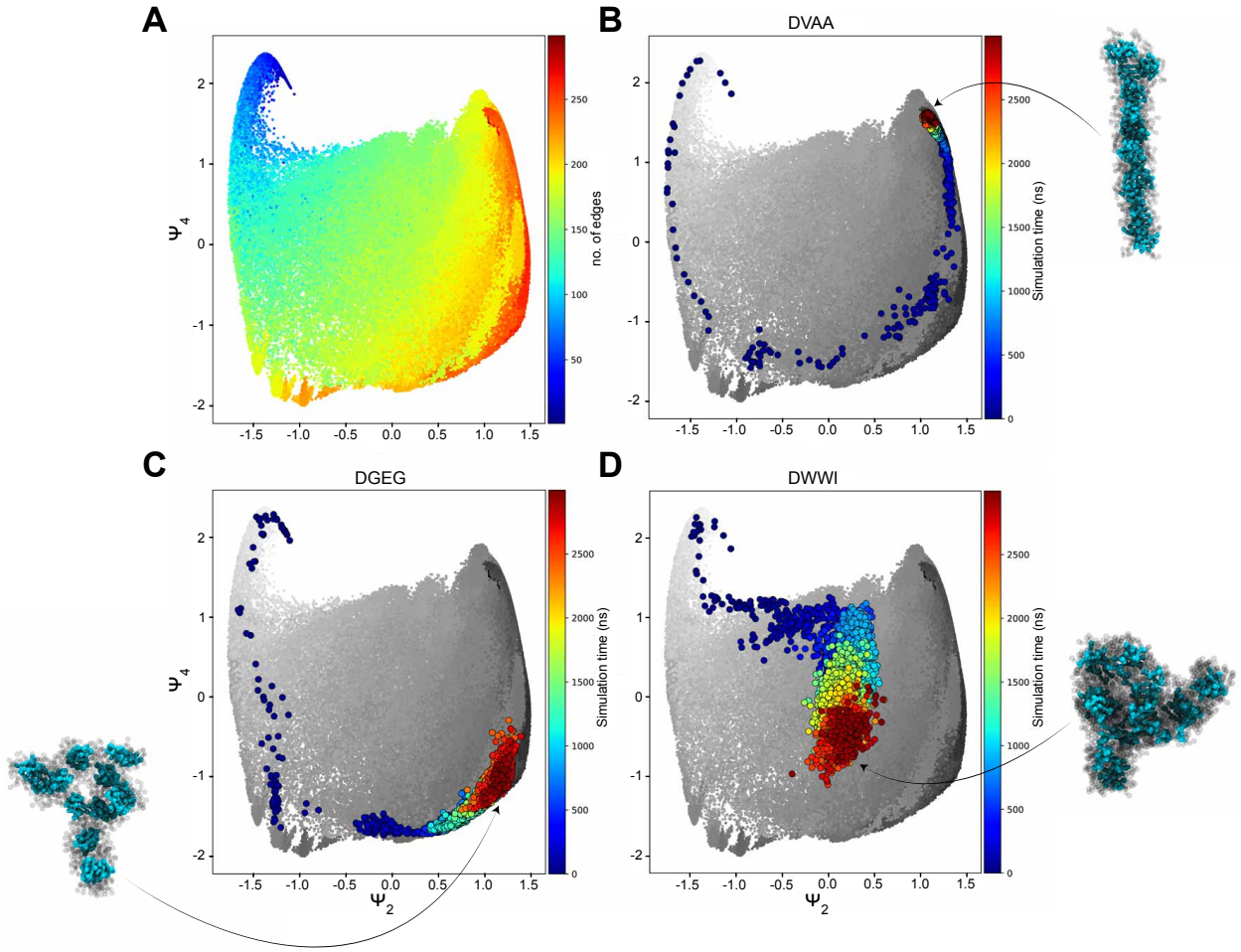


Figure 2.9: Diffusion map embeddings into $\psi_2 - \psi_4$ of the $N = 186$ DXXX-OPV3-XXXD molecular simulation trajectories. (a) Composite embedding of all 558,000 simulation snapshots. Each point represents a snapshot from one of the simulation trajectories and points are colored by the total number of edges in the corresponding molecular interaction graph (Section 2.3.2). Temporal assembly courses of selected molecules over the 2D manifold where points are colored by simulation time: (b) DVAA-OPV3-AAVD (good assembler: $\overline{\kappa} = 5.85 \pm 0.03$, rank = 4/186), (c) DGEG-OPV3-GEGD (intermediate assembler: $\overline{\kappa} = 5.02 \pm 0.02$, rank = 89/186), (d) DWWI-OPV3-IWWD (poor assembler: $\overline{\kappa} = 3.74 \pm 0.01$, rank = 175/186).

2.4.3 Unsupervised spectral clustering into assembly classes

The diffusion map embedding of the assembly trajectories presents a means to identify groups of molecules with similar assembly behaviors and extract design precepts to promote good assembly behavior. We map each of the $N = 186$ DXXX-OPV3-XXXD molecules

in the diffusion map embedding to a single 3D point by averaging over the locations of the final 50 ns of simulation data in the space of the top three nontrivial diffusion map eigenvectors $\{\psi_i\}_{i=2}^4$. We then perform agglomerative hierarchical clustering using Ward's method [238]. We cut the resulting dendrogram to partition in the molecules into three clusters (Fig. S2), and illustrate the clustering of the $N = 186$ molecules within the $\psi_2 - \psi_4$ diffusion map embedding in Fig. 2.10. The three clusters reveal a natural categorization into good, intermediate, and poor assemblers: (i) the green cluster of points in the top-right of the embedding comprises the good assemblers that form pseudo-1D linear stacks, (ii) the red cluster located in the bottom-right of the manifold contains intermediate assemblers that form loosely connected small linear aggregates, and (iii) the orange cluster located in the bulk of the manifold that forms disordered and disconnected clusters with poor π-core stacking. We then propagated the cluster labels defined over these $N = 186$ molecules to the remaining (1,331 - 186) = 1,145 molecules by performing a nearest-neighbor assignation based on distances within the VAE latent space in the terminal round of active learning (Section 2.3.2). A listing of the cluster assignations of each of the 1,331 molecules is provided in Table S3.

Our classification of the 1,331 molecules allows us to perform a statistical analysis of the enrichment or depletion of amino acid residues in good assemblers relative to intermediate or poor assemblers at each of the the three X_i positions in the $DX_1X_2X_3$-OPV3-$X_3X_2X_1D$ sequence (Fig. 2.11). A fuller analysis would account for the complete tripeptide sequence to consider the effects of interactions with the other amino acids, but this simpler one-body analysis is both interpretable and illuminating. Drawing a significance cutoff at 1.5× enrichment or depletion (p-value = 5×10^{-21}, one-tailed Fisher's exact test), within good assemblers at the X_1 position we observe significant enrichment in $\{A, G, I, L, V\}$ residues and depletion in $\{F, W\}$. At X_2, we observe an enrichment in $\{G, I, L\}$ and impoverishment in $\{F, V, W\}$. Finally, X_3 is enriched in $\{G, I, L, V\}$ and impoverished in $\{W, Y\}$.

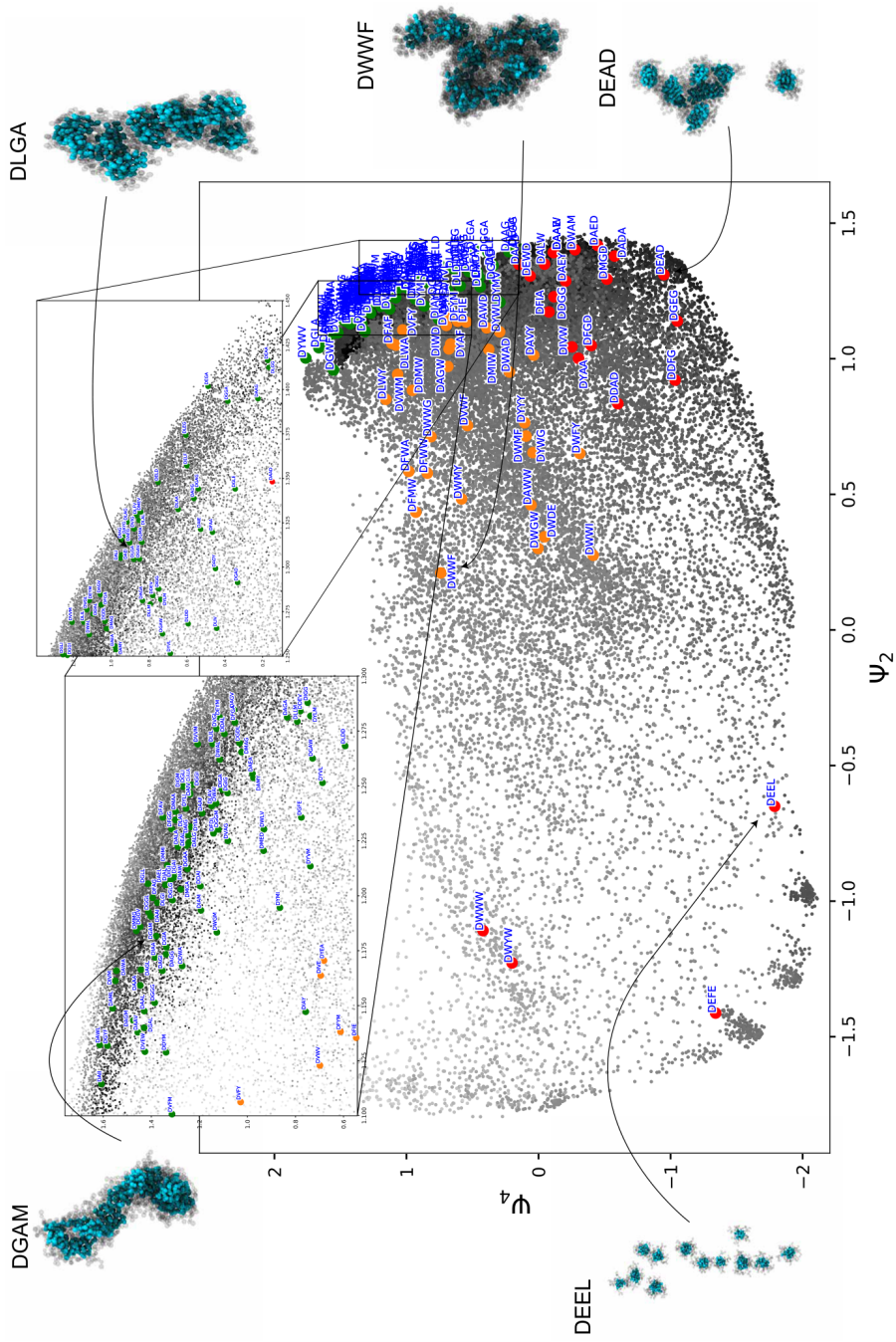


Figure 2.10: Spectral clustering of the N=186 sampled molecules into four clusters which are projected into the top three non-trivial diffusion map eigenvectors $\{\psi_i\}_{i=2}^{4}$. Points are clustered based on agglomerative hierarchical clustering and cutting the resulting dendrogram at the level of three clusters. Grey points represent instantaneous snapshots harvested from all molecular simulations projected into the ψ_2-ψ_4 plane. Colored points represent the average coordinates over the terminal 50 ns of simulation for each of the $N = 186$ molecules. Points are color coded by cluster: green (good assemblers), red (intermediate assemblers), and orange (poor assemblers)

First considering the depleted amino acids, the largest hydrophobic residue W is disfavored in good assemblers at all positions. This can be understood as these bulky aromatic side chains possessing favorable π-stacking interactions with the π-cores, thereby disrupting π-core–π-core stacking. The W residue is most strongly disfavored in the core-adjacent X_3

position, where its bulk and proximity to the core can most effectively disrupt good co-facial core stacking. These observations are consistent with the experimental results in Ref. [185] where smaller UV-vis spectral shifts were observed upon assembly for molecules containing aromatic residues. Of the remaining two aromatic amino acids, F is similarly disfavored, albeit not to the same degree, but the picture for Y is surprisingly nuanced. Y is moderately disfavored at X_1 and strongly disfavored at X_3, but at X_2 it is neither favored nor disfavored. The latter observation was unanticipated, and we currently lack an understanding for why this should be so. This analysis illuminates how location within the tripeptide acts in concert with the inherent physicochemical attributes of an amino acid to modulate its effect.

In regards to the enriched amino acids, the smaller hydrophobic residues G, I, and L are strongly favored at all positions, with I particularly favored in the X_3 position. This preference can be understood as the smaller aliphatic residues enabling closer packing between the peptide wings compared to their bulkier counterparts and their absence of aromatic character reducing interference in the co-facial stacking of π-cores. Residue A is moderately favored at X_1 and X_2, but neither favored nor disfavored at X_3. Contrariwise, V is moderately to strongly favored at X_1 and X_3, but moderately disfavored at X_2.

Finally, there is no strong preferences for residues D, E, and M at any of the three positions, with the exception of a moderate favorability for M at position X_3.

2.5 Conclusions

The primary goal of this work was to employ molecular simulation to identify members of the DXXX-OPV3-XXXD oligopeptide family exhibiting promising assembly behaviors into pseudo-1D nanoaggregates with good optoelectronic properties, and to discover design precepts for the good assemblers. Trial-and-error exploration of the full chemical space is computationally and experimentally intractable, motivating our use of techniques from optimal experimental design and deep representational learning to efficiently traverse the

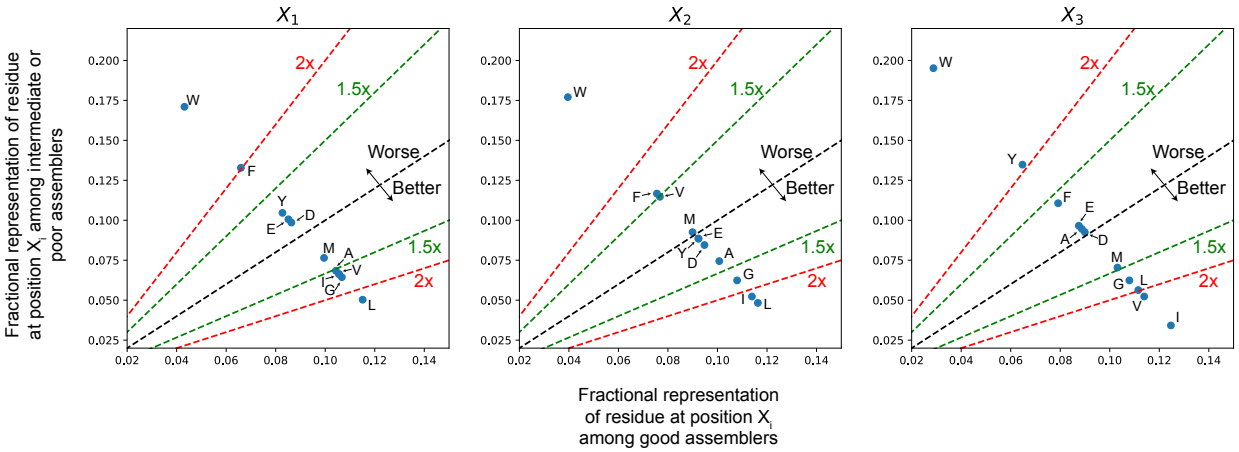


Figure 2.11: Residue enrichment analysis of each position X_i in $DX_1X_2X_3$-OPV3-$X_3X_2X_1D$ within molecules classified as good assemblers relative to those classified as intermediate or poor assemblers. Good assemblers are enriched in amino acids residing below the dashed black line and depleted in those residing above it. Dashed green and red lines show boundaries for 1.5× and 2× differential enrichment and depletion.

space of XXX tripeptide sequences and minimize the number of expensive molecular simulations required to identify the top candidates. Employing a combination of coarse-grained molecular simulation, variational autoencoders, Gaussian process regression, and Bayesian optimization, we define an iterative active learning protocol that constructs surrogate models of assembly behavior based on the simulation data collected to date, and uses these models to optimally direct the next round of simulations. The loop is terminated when the surrogate model ceases to improve with additional simulation data and we can reliably predict the top performing molecules. Using this platform we compute a converged rank ordering of the DXXX-OPV3-XXXD oligopeptides in terms of assembly quality after directly simulating only 2.3% of all possible oligopeptide sequences. The calculated ranking list is consistent with existing understanding of what constitutes good and bad sequences for assembly, but also reveals new promising candidate molecules as superior assemblers that have not previously been considered. Our ranked list presents an inexpensive filtration of the complete DXXX-OPV3-XXXD sequence space to direct expensive experimental synthesis and characterization efforts towards the most promising candidate molecules. To provide context

for the extent of savings afforded by this computational screening our estimates suggest experimentally screening the top-25 candidates determined by this work would take only ∼2 months, while performing active learning with direct experimental feedback instead would require ∼16 months.

A subsequent analysis of the molecular simulation trajectories reveals a low-dimensional manifold within the high-dimensional configurational space over which assembly proceeds. Clustering of the simulation trajectories within this space reveals a natural partitioning of the DXXX-OPV3-XXXD family into good, intermediate, and poor assemblers. Statistical analysis of these classes reveals the good assemblers to be enriched in small and intermediate-sized hydrophobic residues, depleted in large aromatic residues, and that Asp, Glu, and Met do not strongly influence the quality of assembly. The one exception to the latter result is that Met in the X position closest to the π-core does moderately to strongly favor assembly. These design precepts provide understanding of the rankings established by the active learning protocol.

In sum, this work offers a comprehensive investigation of the assembly landscape of the DXXX-OPV3-XXXD family of π-conjugated peptides using Bayesian optimal experimental design to guide expensive coarse-grained molecular simulations over microsecond time scales. Our calculations efficiently furnish a rank ordering of the DXXX-OPV3-XXXD and identify a small number of top-performing candidates. While these predictions are only as good as the accuracy of the (coarse-grained) molecular model, they are consistent with existing physicochemical understanding, and can be viewed as a coarse computational filtration of the complete sequence space that can guide subsequent computation and experiment towards the most promising candidates. Ongoing experimental work will attempt to synthesize and test the optoelectronic properties of the candidates in this work, while future computational studies will generalize the approach to $D(X)_n$-Π-$(X)_n D$ molecules by extending the considered chemical space to include different Π cores, such as perylenediimide (PDI) or oligothiophene

(OT), and varying the length of the peptides. Our platform is also generically extensible to the design of other peptide and peptide-like systems, including antimicrobial peptides, cell-penetrating peptides, intrinsically disordered proteins, and peptoids, where the efficient traversal of chemical space, identification of small numbers of top-performing candidates, and exposure of comprehensible design precepts are prioritized.

Data Availability

The coarse-grained molecular simulation trajectories of the self-assembly of the 186 DXXX-OPV3-XXXD molecules conducted in this work are hosted for free public download at the Materials Data Facility [239], a project affiliated with the NIST Center for Hierarchical Materials Design [240, 241] at `http://dx.doi.org/10.18126/xqiz-hzc2`. Input files and scripts necessary for reproducing the simulation trajectories are available in sim.zip file within the Supporting Information. Python 3 Jupyter notebooks implementing our active search procedures are available on GitHub at `https://github.com/KirillShmilovich/ActiveLearningCG`.

Supporting Information

The Supporting Information is available free of charge at `https://pubs.acs.org/doi/10.1021/acs.jpcb.0c00708`.

CHAPTER 3

HYBRID COMPUTATIONAL-EXPERIMENTAL DATA-DRIVEN DESIGN OF SELF-ASSEMBLING π-CONJUGATED PEPTIDES

3.1 Abstract

Biocompatible molecules with electronic functionality provide a promising substrate for biocompatible electronic devices and electronic interfacing with biological systems. Synthetic oligopeptides composed of an aromatic π-core flanked by oligopeptide wings are a class of molecules that can self-assemble in aqueous environments into supramolecular nanoaggregates with emergent optical and electronic activity. We present an integrated computational-experimental pipeline employing all-atom molecular dynamics simulations and experimental UV-visible spectroscopy within an active learning workflow using deep representational learning and multi-objective and multi-fidelity Bayesian optimization to design π-conjugated peptides programmed to self-assemble into elongated pseudo-1D nanoaggregates with a high degree of H-type co-facial stacking of the π-cores. We consider as our design space the 694,982 unique π-conjugated peptides comprising a quaterthiophene π-core flanked by symmetric oligopeptide wings up to five amino acids in length. After sampling only 1181 molecules ($\sim$0.17% of the design space) by computation and 28 ($\sim$0.004%) by experiment, we identify and experimentally validate a diversity of previously unknown high-performing molecules and extract interpretable design rules linking peptide sequence to emergent supramolecular structure and properties.

This chapter is based on the work reported in: **Shmilovich, K.**, Panda, S. S., Stouffer, A., Tovar, J. D., & Ferguson, A. L. (2022). "Hybrid computational–experimental data-driven design of self-assembling π-conjugated peptides." *Digital Discovery*, 1(4), 448-462.

3.2 Introduction

Self-assembling π-conjugated peptides containing a π-core flanked by peptide wings represent a highly tailorable molecular building block for the bottom-up self-assembly of biocompatible supramolecular networks capable of long-range charge transport [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Specific peptide sequences can promote secondary structures within the multi-molecular assemblies akin to beta sheets and guide the quadrupolar association of the π-cores into π-stacked nanostructures with long-range electronic delocalization. Wielding control over the molecular self-assembly of these nanomaterials to tailor the emergent structural, optical, and electronic properties can enable their functional applications as biocompatible, peptide-based field-effect transistors, photoconductors, or solar cells [7, 4, 2, 28]. The availability of 20 distinct natural amino acids and various π-cores make these systems extremely tunable and versatile in their optical and electronic properties [23, 24, 25]. For example, we have previously tuned the steric volume of amino acids directly adjacent to the π-core to engineer tighter or looser packing of the assemblies [24, 26] and controlled the supramolecular chirality of the nanoaggregates by modulating the length of an alkyl spacer between the peptide wings and the π-core [27]. However, most of these materials have been developed by serendipity, intuition, or minor iterative modifications of existing molecules. Systematic data-driven screening approaches present the potential for much deeper and more efficient exploration of sequence space and the discovery of molecules with superior structural and functional properties.

The primary goal of the present work is to discover members of the X_n-quaterthiophene-X_n (X_n-4T-X_n) family of π-conjugated peptides capable of self-assembling into pseudo-1D nanoaggregates with in-register stacking of the quaterthiophene π-cores guided by the X_n peptide wings containing n=1-5 amino acids. Overlap of the π-cores is a structural pre-requisite to supramolecular π electron delocalization and the emergence of charge transport functionality. We chose to explore the quaterthiophene π-core due to its demonstrated ap-

plications in organic electronic field-effect transistors and photovoltaics [44, 45] and also our previous measurements of high charge mobilities in oligopeptide-quaterthiophene conjugates [23]. The peptide wing containing n=1-5 amino acids is denoted as X_n, where each X is one of the 20 natural amino acids. We limit the wing length to a maximum of five residues in order to simplify synthetic manipulation. We place two additional design constraints on the peptide wings. First, we require the oligopeptides to be head-to-tail invariant such that they are chemically symmetric about the quaterthiophene core: one peptide wing is mirror symmetric to the other and each molecule possesses two C-termini. This symmetry imposes parity between the left and right sides of the molecular building blocks to promote the formation of linear supramolecular aggregates. Second, we require that the X_n sequence contain at least one acidic residue (i.e., Asp, Glu) and no basic residues (i.e., Arg, His, Lys). The acidic side-chains together with the two carboxyl C-termini allows us to wield pH control of the oligopeptide protonation state such that they possess a formal negative charge of at least (-4)e at high pH and are formally charge neutral at low pH. It has been previously shown through molecular modeling calculations and fluorescence correlation spectroscopy that intermolecular Coulombic repulsion and enhanced molecular solvation at high pH disfavors assembly and maintains the system as a mixture of monomers and small oligomers [34]. Acidification protonates the ionizable groups to eliminate the Coulombic repulsion and serves as a trigger for large-scale supramolecular assembly. Under these two constraints, the X_n-4T-X_n family comprises 694,982 unique molecules possessing oligopeptide wings containing between n=1-5 amino acids. The design challenge is to discover the members of this design space that self-assemble into the most highly ordered linear nanoaggregates.

The large volume of sequence space means that no more than a tiny fraction of molecular candidates can be experimentally explored due to the time and labor costs associated with oligopeptide synthesis and assays. Edisonian trial-and-improvement experimental search is therefore highly inefficient and limited. Chemical intuition can help focus the search, but

prior knowledge is restricted to a small number of previously studied candidates and also introduces human bias that can impede the discovery of high-performing, non-intuitive solutions. Computational modeling presents a means to conduct high throughput *in silico* screening of molecular space. For example, Frederix et al. identified design rules for the assembly of tripeptide sequences by exhaustively simulating all possible amino acid combinations using coarse grained molecular dynamics simulation [242]. For more complicated molecules and larger molecular search spaces, exhaustive enumeration becomes intractable and it is profitable to combine computational screening with data-driven modeling and active learning. The essence of this approach is to train on-the-fly sequence-property relationships over all computational screening data collected to date and use these models to guide subsequent rounds of the computational screen within a virtuous feedback loop [39, 40, 243, 244, 245]. For example, Li et al. used machine learning algorithms such as random forests, gradient boosting, and logistic regression to predict the assembly and formation of hydrogels from possible peptidic precursors [246]. Nagasawa et al. employed artificial neural networks and random forests for the discovery of conjugated polymers for organic photovoltaic applications [247]. In the context of π-conjugated peptides, we previously combined coarse-grained molecular simulation with deep representational learning and Bayesian optimization to identify molecules predicted to exhibit superior assembly into pseudo-1D linear aggregates [131] and we recently synthesized and tested perylene dimide based peptide-π conjugated materials based on quantitative structure property relation models trained over molecular simulation data [248, 26].

A deficiency of data-driven virtual screening is the weak coupling between computation and experiment. High-throughput virtual screening using computation is used as an initial coarse filtration of the design space that identifies a manageably small number of candidates for synthesis and testing in a subsequent low-throughput experimental screen [249, 250]. The serial nature of this process means that there is no provision to incorporate experimental

feedback into the data-driven search of the design space. This is a lost opportunity since the experimental data can serve as a source of high-quality information to better guide the search and correct for approximations and uncertainties inherent in the computational models. A hybrid data-driven search comprising parallel computational and experimental screens has the potential to offer the best of both worlds – high-throughput approximate computation to achieve broad coverage of the design space and low-throughput experimentation directed towards the most promising candidates. The enabling component of such a procedure is a data-driven model capable of constructing on-the-fly sequence-property relationships from experimental and computational screens that operate asynchronously and in parallel and measure/predict different properties of the molecular system. The trained model is then used within an active learning paradigm to select the most promising molecules for subsequent rounds of computational and experimental screening.

In this work, we develop and deploy a hybrid computational/experimental active learning approach for the data-driven design of X_n-4T-X_n π-conjugated oligopeptides capable of self-assembling into pseudo-1D linear aggregates. We perform high-throughput computational screening using all-atom molecular dynamics simulations that predict the structural morphology of the self-assembled oligopeptide nanoaggregates. We conduct low-throughput experimental oligopeptide synthesis and characterize their assembly using UV-visible spectroscopy. We integrate the computational and experimental screening results to construct on-the-fly sequence-property models that performs asynchronous on-demand selection of the next batch of samples for computational or experimental screening. After sampling only 1181 ($\sim$0.17%) of the 694,982 molecules in the design space by computation and only 28 ($\sim$0.004%) by experiment, we discover and experimentally validate a diversity of previously unknown high-performing oligopeptides capable of spontaneously assembling supramolecular aggregates with a high degree of H-type character and extract interpretable design rules linking peptide sequence to emergent supramolecular structure and properties.

3.3 Methods

In this section, we first present a high-level overview of the hybrid computational/experimental active learning workflow schematically illustrated in Fig. 4.2. We then present the methodological details of each component. For the interested reader, a more comprehensive explication of the theoretical underpinnings of these techniques, discussion of their numerical implementation, and the codes used to conduct this work are provided in the ESI†: Supporting Methods [251].

3.3.1 Overview

We operate a high-throughput computational screening loop to perform all-atom molecular dynamics (MD) simulations to predict the structural morphology of the self-assembled nanoaggregates produced by particular X_n-4T-X_n sequences. The results of the computational screen are used to fit two surrogate sequence-property models relating oligopeptide sequence to the radius of gyration R_g and number of intermolecular contacts κ within the structure. The computational loop seeks to simultaneously maximize κ and R_g to produce pseudo-1D nanoaggregates with in-register π-stacking. We do not *a priori* know the appropriate relative weights of κ and R_g and so adopt an *a posteriori* optimization strategy in which we map out the family of Pareto optimal solutions populating the κ-R_g Pareto frontier [253]. We construct the sequence-property models over a low-dimensional embedding of the molecular design space extracted using regularized autoencoders [254] (RAE) and fit the models using Gaussian process regression [58] (GPR). The GPR predictions are passed to multi-dimensional Bayesian optimization [55, 255] (BO) routines to select the next most promising molecules to simulate for the computational screen.

We simultaneously operate a low-throughput experimental screening loop. In this loop, we synthesize X_n-4T-X_n oligopeptides and characterize their assembly by the blue-shift λ in the mode of the UV-visible spectrogram between the unassembled (high-pH) and assembled

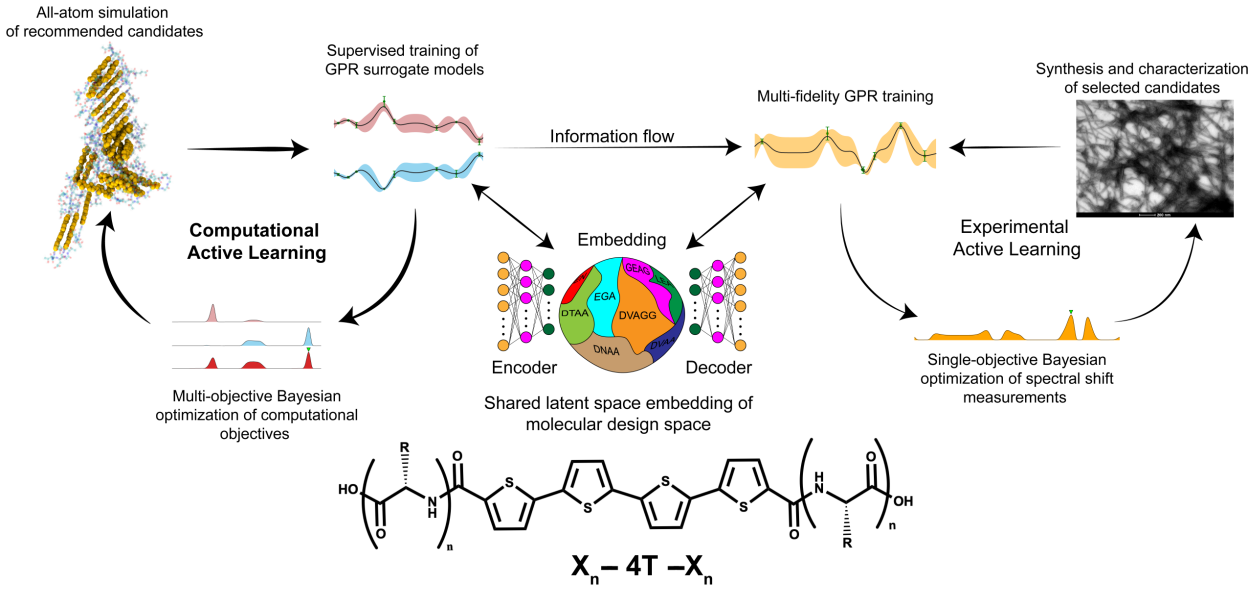


Figure 3.1: Schematic of the hybrid computational/experimental active learning workflow for the discovery of self-assembling X_n-4T-X_n π-conjugated peptides. We perform separate computational and experimental active learning loops within a shared low-dimensional latent space embedding of the molecular design space learned using regularized autoencoders (RAE). Each active learning loop then consists of three parts. (i) Evaluating the quality of a given molecule k in the X_n-4T-X_n design space by either performing high-throughput all-atom molecular dynamics simulations to measure the average number of contacts per molecule $\kappa^{(k)}$ and radius of gyration ${R_g}^{(k)}$ of the self-assembled nanoaggregate or low-throughput experimental synthesis and measurement of the blue-shift $\lambda^{(k)}$ in the mode of the UV-vis spectrum. (ii) Fitting surrogate models using Gaussian process regression (GPR) to predict the performance of untested candidates given the accumulated simulation and experimental data collected to date. Two separate GPRs are maintained for the two computational objectives GPR_κ and GPR_{R_g}. We build a multi-fidelity GPR (mfGPR) as our experimental surrogate model GPR_λ that also incorporates data from the computational GPRs to improve prediction accuracy beyond what would be possible from the limited experimental data alone. (iii) Employing Bayesian optimization (BO) to interrogate the GPR model and select the next most promising molecular candidates for computation as those lying on the κ-R_g Pareto frontier and for experimentation as those with large values of λ. The molecular renderings in this figure, and throughout the paper, are generated using the Visual Molecular Dynamics (VMD) software [252].

(low-pH) states as a quantitative measure of the degree of H-type (i.e., co-facial) aggregation. Again we construct a surrogate sequence-property model relating oligopeptide sequence to the spectral shift where we use the same low-dimensional embedding furnished by the RAE. Importantly, the regression model we use to fit this relationship is a multi-fidelity GPR

(mfGPR) [256] that we train over both the experimental measurements and the computational predictions. Despite measuring different observables of the molecular system – κ and R_g vs. λ – the mfGPR can make use of the voluminous computational predictions to supplement the scarce experimental measurements to furnish a higher accuracy sequence-property model for the spectral shifts than would be possible using the experimental data alone. The computational and experimental screening loops operate asynchronously and in parallel and the data-driven GPR models are continually updated with each new batch of screening data.

The goal of the hybrid computational/experimental screening process is to discover and experimentally validate X_n-4T-X_n molecules with unprecedentedly large values of λ indicative of exceptional in-register π-stacking and H-type character that is a prerequisite for supramolecular electronic delocalization and emergent optical and electronic functionality.

3.3.2 All-atom molecular dynamics simulations

All-atom molecular dynamics simulations of X_n-4T-X_n molecules were conducted using the GROMACS 2019.2 simulation suite [257]. Simulations were initialized in the unassembled state by randomly placing 24 X_n-4T-X_n molecules within a $10\times10\times10$ nm^3 simulation box with three-dimensional periodic boundary conditions and then solvating the system with TIP3P water [258]. Peptides were modeled in the electrically neutral state corresponding to low-pH conditions and treated with the AMBER99SB-ILDN forcefield [259]. The system was relaxed to $T = 300$ K and $P = 1$ bar by steepest descent energy minimization, NVT equilibration [260] and NPT [261] equilibration. We subsequently conducted 200 ns NPT production runs to observe the spontaneous assembly of supramolecular nanoaggregates. Production runs were of sufficient duration that structural metrics of nanoaggregate formation stabilized over the course of the run (Fig. S1 in the ESI†). Simulation snapshots were extracted and saved every 1 ps for analysis.

The fitness of a particular X_n-4T-X_n molecule was defined within our molecular simula-

tions based on its ability to form pseudo-1D linear nanoaggregates with in-register stacking of the π-cores. We quantified this structurally via the average number of contacts per molecule κ and the radius of gyration R_g [262] of the self-assembled nanoaggregates averaged over the terminal 50 ns of our production runs. An intermolecular contact is defined according to our previously reported "optical distance" $d_{i,j}^{optical}$ that measures the minimum intermolecular distance between any pair of thiophene rings within the π-cores of molecules i and j [210, 37, 36, 131]. We have previously shown that adopting a threshold of $d_{i,j}^{optical} < 0.7$ nm by which to define an intermolecular contact assures close proximity and in-register π-stacking between at least one pair of thiophene rings [36, 37, 131, 38]. Our computational active learning loop ultimately aims to explore the Pareto frontier of molecules and discover well assembling π-conjugated peptides that possess good thiophene π-π stacking (i.e., large κ) within high-aspect ratio linear nanoaggregates (i.e., large R_g). In general, we found maximization of either objective function alone was insufficient to promote in-register stacked pseudo-1D nanoaggregates: high κ in the absence of high R_g corresponds to globular structures with promiscuous multi-molecular π-stacking, whereas high R_g in the absence of high κ corresponds to weakly associated elongated threads lacking π-core overlaps. As we will show, the hybrid computational/experimental active learning framework learns the appropriate balance between κ and R_g that is most predictive of high-performing experimental candidates with large values of λ.

3.3.3 Chemical space embedding

Each X_n-4T-X_n molecule in our design space is differentiated by the identity of the peptide wing containing between one and five amino acids. We represent each candidate molecule as a linear amino acid graph where each node is an amino acid and the edges reflect their linear connectivity (Fig. S2 in the ESI†). The molecular design space of 694,982 molecules is large, discrete, and high-dimensional. It is possible to perform active learning directly

over this space using, for example, kernel models [263, 264, 265, 266, 267, 268, 39], but superior search efficiencies can be achieved by first projecting the molecular design space into a smooth, continuous, low-dimensional space that is more amenable to the construction of robust regression models and deployment of optimization algorithms [131, 244]. We learn a bespoke latent space for the X_n-4T-X_n family using a regularized autoencoder [254] (RAE) trained on our entire molecular design space (Fig. S3 in the ESI†). An RAE is a deterministic variant of the popular variational autoencoder (VAE) architecture [269] that possesses the potential advantage of enabling more expressive flexibility in the data distribution within the low-dimensional embedding by not assuming or enforcing a prior (generally Gaussian) distribution over the latent dimensions. We do not perform a training/validation/testing split since the purpose of our RAE model is to provide a low-dimensional latent space embedding of the complete design space of all 694,982 candidate molecules for downstream active learning, not as a predictive tool for out-of-sample inference beyond this design space. As such, we do not use the decoder to generate new candidate molecules and it is therefore not a disadvantage for our applications that the latent distribution is not approximately Gaussian distributed, which is generally desirable for efficient generative sampling. We have previously demonstrated an application of RAEs in the data-driven design of small drug-like cardiolipin-selective molecules [132]. This encoder-decoder architecture consists of a message passing neural network [270, 271] encoder and a decoder performing explicit graph matching to ensure end-to-end invariance of node permutations. The encoder [272, 273, 274] accepts graphical representations of the X_n peptide sequences the nodes of which are featurized by the 553 single amino acid physiochemical properties contained in the AAindex database and the edges of which are featurized by the 135 pairwise amino acid contact potentials and mutation matrices [275]. The bottleneck layer defining the interface between the trained encoder and decoder contains a nonlinear latent embedding of the molecular design space $\boldsymbol{\xi}$ learned by the encoder that preserves relationships between molecular candidates and

is sufficiently informative for the decoder to accurately reconstruct the molecular graphs (Fig. S4 in the ESI†). We determine an appropriate dimensionality d=32 for the latent space using exploratory hyperparameter tuning optimizing the RAE reconstruction loss [217, 276]. This low-dimensional latent space defines a smooth and continuous representation $\boldsymbol{\xi}$ of the molecular design space over which fit (mf)GPR surrogate models and conduct BO-enabled active learning.

3.3.4 Computational active learning

The goal of the computational active learning loop is to drive computational discovery of candidate X_n-4T-X_n molecules that simultaneously maximize the average number of contacts per molecule κ and the radius of gyration R_g of the self-assembled aggregates. This amounts to searching for molecules residing upon the κ-R_g Pareto frontier, for which the dual objective functions necessitate a multi-objective optimization strategy. We seeded the active learning screen by conducting MD simulations of 228 initial candidate molecules. To ensure broad initial sampling of the candidate space, we selected 100 molecules as those residing closest to the centroids of a 100-cluster k-means partitioning of the RAE latent space. The remaining 128 molecules were hand-selected to comprise a diversity of peptide wing lengths, residue hydrophobicity, aromaticity, polarity, and presence of heteroatoms. We used the computational predictions of κ and R_g to train two independent Gaussian process regression (GPR) surrogate models [58] $\hat{\kappa} = f(\boldsymbol{\xi})$ and $\hat{R_g} = g(\boldsymbol{\xi})$ that perform supervised learning of mappings from the latent space coordinates to the two structural observables. The predictions of the two GPRs are passed to a multi-objective BO routine that seeks to simultaneously maximize κ and R_g using the method of random scalarizations [255, 55]. The trained GPRs for κ and R_g are used to construct two independent BO upper confidence bound (UCB) [277] acquisition functions defining the relative desirability of each candidate X_n-4T-X_n molecule in maximizing each of these two objectives. We then collapse these two

acquisition functions into a single scalarized acquisition function constructed as a randomly weighted linear sum. The scalarized acquisition function is then used to perform univariate Bayesian optimization. A particular random scalarization corresponds to a particular choice of relative weightings between the two design objectives and defines a vector within the 2D κ-R_g space along which to maximize. Under sufficiently many repeated random scalarizations, the random vectors span the κ-R_g Pareto frontier to discover a family of Pareto optimal solutions. We perform batched selection over different random scalarizations and choices of the UCB hyperparameter balancing the BO exploit-explore tradeoff to propose 25 new candidate molecules per round. MD simulations of these 25 molecules are performed and the three part active learning cycle is iteratively repeated. We assess convergence of the iterative screen by monitoring the set of Pareto optimal points that define the κ-R_g Pareto frontier and terminate sampling once the Pareto frontier ceases to advance with additional rounds of sampling. We conducted 38 rounds of computational screening over which we considered 1181 candidate molecules. A full accounting of the molecules identified in each round of the computational active learning loop is provided in the Data Availability statement [251].

3.3.5 Experimental active learning

The experimental active learning loop aims to maximize the blue-shift λ in the mode of the UV-visible spectrogram between the unassembled (high-pH) and assembled (low-pH) states. The magnitude of this spectral shift λ has been experimentally shown to correlate with co-facial H-type assembly that results in formation of the desired pseudo-1D linear stacks [46, 47, 23]. We seeded the experimental search with a set of 11 molecules hand selected from the 228 molecules comprising the initial computational round to comprise a diversity of oligopeptide wing lengths and predicted values of κ and R_g. We note that these 11 initial molecules originated from the subset of 128 human-selected molecules from our initial computational round, rather than from the subset of 100 molecules identified

via k-means clustering. We trained a GPR model $\hat{\lambda} = h(\boldsymbol{\xi})$ to predict the spectral shift λ as a function of latent space coordinates. In this case we have only a single objective function λ but we wish to construct a multi-fidelity surrogate model incorporating both direct experimental measurements of λ and computational predictions of κ and R_g. The rationale is that the computational predictions for κ and R_g should be correlated with and predictive of the experimental measurements of λ. This is expected to be the case since κ and R_g are structural measures of the degree of in-register π-stacking in elongated nanoaggregates that are prerequisites for H-aggregate character manifested in measurements of λ. A multi-fidelity model trained to learn a nonlinear mapping from the low-fidelity computational predictions to high-fidelity experimental measurements can take advantage of the abundant computational data to produce a superior model than that obtained by training over only the sparse experimental data alone. Indeed, by the terminal round of experimental active learning, incorporation of computational screening data within the multi-fidelity paradigm leads to a $\sim$27% improvement in the predictive accuracy of our surrogate model compared to a single-fidelity model fitted only over the experimental observations (Fig. S5 in the ESI†). This predictive improvement highlights the capabilities of our mfGPR to leverage plentiful low-fidelity data from our computational κ and R_g metrics to improve our predictive performance for our target high-fidelity objective in λ. We observe that the flexibility of the mfGPR framework enables this information transfer even in situations where the low- and high-fidelity observables are measuring different properties, are only weakly correlated or even anti-correlated, and where the degree of correlation may change over the domain [278]. The only requirement is that the low-fidelity response surface is informative of the high-fidelity response surface and that this relationship can be extracted from the data within the mfGPR model.

We construct multi-fidelity surrogate models fusing the computational (low-fidelity) and experimental (high-fidelity) data using the multi-fidelity Gaussian process regression (mfGPR)

formalism [256]. The mfGPR model is then passed to a standard BO routine [55] employing an expected improvement (EI) acquisition function [55, 67, 65] and the Kringing believer [279] batched sampling. We use the BO to propose a batch of molecules for the next round of sampling, which we manually down-select to 8-9 molecules for experimental synthesis and characterization. By incorporating "human-in-the-loop" curation of the selected molecules we hope to balance purely data-driven candidate proposal with chemical intuition and thereby incorporate some degree of prior knowledge and human experience into the search process without, we hope, imposing too much bias on the search. The success of this collaborative human-machine paradigm has been previously demonstrated in the data-driven discovery of molecular organic light emitting diodes [250]. The experimental measurements are fed back into the low-throughput experimental active learning loop that is executed asynchronously and in parallel with the high-throughput computational loop. We execute three rounds of experimental active learning over the course of the 38 rounds of computational active learning that are executed at computational rounds 0, 14, and 22. Given the good performance of the candidates studied in the third experimental round together with the relatively modest advances in the Pareto frontier observed over computational rounds 23-38, we elected to terminate our experimental screen after its third round. A full accounting of the molecules identified in each round of the experimental active learning loop is provided in Table 3.1 and have been made available as detailed in the Data Availability statement [251].

3.3.6 Nonlinear manifold learning of low-dimensional assembly pathways

After completing the hybrid computational/experimental screen we subjected the ensemble of 1181 molecular simulation trajectories of X_n-4T-X_n candidate molecules to nonlinear dimensionality reduction in order to resolve the structural assembly pathways. In doing so, we sought to gain mechanistic understanding of the molecular assembly mechanisms differentiating the top performing molecules identified by our screen. We performed nonlinear

dimensionality reduction using diffusion maps manifold learning [280, 281] to project the configurational coordinate space into a low-dimensional space preserving the leading high-variance collective dynamics of the system [131, 221, 222, 223, 210, 282]. Diffusion maps take as an input a pairwise distance matrix measuring the configurational similarity between all 118,100 simulation snapshots harvested from the ensemble of 1181 simulation trajectories. We define these pairwise distances using the smooth overlap of atomic positions (SOAP) kernel [283, 284, 285, 286] between the heavy atoms constituting the 4T π-cores as a distance metric that is naturally invariant to rotations, translations, and permutations of atoms, and which – as a π-core-centric metric – can be applied between oligopeptides with different wing lengths. The influence of the wings is implicitly retained through their impact on the configurations adopted by the π-cores. A density-adaptive variant of diffusion maps [231] is then applied to furnish embeddings of the assembly trajectories into a 2D manifold that exposes the assembly pathways and mechanisms followed by the various X_n-4T-X_n molecules. We note that our diffusion maps furnish a low-dimensional embedding of the *configurational coordinate space* traversed by our MD simulations, and that this embedding is completely independent of the low-dimensional embedding of the *molecular design space* furnished by the RAE that is employed within our active learning protocol.

3.3.7 Oligopeptide synthesis

General Information. *N*,*N*-Dimethylformamide (DMF) was purchased from Sigma-Aldrich. *N*-Methyl-2-pyrrolidone (NMP) was obtained from Advanced ChemTech. Dichloromethane (DCM) and *n*-hexane were freshly distilled prior to storage. All solvents were stored over 4 Å molecular sieves and were subsequently degassed by sparging with nitrogen gas at least 30 min prior to use. O-(Benzotriazol-1-yl)-*N*,*N*,*N*′,*N*′-tetramethyluronium hexafluorophosphate (HBTU) was purchased from Oakwood Products Inc. Tetrakis(triphenylphosphine)palladium was obtained from Strem Chemicals. Wang resin (preloaded with amino acid) and Fmoc-

protected amino acids were obtained from Advanced Chem Tech. 5-Bromo-2-thiophenecarboxylic acid was obtained from Accela Chem Bio Co. Ltd. All other reagents and starting materials were obtained from Sigma-Aldrich and were used as received. Details of the synthesis for individual oligopeptides is provided in the ESI†: Peptide Synthesis.

Electrospray Ionization Mass Spectrometry (ESI-MS). ESI samples were collected using a Thermo Finnigan LCQ Deca Ion Trap Mass Spectrometer in negative mode. Samples were prepared in a 1:1 MeOH:water solution with 1% ammonium hydroxide. ESI spectra for each synthesized peptide are provided in the ESI†: ESI Spectra.

Reverse-Phase HPLC. HPLC purification was performed on an Agilent 1100 series (semipreparative/analytical) and a Varian PrepStar SD-1 (preparative) instrument using Luna 5 μm particle diameter C8 with TMS end-capping columns with silica solid support. An ammonium formate aqueous buffer (pH 8) and acetonitrile were used as the mobile phase. HPLC traces for each synthesized peptide are provided in the ESI†: Analytical HPLC Traces.

General Solid-Phase Peptide Synthesis (SPPS). All peptides were synthesized using the standard Fmoc solid-phase technique with Wang resin preloaded with Fmoc-protected amino acids. To the resin in a peptide chamber, Fmoc-deprotection was accomplished by adding a 20% piperidine solution in DMF twice (successive 5- and 10-min treatments) followed by washing with NMP $\times$ 3, methanol $\times$ 3, and DCM $\times$ 3. For the amino acid couplings, 3.0 equiv of the Fmoc-protected amino acid was activated with 2.9 equiv of HBTU and 10 equiv of diisopropylethylamine (DIPEA) in NMP, and this solution was added to the resin beads. The reaction mixture was allowed to mix for 45-60 min, after which the beads were rinsed with NMP, methanol, and DCM (3 times each). The completion of all couplings was monitored using a Kaiser test on a few dry resin beads, repeating the same amino acid coupling if needed. The general procedure for amino acid coupling was repeated for each additional amino acid until the desired peptide sequence was obtained.

General *N*-Acylation Procedure for Peptides. Following our previous proce-

dure [27], a solution containing 3 equiv of 5-Bromo-2- thiophenecarboxylic acid, HBTU (2.9 equiv), and DIPEA (10 equiv) in NMP was mixed with the oligopeptide-bound resin for 3 h. The completion of the *N*-acylation was assessed using a Kaiser test on a few dry resin beads. The resin was washed with NMP, methanol, and DCM (3 times each).

General on-Resin Stille Coupling Procedure. The solid-supported *N*-acylated oligopeptide (1 equiv) was transferred to a Schlenk flask equipped with a reflux condenser. The resin was dried under vacuum. $Pd(PPh_3)_4$ (4 mol %, relative to resin loading) was added to the reaction vessel. An approximately 15 mM solution of the 5,5′-bis-trimethylstannyl-[2,2′]-bithiophene (0.50 equiv) was prepared in DMF and was added to the reaction flask via syringe. The mixture was heated to 80°C for 18 h and was agitated constantly by bubbling nitrogen through the solution. The mixture was allowed to cool to room temperature. The peptide was subjected to the general cleavage and workup procedure to yield the crude product and then further purified by HPLC.

General Peptide Cleavage and Workup Procedure. Following dimerization, the resin was returned to the peptide chamber and again subjected to a wash cycle: 2×NMP, 2×methanol, and 2×DCM. The resin was then treated with 9.5 mL of trifluoroacetic acid, 250 μL of water, and 250 μL of triisopropylsilane for 3 h. The peptide solution was filtered from the resin beads, washed three times with DCM, and concentrated by evaporation under reduced pressure. The crude peptide was then precipitated from the solution with 40-50 mL of diethyl ether and isolated through centrifugation. The resulting pellet was triturated with diethyl ether to yield the crude product, which was dissolved in approximately 20-25 mL of water. 30 μL of potassium hydroxide (1 M) was added if needed to solubilize the peptides in water and lyophilized.

3.3.8 UV-visible spectroscopy

UV-vis spectra were obtained using a Varian Cary 50 Bio UV-vis spectrophotometer. Spectroscopic samples were prepared by diluting the peptide solution to the appropriate concentration in Millipore water to achieve an optical density near 0.1, 0.2, and 0.3 in the monomeric/basic solution. The pH was then adjusted by adding 20 μL of 1M KOH (basic) followed by addition of 40 μL of 1M HCl (acidic). Approximate concentration of the peptides were 2.25 μM, 4.50 μM, and 6.75 μM for optical density of 0.1, 0.2, and 0.3 respectively.

3.4 Results and Discussion

3.4.1 Hybrid computational/experimental active learning discovers novel high-performing oligopeptides

We report in Fig. 4.4 the results of our hybrid computational/experiment active learning screen to the molecular design space of 694,982 X_n-4T-X_n candidate molecules. We conduct 38 rounds of computational screening to simulate a total of 1181 X_n-4T-X_n molecules interleaved with three rounds of experimental screening in which we synthesize and test a total of 28 molecules. A full accounting of the molecules identified in each round of the computational and experimental active learning loops are provided as detailed in the Data Availability statement [251]. Round 0 of the computational and experimental screens commence simultaneously, respectively screening 228 and 11 molecules. The two screens then proceed asynchronously and in parallel. The high-throughput computational loop considers 25 candidate molecules per round and iterates more rapidly than the low-throughput experimental screen that considers 8-9 molecules per round. We track the progress of the experimental screen via the measured spectral blue shifts λ upon assembly as a measure of the prevalence of H-type co-facial π-stacking (Fig. 4.4a). We track the progress of the computational screen via the advancement in the κ-R_g Pareto frontier that we quantify through

the mean distance from the origin of all n_i molecules cumulatively simulated over the first i rounds (Fig. 4.4a,b),

$$d_{\text{Pareto}}^{(i)} = \frac{1}{n_i} \sum_{k=1}^{n_i} \sqrt{(\kappa^{(k)})^2 + (R_g^{(k)})^2}. \tag{3.1}$$

We recall that large values of the average number of π-core contacts per molecule κ and radius of gyration of the self-assembled aggregates R_g are anticipated to lead to correlate with the formation of pseudo-1D nanoaggregates with high degrees of H-type character.

Round 1 of the experimental screen is performed upon completing computational Round 14, at which time a total of 578 candidate molecules have been computationally assessed. These computational screening data are passed to the experimental surrogate model and Bayesian optimization within the multi-fidelity hybrid computational/experimental active learning framework in order to better inform the design of experimental Round 1 than would be possible by analyzing the 11 experimental data points alone. Under our human-in-the-loop selection protocol, we selected nine molecules for Round 1 of experimental screening by filtering a 35-molecule list outputted from our BO routine. Down-selection was performed on the basis of high anticipated performance based on our previous experimental and computational work [131, 24] and maintenance of a diversity of peptide wing compositions and lengths. This human-in-the-loop selection process serves as a simple means to inject prior knowledge into the data-driven search process, which can be particularly valuable in the early stages of the search where the models are trained over small numbers of data points, by directing the search process to regions of molecular design space that are anticipated to be particularly profitable.

Round 2 of the experimental screen is conducted after completing computational Round 22, at which point we have simulated 780 candidate molecules. Again, the totality of these computational screening data are used to augment the multi-fidelity experimental surrogate model and used to pick eight candidate molecules for experimental testing down-selected from the 75 top candidates proposed by the BO routine. We inject one addition piece of human

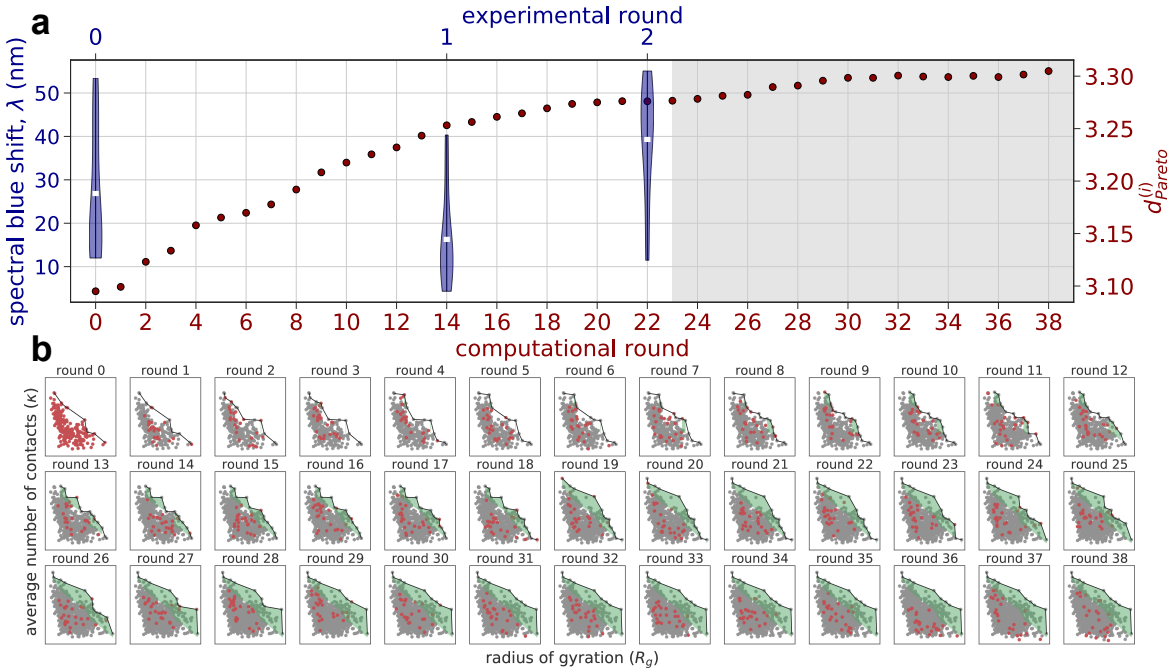


Figure 3.2: Progress and convergence of the hybrid computational/experimental active learning screen for high-performing X_n-4T-X_n molecules within the design space of 694,982 candidates. (a) A total of 38 computational rounds of screening are performed and interleaved with three rounds of experimental screening. The experimental screens are conducted at computational Rounds, 0, 14, and 22. Convergence of the computational screen is tracked by $d^{(i)}_{\text{Pareto}}$, the mean distance from the origin of all molecules cumulatively simulated over the first i rounds within the computational κ-R_g Pareto plot (see panel b), as a measure of the advancement of the optimal frontier (red circles). Convergence of the experimental screen is tracked by the measured spectral blue shift λ quantifying the degree of H-type cofacial π-stacking within the self-assembled nanoaggregates (blue violin plots). Violin plots enclose the range of λ values measured each experimental round and the central white tick denotes the mean λ value for the round. The gray shaded area represents the computational rounds used to verify convergence of the active learning but not used to inform any additional rounds of experimental design. (b) Round-by-round advancement of the computational κ-R_g Pareto frontier over the course of the 38 screening rounds. Within each frame, the points corresponding to X_n-4T-X_n candidates collected within that round are shown in red, those points collected in previous rounds are shown in gray, the Pareto frontier defined by the Pareto optimal points is shown as a black solid line, and the shaded green area indicates the advancement of the Pareto frontier relative to Round 0.

intuition into the down-selection process by making a single Tyr to Ala amino acid mutation of one of the predicted sequences – YEVGA to AEVGA – based on prior understanding that aromatic side chains are known to π-stack with the π-cores and therefore liable to disrupt

favorable in-register 4T stacking [131]. This modification is substantiated by both observing low-ranking candidates possessing bulky aromatic residues in the first two experimental rounds (EYIQG: rank 18/28, VEF: rank 23/28, GFGFD: rank 25/28) along with previous experimental work noting the presence of aromatic residues resulting in reduced UV-vis blue-shifts [24].

We continue to conduct an additional 16 rounds of computational screening (Rounds 23-38) while experimental Round 2 is being completed in anticipation of possibly conducting a fourth experimental round. As illustrated in Fig. 4.4a, we observe a relatively rapid expansion of the computational exploration of the κ-R_g design space over the course of computational Rounds 0-20, which we quantify by $d_{\text{Pareto}}^{(i)}$ defining the mean distance from the origin of all molecules cumulatively simulated over the first i rounds. This trend, however, plateaus at $d_{\text{Pareto}} \approx 3.3$ by Round 29 and exhibits only a 0.7% increase in d_{Pareto} relative to Round 22. This observation is mirrored by a relatively modest advancement of the Pareto frontier between Rounds 23-38 (Fig. 4.4b). Experimentally, the mean spectral shift λ in experimental Round 2 is 46% better than Round 0, and the top performing Round 2 candidate has a 3% better spectral shift compared to that in Round 0. The diminishing returns evinced by the computational d_{Pareto} and the successful discovery of a molecule with superior λ impelled us to terminate our search after experimental Round 2 / computational Round 38.

In all, we simulated 1181 molecules comprising $\sim$0.17% of the 694,982 molecules constituting the X_n-4T-X_n design space, corresponding to 236.2 μs of simulation time, and requiring $\sim$4.97 GPU-years of parallel compute. Experimentally, we synthesized and characterized a total of 28 X_n-4T-X_n molecules over the course of the course of eight months corresponding to exploration of 0.004% of the molecular space.

We present in Fig. 3.3 an embedding of all 1181 simulated molecules and 28 experimentally tested molecules into the κ-R_g objective function space used to identify high-performing molecules in the computational screening loop. Molecular renderings of the self-assembled

nanoaggregates provides qualitative visual conformation that X_n-4T-X_n molecules producing aggregates with large values of both κ and R_g do indeed tend to self-assemble into pseudo-1D structures with good stacking of the 4T π-cores.

The primary result of our hybrid computational/experimental active learning screen are experimental measurements of 28 X_n-4T-X_n molecules reported in Table 3.1. The 11 Round 0 molecules were selected based on human intuition and used to seed the experimental active learning. The nine Round 1 and eight Round 2 molecules are the result of our multi-fidelity active learning search. Of these molecules, 26 are completely novel and on par with known high-performing sequences [47, 190, 23] while also possessing greater diversity amino acid sequences previously unknown to correlate with good spectral blue shifts λ. The high values of λ for these molecules are indicative of a high degree of H-type co-facial stacking and the potential for long-range supramolecular electronic delocalization and emergent optoelectronic functionality. Additionally, we were encouraged that our active learning procedure spontaneously discovered DGG-4T-GGD as a previously known high-performing candidate [47, 190]. A complete list of predicted κ and R_g values from the terminal computational surrogate model and predicted spectral shift measurements λ from the terminal experimental surrogate model for all 694,982 molecules within the X_n-4T-X_n design space have been made available as detailed in the Data Availability statement [251].

3.4.2 Molecular design rules

Our rank-ordered list of 28 experimentally assayed candidates exposes a number of oligopeptide design precepts, that is relationships between the placement/omission of particular amino acids at specific positions along the oligopeptide wing and the magnitude of the spectral blue shift λ quantifying the degree of H-type co-facial association within the self-assembled nanoaggregates. Despite the relatively small size of the experimental data set, we were able to extract three statistically significant design rules. First, the nine top-ranked

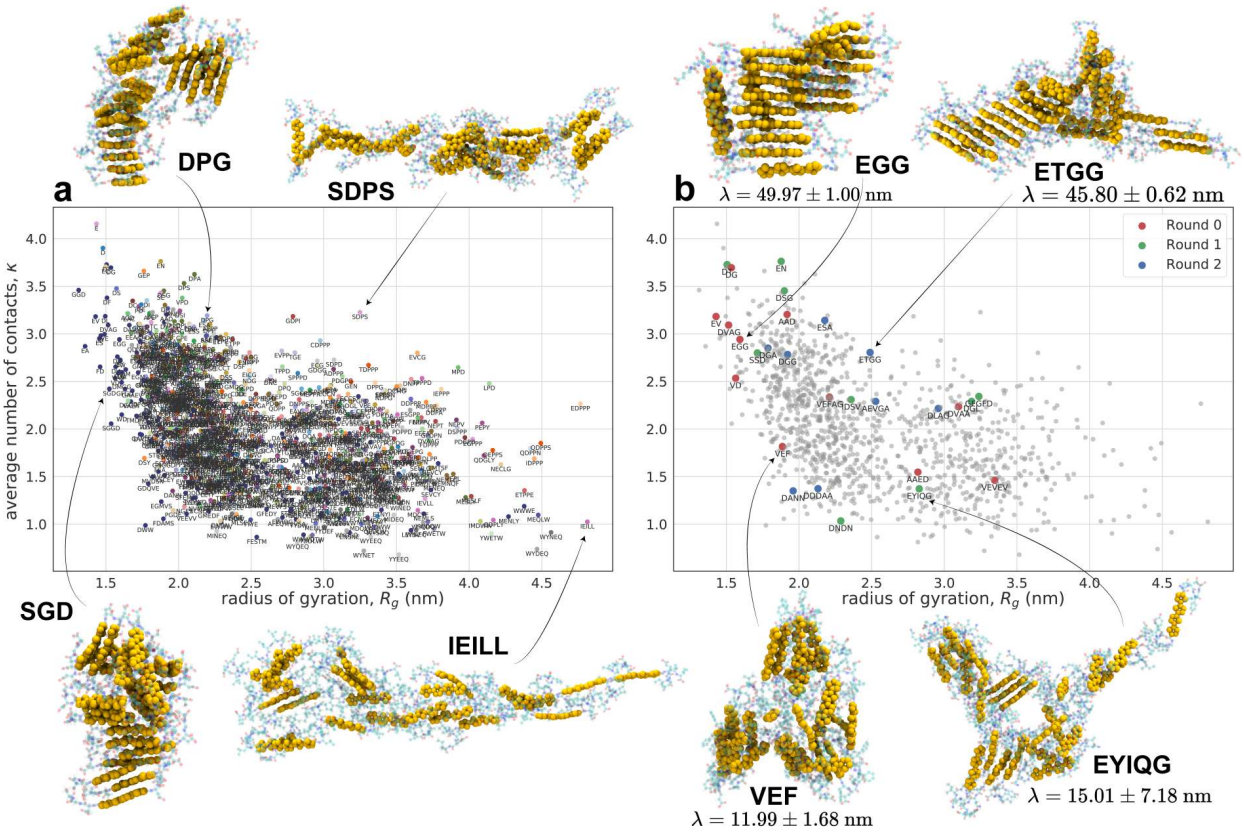


Figure 3.3: Embedding of the computational and experimental molecules sampled in the active learning screen into the κ-R_g objective function space. (a) Embedding of the 1181 X_n-4T-X_n molecules explored in the computational screen. Points shown in the same color were sampled during the same computational active learning round. The labels associated with each point corresponds to the X_n peptide wing sequence. (b) Highlighting the 28 experimentally tested X_n-4T-X_n molecules (colored points) superposed onto all 1181 computationally simulated points (grey points). The color indicates the round of experimental active learning within which the molecule was tested. Encircling the plot are snapshots from our molecular simulation trajectories showing the terminal self-assembled nanoaggregates. The heavy atoms constituting 4T π-cores are rendered as gold space-filling spheres and the X_n peptide wings as semi-transparent ball-and-stick representations. Molecules producing aggregates with large values of both κ and R_g tend to self-assemble into pseudo-1D structures with good stacking of the 4T π-cores. For the experimentally tested candidates we also report the measured values of the spectral blue shift λ. A full accounting of the computed κ and R_g values and measured λ values for all molecules considered in our screen have been made available as detailed in the Data Availability statement [251].

Table 3.1: Rank-ordered list of the 28 experimentally tested X_n-4T-X_n molecules sampled over the course of the active learning screen.

Rank	Peptide wing, X_n	Measured spectral shift, λ (nm)	Discovery round	Previously known?
1	DGG	55.06 ± 1.00	2	Yes [47, 190]
2	DG	53.42 ± 4.16	0	N
3	ESA	50.01 ± 0.53	2	N
4	EGG	49.97 ± 1.00	0	Yes [23]
5	ETGG	45.80 ± 0.62	2	N
6	DGA	44.15 ± 1.49	2	N
7	DDDAA	42.07 ± 0.46	2	N
8	DVAA	41.65 ± 1.15	0	N
9	DSG	40.32 ± 1.15	1	N
10	AEVGA	40.15 ± 1.12	2	N
11	DVAG	35.70 ± 1.49	0	N
12	DNDN	29.58 ± 4.76	1	N
13	DANN	25.70 ± 0.32	2	N
14	VEFAG	21.75 ± 2.07	0	N
15	VEVEV	18.43 ± 0.62	0	N
16	VD	18.02 ± 2.66	0	N
17	AAD	15.98 ± 1.00	0	N
18	EYIQG	15.01 ± 7.18	1	N
19	EV	14.70 ± 1.20	0	N
20	DT	14.70 ± 1.20	1	N
21	AAED	13.68 ± 1.46	0	N
22	SSD	13.68 ± 1.15	1	N
23	VEF	11.99 ± 1.68	0	N
24	DLAG	11.49 ± 0.46	2	N
25	GFGFD	10.97 ± 1.77	1	N
26	DGL	10.25 ± 1.20	1	N
27	IDSV	7.70 ± 3.83	1	N
28	EN	4.33 ± 1.49	1	N

molecules within the 28 assayed candidates contain a distal Asp or Glu residue at the C-terminus and a Gly or Ala residue in the position most proximate to the π-core. A statistical analysis using a one-sided Mann-Whitney U test [287] reveals a statistically significant (p=0.0001) increase in the measured blue shifts λ associated with the presence of the (D/E)X_n(A/G) motif. In our prior work on π-conjugated oligopeptides, we typically synthesized the peptide wings with the ionizable residue responsible for actuating pH-triggered assembly located at the C-terminus to locate it as far away as possible from the π-core: our

motivation for this design choice was that the hydrophilic and polar nature of these residues which, together with their steric bulk, was anticipated to disrupt good supramolecular assembly of the π-cores [288, 23, 47, 24, 49, 234, 289]. Interestingly, our active learning screen appears to have also learned this design rule without any explicit human instruction and furnished *post hoc* support for this intuitive choice. Similarly, our recent computational and experimental work [131, 248, 290]. is consistent with prior chemical intuition [288] that the placement of small non-polar residues adjacent to the π-core should promote good co-facial stacking of the cores. Again, the active learning screen appears to have also learned this design rule within the sequence-property surrogate model and is consistent with a physical rationale that the small steric volume of these amino acids is conducive to in-register π-stacking of the 4T cores. It is more challenging, however, to differentiate between the performance of Gly vs. Ala, with Gly leading to more favorable spectral shifts within a DGX motif – λ_{DGG} = (55.05±1.00) nm and λ_{DGA} = (44.15±1.49) nm – whereas Ala performs better within a DVAX motif – λ_{DVAG} = (35.70±1.49) nm and λ_{DVAA} = (41.65±1.15) nm. The absence of a simple modular decomposition of the influence of each amino acid position in the X_n wing reflects the complexity of the self-assembly process and the important role of multi-body interactions, amino acid context, and wing length.

Second, consistent with the favorability of core-adjacent Gly and Ala residues, the non-C-terminal amino acids within the X_n peptide sequences of the top-performing molecules tend to also be enriched in small hydrophobic residues such as Ala, Gly or Val (one-sided Mann-Whitney U test, p-value=0.004). Interestingly, residues containing polar hydroxyl groups such as Ser and Thr are also over-represented within high-performing sequences when Ser or Thr are non-terminal residues and Asp or Glu are C-terminal such as ESA: rank 3/28, ETGG: rank 5/28, and DSG: rank 9/28 (one-sided Mann-Whitney U test, p-value=0.03). Other polar residues like Asn also perform relatively well in the π-core proximate position when Asp occupies the distal slot (DNDN: rank 12/28; DANN: rank 13/28).

Third, the presence of larger hydrophobic and bulky aromatic residues such as Leu, Ile, Phe, and Tyr at any location are correlated with poorer performing candidates such as DLAG: rank 24/28, DGL: rank 26/28, EYIQG: rank 18/28, and IDSV: rank 27/28, with the poorest-performing candidates enriched in these four amino acids (one-sided Mann-Whitney U test, p-value=0.002). We have previously shown that favorable interactions between these large hydrophobic residues and the aromatic π-cores can disrupt supramolecular association between the cores [131]. This observation is also consistent with our prior observation that oligopeptides possessing oligophenylenevinylene π-cores exhibited larger spectral blue shifts when the peptide wings contained small Gly and Ala residues compared to larger Ile and Val residues [24].

Finally, although no Pro containing molecules were sampled in the experimental screen, we note that a number of simulated molecules containing a Pro residue lie at or near the κ-R_g Pareto frontier and were highly ranked in predicted blue shift λ by the terminal mfGPR surrogate model (e.g., DPG: rank 329/694,982, AEPP: rank 183/694,982, SDPD: rank 10/694,982, EAP: rank 13/694,982, DDPA: rank 23/694,982, GEPG: rank 15/694,982). This finding is somewhat surprising because Pro has been largely unexplored in previous experimental and computational π-conjugated peptide studies. Proline, with its unique conformational properties including its conformational rigidity and absence of hydrogen bond donor capacity, appears to be quite favorable in promoting good in-register stacking between the π-cores and the formation of high-aspect ratio nanoaggregates. We suggest that experimental testing and further computational exploration of the molecular mechanisms underpinning these predictions may be a profitable avenue for future investigations.

3.4.3 Molecular assembly pathways

Having extracted design rules linking amino acid sequence to the degree of H-type stacking within the nanoaggregates, we then sought to analyze our library of molecular simulation tra-

jectories to resolve the molecular self-assembly pathways promoted by the high-performing peptide sequences to gain mechanistic understanding of the link between sequence and the emergent supramolecular structure. We hypothesized that the ensemble of simulation trajectories for 1181 different X_n-4T-X_n molecules collected over the course of our computational screen may admit a low-dimensional clustering within the configurational phase space of assembly pathways, and that the high-performing molecules should follow similar assembly pathways to reach the terminal pseudo-1D nanoaggregates. We report in Fig. 3.4a-b a 2D embedding into the leading two collective variables $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ discovered by diffusion maps. By correlating these collective variables with candidate physical observables, we find $\boldsymbol{\psi}_2$ to be strongly correlated with the instantaneous radius of gyration R_g of the system ($\rho(\boldsymbol{\psi}_2, R_g) = 0.93$) and $\boldsymbol{\psi}_3$ moderately strongly correlated with the instantaneous number of contacts per molecule κ ($\rho(\boldsymbol{\psi}_3, \kappa) = 0.69$). In addition to providing good physical interpretability of the low-dimensional manifold containing the molecular assembly pathways learned by diffusion maps, the emergence of two collective variables strongly correlated with κ and R_g provides *post hoc* support for our selection of two observables as the dual objective functions of our computational screen as the leading variables governing the long-time self-assembly dynamics.

In Fig. 3.4c-f we illustrate the temporal evolution of the self-assembly pathways for particular X_n-4T-X_n molecules over the $\boldsymbol{\psi}_2 - \boldsymbol{\psi}_3$ manifold. Each molecular trajectory begins at the right-most edge of the of the manifold corresponding to the initial monodisperse state. Lateral leftward movement across the manifold correspond to condensation of the system to smaller R_g values due to the formation of nanoaggregates. Vertical upward movement corresponds to the accumulation of inter-molecular contacts between the π-cores and an elevation in κ. The assembly pathways of the top-performing candidates typically terminate in the upper-left corner of the manifold that contains pseudo-1D nanoaggreates containing $\kappa \approx 3$ intermolecular contacts and $R_g \approx 2$ nm corresponding to elongated linear stacks. We observe

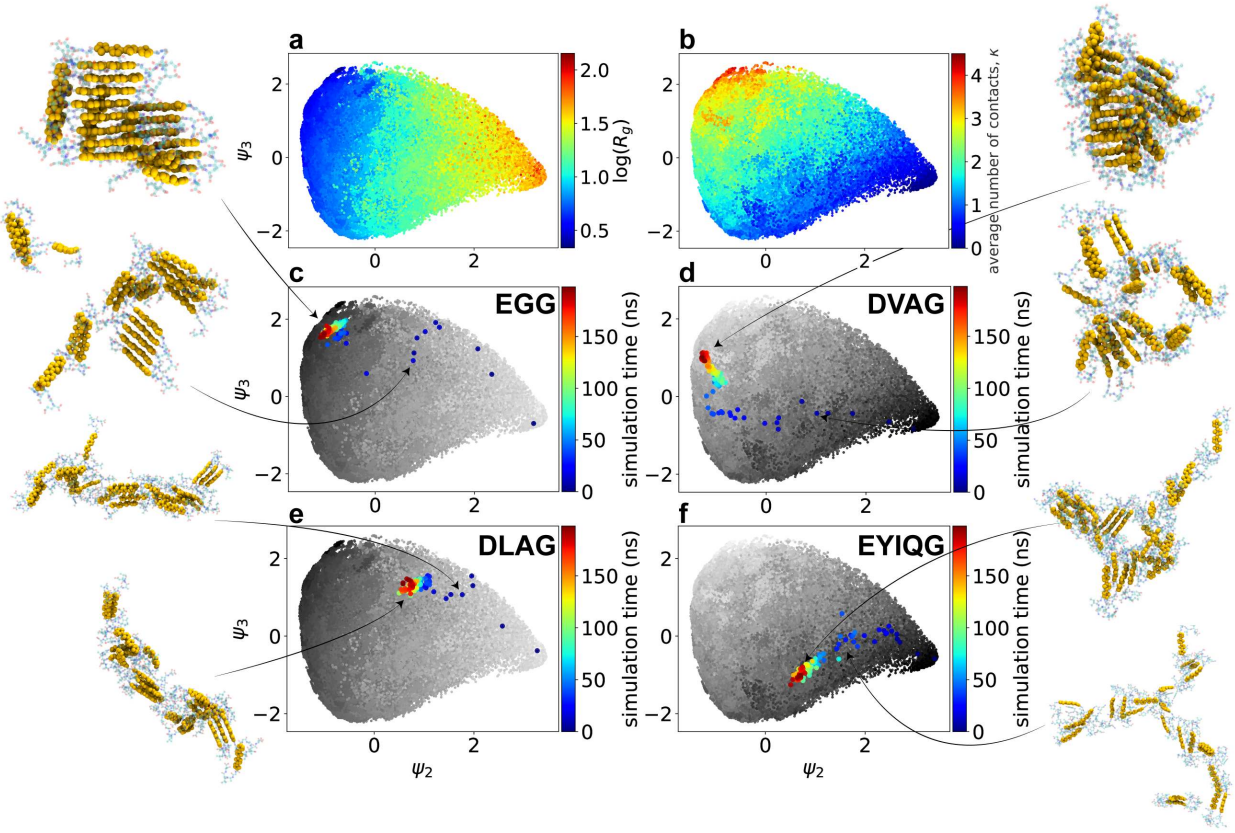


Figure 3.4: Two-dimensional diffusion map embedding of 118,100 simulation snapshots harvested from the molecular simulations of 1181 X_n-4T-X_n molecules collected over the course of the computational active learning screen. Each point represents one simulation snapshot projected into a 2D low-dimensional manifold spanned by the leading collective variables $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ learned by diffusion maps. Coloring the points by (a) the log radius of gyration $\log(R_g)$ of the self-assembled nanoaggregate and (b) average number of contacts per molecule κ within the aggregate exposes the strong correlation of the two learned collective variables with these two physical observables ($\rho(\boldsymbol{\psi}_2, R_g) = 0.93$, $\rho(\boldsymbol{\psi}_3, \kappa) = 0.69$). All assembly trajectories commence in a monomeric dispersion contained at the rightmost edge of the manifold. Progression from right-to-left corresponds to a reduction in R_g as the system self-assembles, and progression from bottom-to-top to the formation of more molecular contacts. The temporal progression of the self-assembly pathways over the manifold for molecules (c) EGG-4T-GGE, (d) DVAG-4T-GAVD, (e) DLAG-4T-GALD, and (f) EYIQG-4T-GQIYE, in which points are colored temporally. Grey points in panels c-f represent the embedding of 118,100 simulation snapshots shaded in panels c and e by $\log R_g$, and panels d and f by κ. Insets show representative molecular renderings throughout the trajectory.

that R_g necessarily decreases as the system self-assembles from dispersed monomers and that this behavior is not in conflict with the active learning goals of maximizing R_g and κ of

the self-assembled nanoaggregates in order to promote high-aspect ratio linear morphologies with good π-π stacking. One prototypical class of assembly pathways for high-performing molecules is exemplified by EGG-4T-GGE (rank 4/28), which traverses the upper edge of the manifold (Fig. 3.4c). This assembly route corresponds to the rapid formation of small oligomeric stacks, the formation of which is likely promoted by the small size of the peptide wing, that ultimately associate into an elongated aggregate with good in-register and global stacking. DVAG-4T-GAVD (rank 11/28) is another high-performing candidate that is prototypical of a different assembly route followed by high-performing molecules (Fig. 3.4d). This pathway commences with an initial rapid hydrophobic aggregation of the system corresponding to a rapid leftward lateral motion over the manifold. The absence of any early upward vertical motion is indicative of no initial substantive increase in κ due to the larger peptide wings seemingly preventing good π-core stacking. This initial collapse is, however, then followed by a more gradual structural ripening as the cores do achieve good stacking and we observe late upward motion over the manifold corresponding to an increase in κ.

Trajectories that terminate within the bulk of the manifold and far from the upper-left corner typically fail to form nanoaggregates containing globally connected pseudo-1D stacks. DLAG-4T-GALD (rank 24/28) is emblematic of a poor-performing molecule that initially builds a reasonable number of intermolecular contacts, but then fails to further condense into an in-register stacked nanoaggregate (Fig. 3.4e). Differing only in a V to L mutation relative to the high-performing DVAG-4T-GAVD, the presence of the bulkier hydrophobic Leu residue appears to preclude structural ripening into the desired elongated stack. Finally, molecules rich in large hydrophobic side chains such as EYIQG-4T-GQIYE (rank 18/28) tend to exhibit moderate leftward motion over the manifold corresponding to hydrophobic collapse but accompanied with unfavorable downward motion indicative of the formation of very few intermolecular π-core contacts (Fig. 3.4f). This behavior can be attributed to the bulky aromatic hydrophobes that stack against the π-cores and prevent the formation of core-core

contacts.

Whereas Fig. 3.4 provided anecdotal insights into the self-assembly trajectories traced out by particular representative X_n-4T-X_n molecules, in Fig. 3.5 we present the entire distribution of trajectory end points for all molecules considered in our active learning screen. In Fig. 3.5a-b we illustrate the end points of the 1181 molecules sampled in our computational screen colored by the R_g and κ values of the terminal nanoaggregates and in Fig. 3.5c the 28-molecule subset of these candidates that were experimentally tested colored by the measured spectral blue shift λ. Focusing on the 28 experimental molecules, we observe observe a clustering of 17 molecules in the upper left region of the manifold that we bound by a purple box. As anticipated by the understanding exposed by the diffusion map, these molecules tend to be high-performers comprising nine of the top 11 experimentally-tested molecules with spectral blue shifts $\lambda \geq$35 nm. Further, the molecules within the box possess a mean spectral shift of λ=31 nm compared to the mean value for those outside the box with λ=21 nm (one-sided Mann-Whitney U test, p-value = 0.07). The correlation between (R_g,κ) and λ within the diffusion map embedding provides a strong *post hoc* substantiation for the use of the former measures as a computational proxy for the latter within the active learning screen, and demonstrates the power and value of the high-throughput computational screen in focusing and guiding the low-throughput experimentation.

3.5 Conclusions

In this work, we have reported an integrated computational/experimental iterative design strategy to discover synthetic π-conjugated oligopeptides within the X_n-4T-X_n family with the capacity to self-assemble into highly-ordered linear aggregates with in-register stacking of the π-cores. These supramolecular assemblies are desirable as biocompatible nanoaggregates possessing emergent optoelectronic properties and potential applications as peptide-based field-effect transistors, photoconductors, or solar cells. The X_n-4T-X_n design space consist-

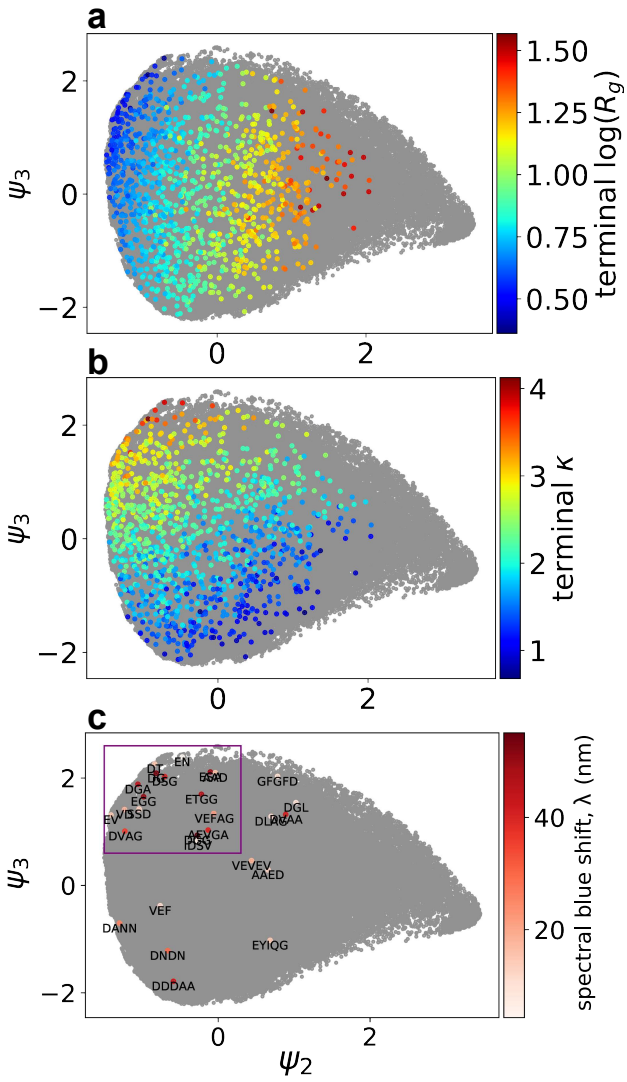


Figure 3.5: Terminal locations of the X_n-4T-X_n self-assembly trajectories over the 2D diffusion map manifold. End points of the 1181 molecules considered in our computational screen colored by the (a) log of the radius of gyration $\log(R_g)$ and (b) average number of contacts per molecule κ computed over the terminal 50 ns of the trajectory. The 118,100 simulation snapshots used to construct the diffusion map embedding are shown in grey. (c) Terminal locations of the 28 experimentally tested molecules colored by their measured spectral blue shift λ and annotated with the sequence of the peptide wing. The purple box bounds a cluster of high-performing experimental candidates residing in the upper-left region of the manifold possessing high values of λ.

ing of symmetric oligopeptide wings containing between one and five amino acids comprises 694,982 candidate molecules, making its exhaustive exploration impracticable by either simulation or experiment. By fusing computational and experimental data streams within an

integrated computational-experimental active learning framework, we perform a data-driven efficient traversal the space of X_n-4T-X_n peptides that minimizes computational and experimental burden required to discover and validate new high-preforming candidates. Our platform employs a combination of all-atom molecular dynamics simulations, deep representational learning, single- and multi-fidelity Gaussian process regression, and single- and multi-objective Bayesian optimization. A computational active learning loop serves as a high-throughput and cheaply available experimental proxy used to refine a surrogate model that predicts the experimental performance of untested candidates. Using this platform, we discovered a diversity of high-performing new molecules experimentally validated to form pseudo-1D linear nanoaggregates after sampling only 1181 molecules ($\sim$0.17% of the design space) by computation and 28 ($\sim$0.004%) by experiment. Subsequent interrogation of our experimental screening data exposed molecular design rules linking sequence to the emergent structure and function of the self-assembled nanoaggregates. Analysis of the computational screening results revealed two prototypical assembly mechanisms and pathways shared by the high-performing molecules: (i) hierarchical assembly of small in-register supramolecular oligomers that undergo further assembly into a single linear aggregate with ordered π-stacking and (ii) rapid hydrophobic collapse followed by slow structural ripening and the emergence of in-register ordering of the π-cores.

This work exposes new understanding of how variation in oligopeptide sequence in π-conjugated peptides impacts assembly behavior using an integrated experimental-computational active learning platform. Our findings corroborate prior physico-chemical understanding and chemical intuition of π-conjugated peptide assembly, but also reveals new design rules and understanding of molecular assembly mechanisms. Our hybrid computational/experimental active learning platform demonstrates the power of tightly integrated collaboration between theory and experiment, and this paradigm is transferable to other generic molecular design and discovery applications.

Data Availability

Data providing a full accounting of all molecules simulated and experimentally tested throughout the active learning process with associated measurements for the average number of contacts κ, radius of gyration R_g and spectral blue shift λ, and terminal GPR and mfGPR predicted κ, R_g and λ; neural network weights and training codes; GPR training codes; RAE embeddings; active learning workflow [251].

Supporting Information

Electronic supplementary information (ESI) available. See `https://doi.org/10.1039/d1dd00047k`

CHAPTER 4

DATA-DRIVEN DISCOVERY OF CARDIOLIPIN-SELECTIVE SMALL MOLECULES BY COMPUTATIONAL ACTIVE LEARNING

4.1 Abstract

Subtle variations in the lipid composition of mitochondrial membranes can have a profound impact on mitochondrial function. The inner mitochondrial membrane contains the phospholipid cardiolipin, which has been demonstrated to act as a biomarker for a number of diverse pathologies. Small molecule dyes capable to selectively partitioning into cardiolipin membranes enable visualization and quantification of the cardiolipin content. Here we present a data-driven approach that combines a deep learning-enabled active learning workflow with coarse-grained molecular dynamics simulations and alchemical free energy calculations to discover small organic compounds able to selectively permeate cardiolipin-containing membranes. By employing transferable coarse-grained models we efficiently navigate the all-atom design space corresponding to small organic molecules with molecular weight less than ≈500 Da. After direct simulation of only 0.42% of our coarse-grained search space we identify molecules with considerably increased levels of cardiolipin selectivity compared to a widely used cardiolipin probe 10-*N*-nonyl acridine orange. Our accumulated simulation data enables us to derive interpretable design rules linking coarse-grained structure to cardiolipin selectivity. The findings are corroborated by fluorescence anisotropy measurements of two compounds conforming to our defined design rules. Our findings highlight the potential of coarse-grained representations and multiscale modelling for materials discovery and design.

This chapter is based on the work reported in: Mohr, B., **Shmilovich, K.**, Kleinwächter, I. S., Schneider, D., Ferguson, A. L., & Bereau, T. (2022). "Data-driven discovery of cardiolipin-selective small molecules by computational active learning." *Chemical Science*,

13(16), 4498-4511.

4.2 Introduction

Mitochondria are double-membrane-bound organelles found in most eukaryotic cells (Fig. 4.1a). Their principal function is to generate the cell's supply of adenosine triphosphate (ATP), the main source of chemical energy used to drive biochemical reactions. [75] Through their key role in bioenergetics, mitochondria play an essential part in controlling cell proliferation, and they are involved in cell signaling and the activation of apoptosis. [291, 292, 293] The inner mitochondrial (IMM) membrane partitions the intermembrane space from the matrix and contains the membrane-bound ATP synthase proteins responsible for ATP generation. The composition of this membrane includes approximately 20% cardiolipin (CL) (Fig. 4.1b), a phospholipid comprising four acyl chains that is exclusively found in energy-generating membranes.[75, 76] Abnormalities in CL composition of inner mitochondrial membranes are linked to pathologies including Barth syndrome, Tangier disease, heart failure, and neurodegeneration.[76, 77]

The CL content of a membrane can be experimentally assayed using molecular stains to visualize and quantify the presence of CL. A primary challenge in the molecular design of cardiolipin-selective small molecules [294] is the high degree of structural similarity between CL and other common phospholipids. The most chemically similar phospholipid is phosphatidylglycerol (PG) (Fig. 4.1c), a precursor of which CL is essentially a dimer.[78, 79, 80] Apart from its role in the CL synthesis pathway, PG is prevalent in bacterial membranes and a minor constituent of eukaryotic membranes. [295] There are two principal differences in the CL and PG headgroups: CL possesses two chemically distinct phosphatidyl moieties whereas PG has only one, and PG possesses two hydroxyl groups while CL retains only one. The dianionic CL structure thereby stands out as the main difference to the monoanionic PG. A number of selective probes have been described for the purpose of detecting

CL,[86, 87, 296, 297, 298, 299] including the fluorescent dye 10-N-nonyl acridine orange (NAO) (Fig. 4.1d),[86, 87] but the degree to which these probes are optimally selective for CL relative to PG is not clear.[87, 300, 301, 302, 303, 304, 305]

Engineering selectivity into a molecular probe requires manipulation of the structure and physicochemical properties to stabilize favorable interactions with CL relative to other phospholipids, in particular, its chemically similar precursor PG. Due to the structural similarity of the two lipids, the design process must exploit the subtle differences in binding affinity mediated by the existence of two phosphate groups due to the four-acyl chain structure of CL relative to its two-acyl chain PG competitor. This motivates a design strategy targeting the lipid headgroups as the site of these distinguishing characteristics. The challenges posed by this subtle structural distinction are further compounded by the absence of a clear binding site—as might be anticipated in, for example, protein-ligand interactions—and the anticipated importance of inter-molecular, multi-body interactions between phospholipid chains within the membrane.

The objective of this work is to engage the design challenge posed by the subtle chemical differences between phospholipids and the lack of a defined binding site in lipid structures. To this end, we use a combination of a high-throughput virtual screening approach and data-driven active learning. The virtual screening is based on coarse-grained molecular dynamics simulations and alchemical free energy calculations, and the active learning employs deep representational learning using neural networks, Gaussian process regression surrogate models, and Bayesian optimization. As our objective function, we choose to maximize the difference $\Delta\Delta G$ in partitioning free energies between a CL and a PG membrane environment to maximize the thermodynamic preference of the molecule for CL relative to PG. By performing our screening using coarse-grained modeling we can substantially reduce the size of our search space without sacrificing coverage of chemical space. Because transferable coarse-grained models rely on a finite set of interaction types, many molecules map to

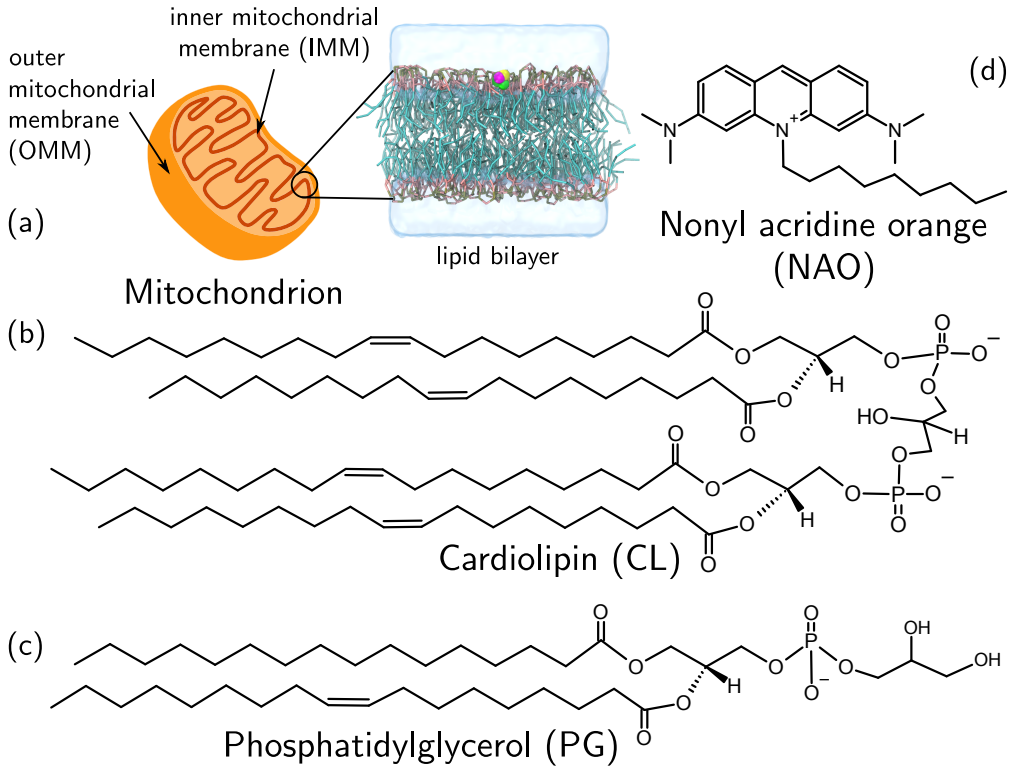


Figure 4.1: (a) Schematic representation of a mitochondrion with lipid bilayer system representing the inner mitochondrial membrane (IMM). (b) Chemical structures of (b) Cardiolipin (CL) and (c) Phosphatidylglycerol (PG) phospholipids. The CL molecule is essentially a dimer of PG. CL contains two phosphatidyl moieties and one hydroxyl group in its headgroup, in comparison one phosphatidyl group and two hydroxyl groups in the PG headgroup. (d) The cardiolipin probe 10-*N*-nonyl acridine orange (NAO). All structures were drawn with ChemSketch.[306]

the same coarse-grained representation.[85] The accuracy and transferability of our coarse-grained model enable us to tractably and efficiently engage the space of $\sim 10^{60}$ small organic molecules with molecular weight up to 500 Da.[84] The multiscale-based many-to-one mapping reduces to only 124,327 unique coarse-grained topologies that maintain the pertinent physicochemical properties. Within this reduced chemical space, we conduct seven rounds of iterative computational screening and surrogate model building linking molecular structure to thermodynamic properties. We discover 242 compounds with up to 184% superior predicted thermodynamic selectivity for CL membranes relative to NAO.

The discovery of high performing compounds after sampling only 0.42% of the CG molec-

ular design space exploits the power of the active learning protocol to efficiently guide sampling towards the most promising regions of chemical space to explore. Our molecular discovery platform also identifies human-interpretable design principles that furnish both novel molecular designs and promote new understanding of the subtle physicochemical molecular properties that lead to high CL selectivity. Specifically, post hoc analysis of the simulation data accumulated over the course of our active learning process enables us to identify chemical motifs that determine CL selectivity by building a sparsity-enforced structure–property regression model. We validate our computational model by performing experimental fluorescence anisotropy measurements of CL selectivity for two molecules selected according to our learned design rules and observe good concordance with our computational predictions. Our computational discovery platform is generically transferable to other molecular engineering applications by modular substitution of the selectivity assay for other addressable properties of interest.

4.3 Computational Methods

4.3.1 Overview

We develop and deploy a computational screening active learning cycle for the data-driven discovery of small molecule probes highly selective to CL membranes (Fig. 4.2). In a nutshell, we (i) conduct coarse-grained molecular dynamics (CGMD) simulations and alchemical free energy calculations to measure the thermodynamic preference $\Delta\Delta G$ of a particular candidate molecule for a CL membrane relative to a PG membrane, (ii) build supervised regression models over a learned low-dimensional latent space to predict the performance of new compounds that have not yet been simulated, and (iii) apply Bayesian optimization to the trained regression models to identify the next most promising compounds to submit for computational screening. After running the active learning cycle for several rounds, we

analyze the learned models in order to extract the physicochemical design rules underpinning the observed performance of the simulated candidates. We provide below the methodological details of each component of this computational workflow. A more comprehensive description of the theoretical bases and numerical implementations of these approaches is provided in the ESI†.

4.3.2 Coarse-grained molecular design space

The molecular design space for our computational screening is formed from organic molecules with molecular weights less than or equal to 500 Da. We impose this upper weight threshold in order to promote high mobility and diffusion into and through membranes.[294] No atom types are excluded from our design space. The number of molecules satisfying these criteria is $\sim 10^{60}$,[84] motivating the use of approximations and simplifications to efficiently screen this subset of chemical space. To this end, we employ the coarse-grained (CG) Martini 2 molecular model, that both greatly reduces the cost of our molecular simulations and significantly reduces the size of chemical compound space by grouping molecules into a smaller number of CG representations.[124, 307, 308] The Martini model was parameterized against thermodynamic data to generate 14 neutral and four charged CG bead types (*14+4* bead-type model) representing most physicochemical interactions relevant in biomolecular settings. This building-block approach allows the rapid generation of new representations without requiring individual reparameterization and provides a good balance between chemical accuracy and computational efficiency. The Martini CG force field has been widely used in the study of membrane organization and dynamics,[309, 310, 311, 312] drug-membrane permeability, [313, 314, 315, 316, 317] and membrane-protein interactions.[194, 318, 319, 320, 321] More recently it has been shown that even coarser models, going as low as five bead types, represent the underlying physical properties comparably well while facilitating more thorough coverage of chemical compound space by reducing the combinatorial complexity.[85]

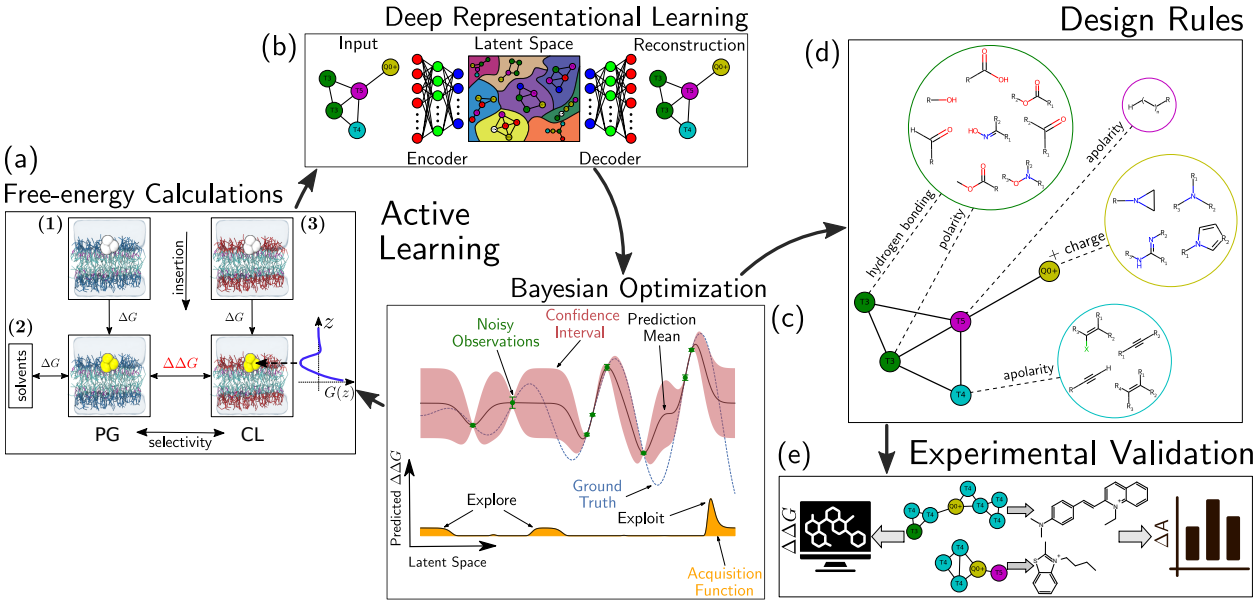


Figure 4.2: Active learning cycle for the automated discovery of small ($\leq$500 Da) molecules with high CL selectivity. (a) Coarse-grained molecular dynamics (CGMD) simulations are performed to calculate the partitioning free energy of a molecule into PG and CL membranes used to evaluate the CL selectivity $\Delta\Delta G$. (b) Deep representational learning using a regularized autoencoder (RAE) is used to construct a fixed-size latent space embedding of the discrete molecular design space encompassing all coarse-grained molecular candidates. Gaussian process regression (GPR) surrogate models are fit using all accumulated simulation data to predict the CL selectivity $\Delta\Delta G$ of all untested compounds within the design space. (c) These surrogate model predictions are then interfaced with a Bayesian optimization platform to select the next most promising compounds for computational simulation. This process iteratively continues until multiple consecutive active learning rounds fail to identify new top-performing compounds. (d) Using our accumulated simulation data we extract design rules linking the inclusion/omission of chemical functional groups to the degree of CL selectivity $\Delta\Delta G$ by building of interpretable linear models using graph representational learning. (e) The design rules are used to select candidate compounds that are subjected to fluorescence anisotropy measurements as well as CG free energy calculations to validate our findings. Structures in this panel are drawn with ChemSketch,[306] the icons were obtained from Flaticon.com.

We created a *5+1* bead-type CG model through extending the five bead-type *5+0* reduced Martini model[85] by one charged bead-type to fully represent the candidate small molecule probes. The additional charged bead-type represents single positive or negative charges (Q0±). We offset this increase in complexity by removing the two non-polar bead types representing only hydrogen bond donor or acceptor properties (T3d, T3a), as the reduced Martini model already contains a nonpolar hydrogen bond donor- and acceptor bead-type (T3) that represents both interaction types simultaneously. The complete *5+1* model comprises the set of five neutral bead types ordered by descending polarity and one charged bead type {T1, T2, T3, T4, T5, Q0±} as illustrated in Fig. S1 in the ESI†. This reduced model allowed us to explore the chemical compound space more efficiently compared to the *14+4* interaction type scheme applied in Martini.[124]The combinatorial candidate space of compounds less than 500 Da was defined by constructing all plausible molecular graphs containing five or fewer *5+1* beads, resulting in 124,327 candidate compounds.

4.3.3 Coarse-grained molecular dynamics (CGMD) simulations

We constructed CL and PG membranes comprising 98 and 118 lipid molecules, respectively, using the CHARMM-GUI Martini maker.[322] The membranes were solvated in 3287 water particles for CL and 1754 water particles for PG. A sufficient number of sodium ions (Na^+) to maintain charge neutrality were added: two Na^+ ions per CL headgroup since the double negative charged CL model was chosen,[323, 324, 325] and one Na^+ ion per PG headgroup. The CL and PG phospholipid membranes were represented using the Martini 2 force field, [124] water using the refined parameters for polarizable water,[308] and ions using the polarizable ion model.[307] Coarse grained molecular dynamics (CGMD) simulations were conducted using GROMACS 2018.6[257] implementing the standard Martini run parameters introduced for GPU acceleration.[326] Full details of the force field and run parameters are reported in the ESI†.

4.3.4 Alchemical free energy calculations

We evaluate CG candidate molecules based on their thermodynamic affinity for a CL membrane environment relative to a PG membrane. It is our goal to maximize the affinity of the probe for CL and simultaneously minimize its affinity for PG in order to maximize CL selectivity. It is computationally intractable to consider all possible competing phospholipid environments in our molecular screen, so we adopt PG as the negative design target as the most chemically similar phospholipid to CL and therefore the most challenging target for negative design. We quantify selectivity by computing the relative partitioning free energy $\Delta\Delta G$ for each candidate molecule between the transfer free energy from the water phase to the interface region of a PG membrane ($\Delta G^{\mathrm{PG}}_{\mathrm{W}\to\mathrm{I}}$) and the same transfer for a CL membrane ($\Delta G^{\mathrm{CL}}_{\mathrm{W}\to\mathrm{I}}$),

$$\Delta\Delta G = \Delta G^{\mathrm{CL}}_{\mathrm{W}\to\mathrm{I}} - \Delta G^{\mathrm{PG}}_{\mathrm{W}\to\mathrm{I}}. \tag{4.1}$$

We focus on the water–membrane interface since it is the water-facing lipid headgroups that are the site of the primary chemical difference between CL and PG, and therefore the region of interest for engineering target selectivity.

All free energies are calculated using CGMD alchemical free energy calculations,[81] using the MBAR method[82] with tools provided by the pymbar [83] package. Full details of the calculation procedure are reported in the ESI†. Calculation of $\Delta\Delta G$ for a single candidate molecule requires approximately ∼24 GPU-h for an uncharged molecule and ∼48 GPU-h for a charged molecule on a single NVIDIA Tesla V100 GPU card. Exhaustive simulation of all 124,327 candidate molecules would therefore require ∼4.8M GPU-h conservatively assuming an average simulation time per molecule of ∼36 GPU-h. To reduce the overall computational cost, we adopted a three-step hierarchy for the calculations that enables early exit for unviable candidate molecules. Only candidates that meet our minimum requirements of easily partitioning into and aligning with the interface region of the PG membrane are subjected to the most computationally expensive calculations in the CL membrane.

1. Assessment of interfacial preference in PG. In the first step, we perform alchemical transformations to compute the free-energy difference ΔG^{PG} of transferring the candidate molecule from a vacuum reference state to the interface of the PG membrane, (Fig. 4.2a and ESI† Fig. S2 (1)). The cumulative probability of the positions of the compound along the membrane normal, z, is evaluated over the trajectory to determine whether a candidate molecule inserted into the membrane will preferentially position itself at the interface (ESI†, Fig. S7 (1)). If it spontaneously moves astray toward the hydrophobic membrane midplane or water phase, we declare the candidate "non-interfacial." For computational efficiency, we assume that this determination can be made by considering only PG membranes so we do not repeat this calculation in CL. We eliminate non-interfacial candidates from further consideration since they do not preferentially reside near the phospholipid headgroups where CL and PG are most chemically differentiated.

2. Calculation of PG transfer free energies. For candidate structures deemed interfacial, the second step consists of alchemical free-energy calculations in bulk liquids: water (ΔG^{W}) and octane (ΔG^{O}). The two environments are used to evaluate the transfer free energies from vacuum into the water phase and the membrane midplane, respectively (Fig. 4.2a and ESI† Fig. S2 and S7 (2)). In line with Menichetti *et al.*, [315] Fig. S6 in the ESI† shows that octane acts as an excellent proxy for the membrane midplane environment, which additionally offers significant computational savings. The free-energy differences ΔG^{PG}, ΔG^{W}, and ΔG^{O} allow us to calculate the transfer free energies $\Delta G^{\mathrm{PG}}_{\mathrm{W}\to\mathrm{I}} = \Delta G^{\mathrm{PG}} - \Delta G^{\mathrm{W}}$ from the water phase to the interface and $\Delta G^{\mathrm{PG}}_{\mathrm{O}\to\mathrm{I}} = \Delta G^{\mathrm{PG}} - \Delta G^{\mathrm{O}}$ from the midplane to the interface. Negative values for both $\Delta G^{\mathrm{PG}}_{\mathrm{W}\to\mathrm{I}}$ and $\Delta G^{\mathrm{PG}}_{\mathrm{O}\to\mathrm{I}}$ confirm a spontaneous insertion of the candidate at the PG interface.

3. Calculation of CL transfer free energies. If the previous bulk simulations confirm the interfacial nature of the compound, we perform a final set of alchemical transformations. We compute the transfer free energy of the compound from vacuum to the interface of the

CL membrane ΔG^{CL} (ESI†, Fig. S2 and S7 (3)). We then use this value to compute the CL transfer free energies $\Delta G^{\mathrm{CL}}_{\mathrm{W}\to\mathrm{I}} = \Delta G^{\mathrm{CL}} - \Delta G^{\mathrm{W}}$ from the water phase to the interface and $\Delta G^{\mathrm{CL}}_{\mathrm{O}\to\mathrm{I}} = \Delta G^{\mathrm{CL}} - \Delta G^{\mathrm{O}}$ from the midplane to the interface. We now have all the ingredients to quantify the compound's thermodynamic preference for the CL membrane relative to its PG counterpart, $\Delta\Delta G = \Delta G^{\mathrm{CL}}_{\mathrm{W}\to\mathrm{I}} - \Delta G^{\mathrm{PG}}_{\mathrm{W}\to\mathrm{I}}$. Larger negative values of the transfer free-energy difference $\Delta\Delta G$ are indicative of stronger thermodynamic selectivity for CL relative to PG.

4.3.5 Chemical space embedding

Each of the 124,327 CG molecular candidates in the design space is represented as a molecular graph composed of up to five beads selected from the possible six bead types and different charge states of the *5+1* model {T1, T2, T3, T4, T5, Q0±}.[85] The nodes of the graph represent the identity of each bead in the molecule and the edges capture the connectivity of the beads within the CG topology. Performing molecular design and optimization directly over the discrete molecular design space has been shown possible using kernel-based methods.[263, 264, 265, 266, 267, 268, 39] Here instead, we choose to project candidate molecules into a learned, smooth, low-dimensional, and continuous embedding. This approach renders our design space amenable to the construction of low-dimensional and robust surrogate models and the use of off-the-shelf Bayesian optimization algorithms.[131, 244] We learn an appropriate latent space embedding in a data-driven fashion by training a regularized autoencoder (RAE),[254] a deterministic adaptation of the variational autoencoder (VAE) architecture,[269] over the corpus of 124,327 CG molecular graphs (Fig. 4.2b). The encoder-decoder architecture is composed of a message passing neural network encoder and a permutation-invariant graph decoder.[270, 271, 272, 273, 274] The network accepts graph-structured molecular representations where the nodes are featurized with one-hot representations of the bead type and the charge state. The edge between two nodes relies on

the corresponding Lenard-Jones 6–12 interaction parameter. It is the objective of the RAE encoder to learn a smooth and continuous latent representation containing the salient information about the input graphs from which the decoder can accurately reconstruct (i.e., auto-encode) the same graph. We achieve good reconstruction performance employing a 16-dimensional bottleneck layer between the encoder and decoder that defines our latent space dimensionality and representation, and within which we define measures of proximity between candidates and construct our surrogate models and perform global optimization via active learning. Although RAEs are generative models that when trained on open data sets are in principle capable of producing never-before-seen reconstructions, our model operates within a fully enumerated and fixed molecular design space such that we do not require this generative functionality. Rather, upon completing training, the decoder is discarded and the encoder alone is purposed for dimensionality reduction mapping our 124,327 candidate design space into a continuous embedding for downstream application within our active learning workflow. RAE models were constructed and trained using PyTorch.[276] The RAE was trained only once over all 124,327 candidate molecules prior to commencing active learning. Training required ∼36 GPU-h on a single NVIDIA Tesla V100 GPU card. Full details of the network architecture and training are provided in the ESI†.

4.3.6 Active learning

The primary goal of our active learning cycle is to efficiently discover CG compounds within the 124,327-member candidate space with high selectivity for CL membranes (i.e., large negative $\Delta\Delta G$). The high computational cost to measure the selectivity of each candidate precludes an exhaustive traversal of these candidates and we must employ a more computationally efficient search strategy. We achieve this by iterating between targeted CGMD simulations of promising candidates designed using a surrogate model relating molecular structure to CL selectivity $\Delta\Delta G$ and the training and interrogation of these surrogate mod-

els constructed over the 16-dimensional RAE latent space. To ensure broad initial coverage of the molecular design space, we seed the first round of active learning by conducting CGMD simulations of 100 compounds residing closest to the centroids of a 100-cluster k-means partitioning of the latent space.[327] We then trained a Gaussian process regression (GPR)[58] surrogate model with a Gaussian kernel to learn a mapping from the 16-dimensional latent space coordinates to the calculated $\Delta\Delta G$. The trained model was then applied to predict the $\Delta\Delta G$ values for all remaining $124,327 - 100 = 124,227$ compounds in the design space along with the predicted model uncertainties. The predictions of the surrogate GPR model were then interfaced with a Bayesian Optimization (BO)[55] framework using the Expected Improvement (EI)[55, 67] acquisition function and Kringing believer batched sampling [279] to identify a 60-molecule batch of candidates with the most promising (i.e., lowest) values of $\Delta\Delta G$ (Fig. 4.2c). Importantly, the EI acquisition function accounts for both the value of the GPR predictions and its estimated uncertainties, enabling the BO-directed active learning search to direct inquiry towards both regions of design space with favorable predictions (exploitation) and under-explored regions with high uncertainties (exploration). This batch of 60 compounds is then subjected to CGMD free-energy calculations to evaluate their associated $\Delta\Delta G$ values and the cycle repeats. Each round of active learning involves updating the training dataset based on all accumulated simulation data, refitting the GPR to predict out-of-training sample $\Delta\Delta G$ selectivity, and BO selection of the most promising compounds to simulate next (Fig. 4.2a-c). Convergence of the active learning cycle is monitored by tracking the distribution of $\Delta\Delta G$ values measured in each round, and the process is terminated when multiple consecutive rounds fail to identify new best performing candidates. As detailed below, we compute $\Delta\Delta G$ values for a total of $N = 439$ interfacial molecular candidates.

4.3.7 Inference of design rules

After completing the active learning screen, we analyzed the collected library of $\Delta\Delta G$ values to extract human-interpretable design rules linking the presence or absence of particular CG structural motifs to the calculated values of the transfer free energy (Fig. 4.2d). We constructed these interpretable models by decomposing each CG topology as a distribution of relative subgraph frequencies and then performed sparse LASSO regression to rank the most impactful subgraph motifs. We begin by enumerating $k = 1608$ topologically unique subgraphs with 1–5 CG beads that are contained within the $N = 439$ interfacial compounds. Each CG topology is then featurized as a length-k vector of subgraph frequencies, reweighted to account for the over counting of smaller subgraphs necessarily contained within larger subgraphs, and normalized to unit length.[328] Assembling these featurizations into a normalized frequency matrix $F \in \mathbb{R}^{(N=439)\times(k=1608)}$ we adopt a simple and interpretable linear model for predicting transfer free energy based on these relative subgraph frequencies,

$$\Delta\Delta G^{(n)}_{\text{predicted}} = \Delta\Delta G_{\text{mean}} + \sum_{k=1}^{1608} \theta_k F_{n,k}. \tag{4.2}$$

$\Delta\Delta G_{\text{mean}}$ is the arithmetic mean of all N=439 $\Delta\Delta G$ values and the regression coefficients $\boldsymbol{\theta} \in \mathbb{R}^{1608}$ assign weights to the different subgraph frequencies. The coefficients $\boldsymbol{\theta}$ are learned by minimizing the LASSO regression loss between the predicted and the calculated free energies, $\Delta\Delta G_{\text{predicted}}$ $\Delta\Delta G_{\text{ref}}$, respectively. The resulting expression yields

$$\begin{aligned} L(\Delta\Delta G_{\text{predicted}}\,, \Delta\Delta G_{\text{ref}}; \boldsymbol{\theta}, \alpha) = \frac{1}{2}\left(\Delta\Delta G_{\text{predicted}} - \Delta\Delta G_{\text{ref}}\right)^2 \\ +\alpha\|\boldsymbol{\theta}\|_1, \end{aligned} \tag{4.3}$$

where α is the L1 regularization weight promoting model sparsity. As $\alpha \to \infty$ the L1 regularization penalty in Eqn. 4.3 dominates corresponding to the null model where $\boldsymbol{\theta} = \mathbf{0}$,

but as $\alpha \to 0$ nonzero elements progressively accumulate within $\boldsymbol{\theta}$ corresponding to the participation of more subgraph motifs into the regression model. This L1 regularization serves to prevent overfitting by identifying and retaining only a small number of the most generalizable features represented in our training dataset. The optimal value for α is selected using cross-validation (ESI†, Fig. S19) and the learned nonzero weights in $\boldsymbol{\theta}$ can be interpreted as the most critical subgraph motifs for transfer free–energy prediction. Further, the linear nature of Eqn. 4.2 admits a simple interpretation to the sign of the learned weights: large negative weights $\theta_k < 0$ correspond to subgraph motifs predictive of good cardiolipin selectivity, while large positive weights $\theta_k > 0$ correspond to subgraph motifs predictive of poor cardiolipin selectivity. Analyzing a rank-ordering of the largest negative/positive weights provides a means for automatic selection and discovery of design rules identifying subgraph motifs most impactful for $\Delta\Delta G$ inference. Lastly, our trained model allows us to extrapolatively predict CL selectivity of unseen and arbitrarily large CG topologies via a decomposition into learned contributions of their constituent 1-5 bead subgraphs.

4.3.8 Functional group analysis

The CG model integrates out atomic-level information to only keep essential physicochemical properties. The ability to backmap from coarse-grained to all-atom resolution would offer chemical insight as to compounds of interest. As such, we aim to identify the diversity of all-atom structures given CG graphs of interest. This problem is more easily addressed by working from high- to low resolution: we consider a large set of molecules and coarse-grain them all. In this work we relied on the Generated DataBase (GDB).[329] We only considered molecules mapping to a single CG bead (1481 molecules of a size up to six heavy atoms).[314] On top of their original neutral form, we used the Calculator Plugin of Marvin[330] by *ChemAxon* to estimate their protonation state at pH 7. In addition, we account for aromatic groups by further considering around 22,000 small five- or six-member

cyclic compounds. The CG representations of cyclic molecules were used to evaluate how heteroatoms or substituents present in cyclic hydrocarbons affected the choice of bead types in the coarse-graining process. To automatically identify functional groups, bead types were linked to chemical information using an algorithm described by Ertl[331] and implemented in RDKit.[332] We extended the Ertl algorithm to recognize and name the most common functional groups found in bioactive molecules[333] as well as additional chemical structures repeatedly encountered in our GDB-derived dataset. More detail of this analysis is provided in the ESI†. Probabilities of functional groups mapping to a CG bead type were inferred through observation frequencies. The probabilities between CG bead type and functional group offer a practical link to translate CG design rules into chemical structures.

4.4 Experimental Methods

4.4.1 Fluorescence anisotropy

1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC), 1,2-dioleoyl-sn-glycero-3-phosphoglycerol (DOPG), and cardiolipin (CL) were purchased from Avanti Polar Lipids (Alabaster, ALA, US). 1,6-Diphenyl-1,3,5-hexatriene (DPH) was purchased from Fluka Analytical (Charlotte, North Carolina, US). The small molecule candidate compounds were purchased from Mcule (Budapest, HUN). Liposome solutions with a total lipid concentration of 300 μM, 0.6 μM of DPH, and 3 μM of the indicated substance, were prepared as follows: the lipids (dissolved in chloroform) were mixed in the described ratio with DPH (dissolved in chloroform) and, if needed, the small molecule candidate compound (dissolved in methanol). The solvents were removed under a gentle nitrogen stream and the lipids were desiccated overnight under vacuum. The next day, the lipids were dissolved in 10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)-buffer (pH 7.4, 150 mM NaCl) via vortexing 2x for 1 min, followed by 5 freeze-thaw circles. 250 μL of each sample was measured at 25°C.

Each lipid composition was measured at least three times using fresh lipid preparations. Anisotropy was measured using a Fluoro-Max-4 fluorescence spectrometer (Horiba (Bensheim, Germany)). Single point anisotropy was measured with 5 nm bandwidth, an excitation wavelength of 350 nm and an emission wavelength of 452 nm. Each sample was measured five times and the results were averaged. Samples in the absence and presence of the small molecule candidates were measured leading to the anisotropy difference,

$$\Delta A = A_S - A_\emptyset,$$

where A_S is the measured fluorescence anisotropy of a lipid membrane in presence of the small molecule, while $A_\emptyset$ is the anisotropy measured for an empty membrane. Experimental errors are given as standard error of the mean calculated from the three repetitions with fresh liposomes for the individual measurements, and calculated based on error propagation for the difference in anisotropy.

4.5 Results and Discussion

4.5.1 Active learning identifies highly cardiolipin-selective compounds

We perform active–learning directed CG alchemical free energy calculations within an RAE-learned chemical space embedding to discover compounds with high thermodynamic selectivity for CL membranes. We conduct seven rounds of active learning during which we consider 520 molecules (100 seeded and 60 per round) within the search space of 124,327 possible CG candidates. The majority (439 of 520) of these molecules were observed to be interfacial and for which we performed the complete alchemical free-energy calculations to determine $\Delta\Delta G$. The remaining 81 compounds failed to partition to the membrane interface, were deemed non-interfacial, and the free energy calculations terminated early as a resource-saving strategy.

The intent of the active learning search was to efficiently navigate chemical space to identify profitable regions of our learned latent space embedding densely populated with high-performing candidates by balancing exploration—compounds predicted with high uncertainty, and exploitation—compounds predicted with high performance. The performance of the active learning search is presented in Fig. 4.3a, which illustrates the distribution of $\Delta\Delta G$ values within each round, and Fig. 4.3b, which illustrates the corresponding location within the latent space of the selected candidates. The distribution of $\Delta\Delta G$ values tends towards more negative values over the course of the search, indicating that the active learning strategy is successfully discovering high-performing candidate molecules. Particular rounds tend to perform more exploration, such as Rounds 6 and 7, reflected by more variance in the sampled $\Delta\Delta G$ distribution, broader latent space sampling, and a higher proportion of selected non-interfacial compounds. Other rounds, such as Rounds 4 and 5, tend to be more exploitative, reflected by more localized latent space sampling and $\Delta\Delta G$ distributions with more weight in the negative tail. While these trends in sampling help provide some intuition for the progress of the active learning process, our use of the Expected Improvement acquisition function within our Bayesian optimization selection procedure means that each round naturally balances exploration and exploitation to select the next most effective compounds to simulate in order to maximize the chances of discovering high-performing candidates within the design space.

We assessed convergence of the active learning by monitoring the performance of the best observed candidates after each round. Throughout the first four rounds we observe an approximately linear decrease in the overall best observed $\Delta\Delta G$, with the following three rounds yielding no overall improvement but nonetheless identifying relatively high performing candidates (Fig. 4.4a). The stagnant improvement over the last three rounds motivates us to terminate sampling after Round 7. Over the course of the seven rounds of active learning we observed a 22.2% improvement in the best calculated $\Delta\Delta G$ after simulating a total of only

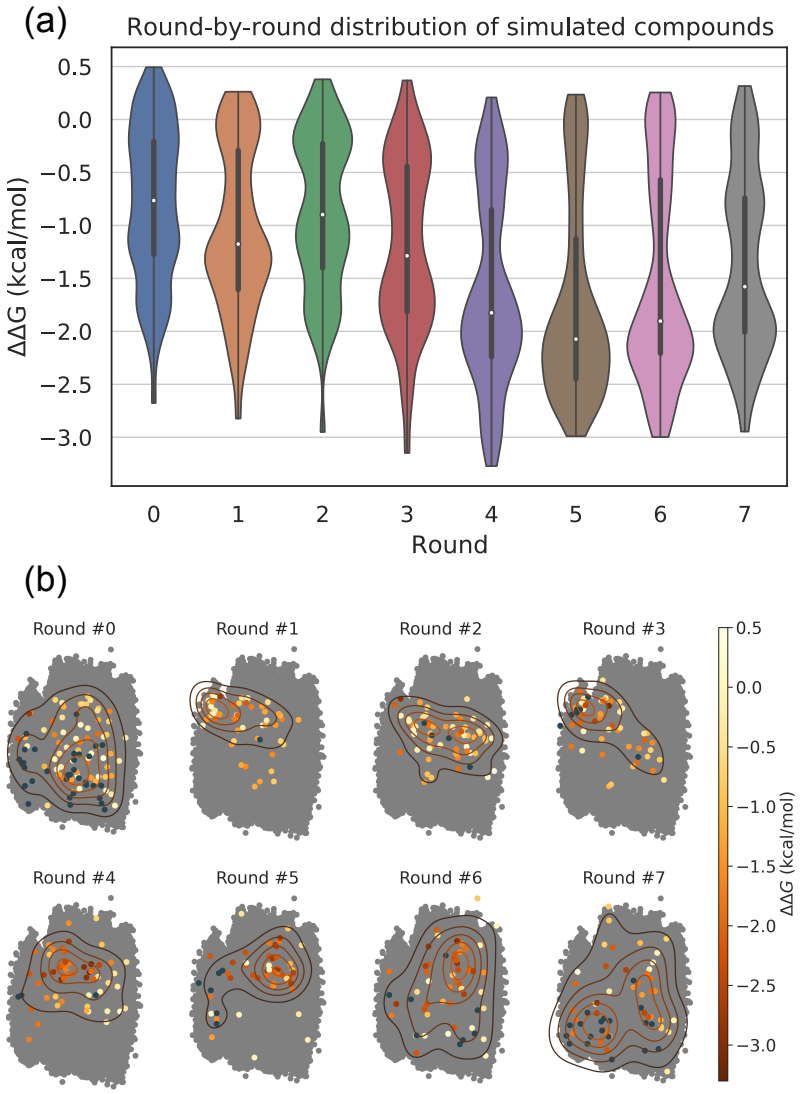


Figure 4.3: **(a)** Distribution of calculated thermodynamic selectivities for CL relative to PG membranes $\Delta\Delta G$ for CG compounds simulated in each active learning round. Each active learning cycle employs Bayesian optimization to select 60 candidate molecules for which we calculate $\Delta\Delta G$ using alchemical free energy calculations and which are used to iteratively retrain and update the model for subsequent rounds of selection. **(b)** Latent space embeddings of simulated compounds visualized in a 2D principal component analysis projection of the 16-dimensional RAE latent space. Points are colored by their calculated $\Delta\Delta G$. Dark blue points represent compounds that fail to spontaneously insert into the membrane bilayer and are deemed not interfacial and hence are terminated early within the workflow without finalizing the $\Delta\Delta G$ calculation. Model predictions become more accurate round by round as data is accumulated enabling the acquisition function to progressively better localize sampling in regions of latent space rich in high-performing compounds.

520 molecules, corresponding to a mere 0.42% of the 124,327 possible candidates. We present in Fig 4.4b the 12 molecules with the best (i.e., most negative) $\Delta\Delta G$ identified over the course of our search. Importantly, the best performing candidate possesses a $\Delta\Delta G = -3.27$ kcal/mol corresponding to a 184% improvement over the NAO fluorescent dye molecule which possess a thermodynamic selectivity of only $\Delta\Delta G = -1.15$ kcal/mol calculated under our CG model. A full accounting of the measured $\Delta\Delta G$ values for all 520 candidates is provided in the Data Availability.[334]

We can approximately quantify the savings afforded by our active learning process in discovering our best performing molecule by comparing against a baseline of naïve random sampling. A simple statistical analysis reveals that performing random selection of candidates within our 124,327 molecule candidate space and still considering 60 molecules per round, we would expect to happen upon our best-performing candidate with a 25% chance of success after ∼518 rounds and with a 50% chance of success after ∼1036 rounds. Although chemical intuition and prior experience could be used to guide the search, the bias and preconceptions that this introduces risks missing non-intuitive but high-performing candidates, and this baseline random search comparison nonetheless effectively highlights the value of our data-driven approach for guided molecular discovery.

4.5.2 Data-driven discovery of chemical design rules

Having completed seven rounds of active learning and calculating $\Delta\Delta G$ values for N=439 compounds, we proceed to use graph representational learning to discover design rules correlating the presence/absence of chemical motifs to CL selectivity. In the interest of simplicity and interpretability, we adopt a simple linear model given in Eqn. 4.2 for predicting the calculated transfer free energy of a molecule via a featurization based on its decomposition into topologically unique structures with 1–5 CG beads. This model is trained using the LASSO regression algorithm to promote sparsity where select elements from the learned

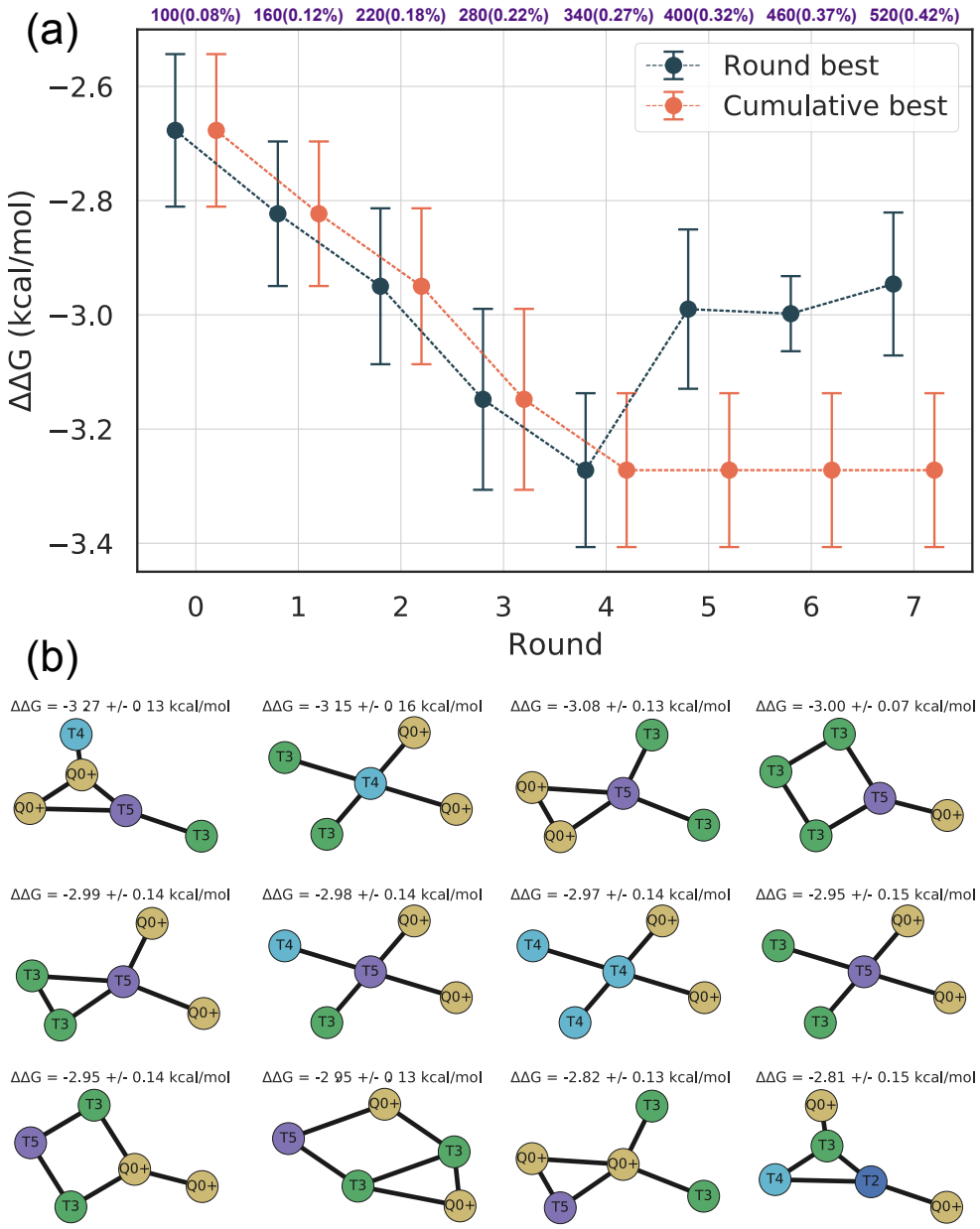


Figure 4.4: **(a)** Best sampled $\Delta\Delta G$ for each active learning round alongside the cumulatively best sampled molecule in each round. The top axis denotes the number of compounds simulated after each round and in parentheses as a percent of the 124,327 molecule design space. Errors represent uncertainly in the measured $\Delta\Delta G$ for the associated compound. The best candidate was selected after round four, with the final three rounds failing to identify any new better performing candidates. **(b)** The 12 best coarse-grained molecules identified over all seven active learning rounds sorted by increasing $\Delta\Delta G$. Large negative values of $\Delta\Delta G$ correspond to high thermodynamic selectivity for CL over PG membranes.

model weights are set precisely to zero $\theta_k = 0$, and the nonzero weights $\|\theta_k\| > 0$ correspond to a small number of the most generalizable features retained by the model. Based on the linear structure of Eqn. 4.2 the sign of learned nonzero weights θ_k suggests representation of the corresponding subgraph motifs as contributing to either more favorable $\Delta\Delta G$ values leading to better CL selectivity if $\theta_k < 0$, or unfavorable $\Delta\Delta G$ values if $\theta_k > 0$. Furthermore, as the subgraph representations $F_{n,k}$ in Eqn. 4.2 are normalized and therefore unitless, the coefficient weights θ_k carry the same units as $\Delta\Delta G$ of kcal/mol. The magnitude of these coefficient weights θ_k can therefore be interpreted as the extent in kcal/mol that the representation of specific subgraph motifs improve, if $\theta_k < 0$, or degrade, if $\theta_k > 0$, upon the average CL selectivity $\Delta\Delta G_{\text{mean}}$. Analyzing the prevailing structural features and characteristics contained in the largest magnitude model weights serves as a data-driven approach for unveiling critical determinants of CL selectivity.

In Fig. 4.5 we present a rank ordering of the largest magnitude nonzero model weights θ_k partitioned by the sign of the weight, $\text{sgn}(\theta_k)$. These largest magnitude weights can be interpreted as corresponding to CG motifs with the highest predictive capacity and therefore the most influential for determining CL selectivity. We note that while this analysis pertains to the CG space our active learning search is performed in, the following Sec. 4.5.3 provides further analysis linking CG beads and structures to atomistic functional groups. This current analysis reveals the importance of subgraph motifs containing positive net charges (Q0+), apolar (T5) bead types, and weakly polar T3 bead with both hydrogen bond donor- and acceptor properties in promoting favorable CL selectivity (i.e., negative values of $\Delta\Delta G$). Contrariwise, negative charges (Q0−) and highly polar beads (T1, T2) tend to impair CL selectivity. These observations can be intuitively rationalized, given the two phosphate groups of CL that are predominantly negatively charged in the physiological pH range,[324, 335] and the close proximity of the hydrophobic lipid tails to the interface region caused by the unique shape of the CL headgroup. The presence of apolar regions in the candidate structures

therefore facilitates easy insertion into the membrane. Electrostatic interactions have been described as primary interaction modes of small molecules with the CL headgroup and the important role of hydrophobic interactions has also been recognized.[86, 336, 87, 296, 337] The small CL headgroup containing a single hydroxy group also suggests that hydrogen bonding may represent an important interaction type for targeting CL, and this correlates well with the frequent appearance of the T3 beads in the identified chemical motifs. Taken together, this analysis allows us to extract the following design rules for highly CL-selective small molecules from the active learning results:

1. At least one, ideally two sites that will carry a positive charge at physiological pH (pH $\approx$ 7.3),

2. Hydrophobic areas in the molecule that induce alignment with or insertion into the lipid bilayer,

3. Functional groups able to form hydrogen bonds with the CL headgroup.

4.5.3 All-atom backmapping

We adopted the *5+1* CG model in order to efficiently screen chemical space at the cost of integrating atomic representations into coarse-grained beads. The top performing candidates identified within our active learning search are therefore represented as bead graphs as opposed to chemical structures. In order to approximately recover these lost degrees of freedom, we perform a backmapping analysis to identify functional groups with physical and chemical properties consistent with each CG bead. Due to the information removed by the coarse-graining procedure, there are multiple all-atom structures consistent with each CG representation, i.e., the mapping is many-to-one. The most prevalent bead to all-atom functional group mappings resulting from our analysis are shown in Fig. 4.6 and Fig. 4.7. The mapping probabilities for cycle-containing groups are presented in Fig. S25 in the ESI†.

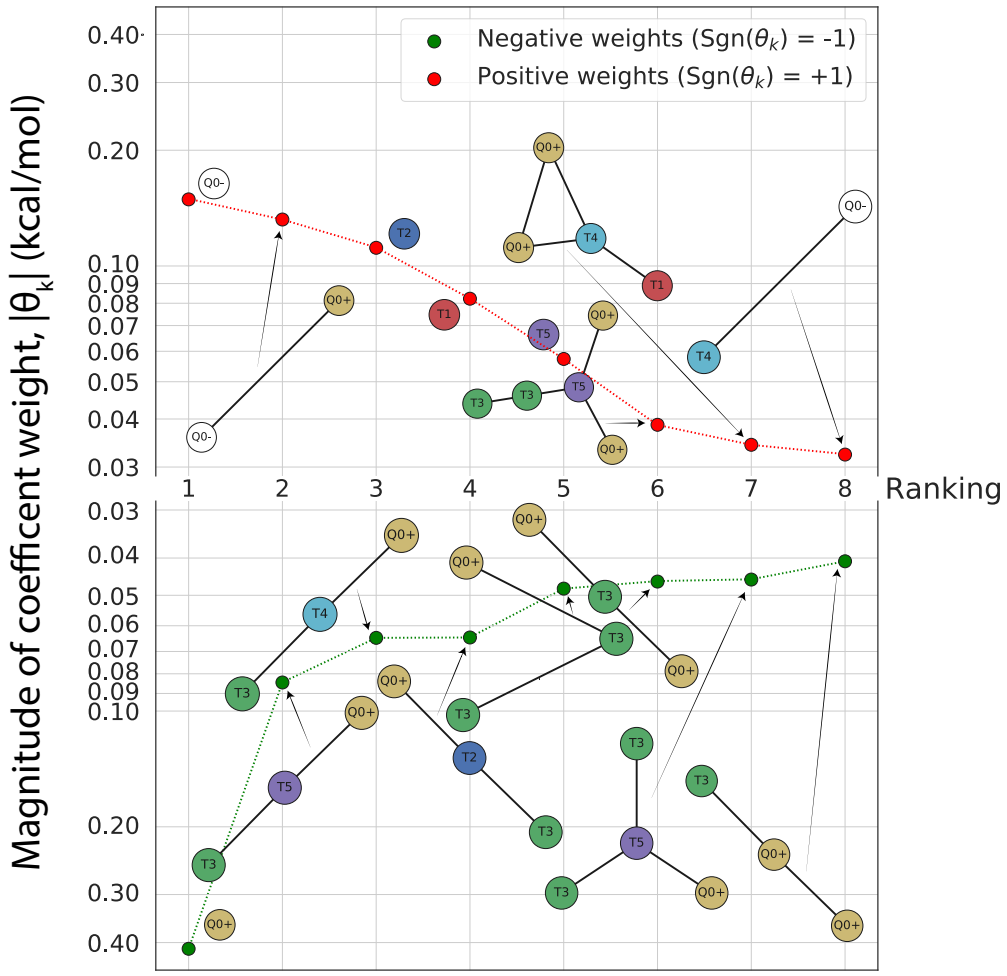


Figure 4.5: Rank-ordering of learned coefficient weights θ_k corresponding to CG motifs for predicting measured $\Delta\Delta G$. Each compound n is featurized by calculating the normalized frequency of all k topologically unique subgraphs consisting of 1-5 CG beads, denoted $F_{n,k}$. Using this featurization we perform LASSO regression to predict the calculated $\Delta\Delta G$ obtained from the alchemical free energy calculations using Eqn. 4.2: $\Delta\Delta G^{(n)}_{\text{predicted}} = \Delta\Delta G_{\text{mean}} + \sum_{k=1}^{k=1608} \theta_k F_{n,k}$. We then extract the associated largest and smallest learned coefficient weights θ_k from the reduced feature set retained by the LASSO model. The most negative coefficient weights correspond to chemical motifs indicative of small $\Delta\Delta G$ and good CL selectivity, while the most positive weights correspond to motifs predicative of large $\Delta\Delta G$ and poor CL selectivity. As the weights θ_k have units of kcal/mol these weights can be interpreted as the magnitude of the influence of the corresponding subgraph motif on CL selectivity.

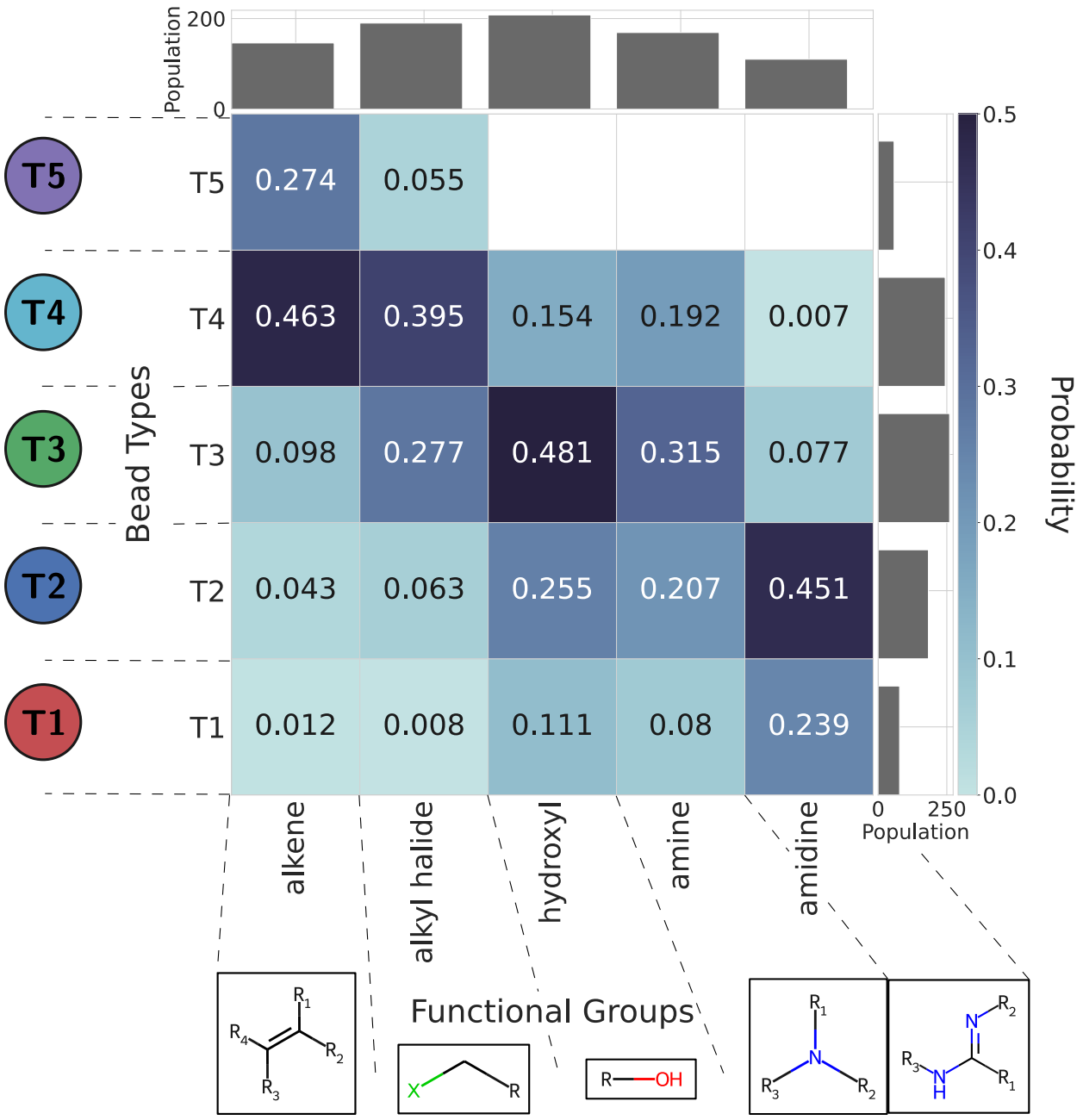


Figure 4.6: Five functional groups most frequently mapped to the uncharged beads of the *5+1* reduced Martini model.[314, 85] Dark colors represent higher mapping probabilities.

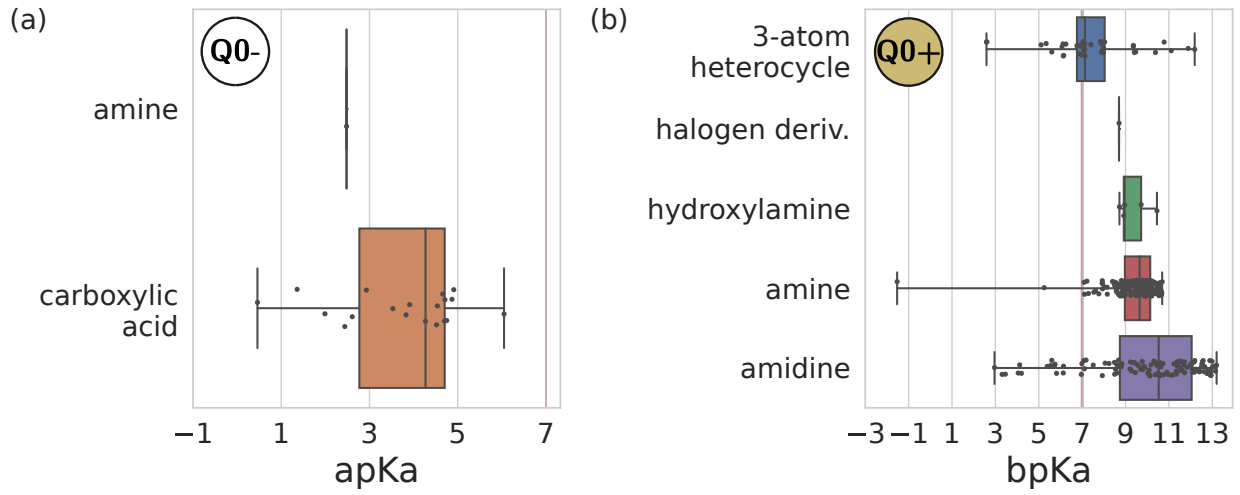


Figure 4.7: Predominant functional groups mapped to the charged bead type Q0± within our *5+1* reduced Martini model.[314, 85] (a) Functional groups with ap$K$a ≤ 7 are most likely to be negatively charged under physiological conditions (pH ≈ 7.3) and are mapped to the Q0- bead. (b) Those with bp$K$a ≥ 7 are most likely to be positively charged and are mapped to the Q0+ bead.

While a rigorous strategy for the backmapping of any CG molecule to all possible atomistic counterparts is beyond the scope of the present work, we illustrate in Fig. 4.8 the diversity of all-atom structures represented by a single coarse-grained molecule by presenting the range of GDB-derived small molecules whose coarse-grained mapping corresponds to a T3-T5 dimer as one of the chemical motifs correlated with increased CL selectivity (see Sec. 4.3.7). We identify a total of 2157 GDB-derived small molecules consistent with a T3-T5 dimer and render the chemical structures of seven of these in an attempt to convey the chemical diversity contained within this single coarse-grained dimer. We intentionally adopted a coarse-grained model for our screening procedure in order to reduce the size of chemical space and accelerate our simulations. This allowed us to efficiently identify CG representations of molecules predicted to have high CL selectivity, but results in enormous redundancy in the number of all-atom structures consistent with a single CG representation. Naturally the CG model is not able to rank compounds that map to the same CG representation. As such we show that CG modeling acts as a funnel to efficiently filter out

uninteresting compounds. The results yield a small set of top performing CG structures, which can be further backmapped to an atomistic resolution. The set of consistent all-atom representations can be filtered to further design other desirable characteristics, such as functional groups that endow fluorescence or other optical responses desirable for imaging agents. They also naturally lend themselves to validation and applications, through all-atom simulations or experiments. In the next section we pursue the latter strategy to validate our computational predictions using fluorescence-anisotropy experiments. In future work, we would also like to pursue all-atom calculations of the top-performing candidates identified in our CG calculations but currently face challenges associated with the absence of validated all-atom force field parameters and exceedingly high computational complexity of all-atom free-energy calculations in membrane systems. [338]

4.5.4 Validating simulations against experimental measurements

We test the validity of our CG predictions of thermodynamic selectivity and of the learned design rules by comparing calculated $\Delta\Delta G$ transfer free energies against experimental measurements of CL selectivity for two molecules: quinaldine red (4-[(E)-2-(1-ethylquinolin-1-ium-2-yl)ethenyl]-N,N-dimethylaniline) and benzothiazolium (3-butyl-2-methyl-1,3-benzothiazol-3-ium) (Fig. 4.9). We selected these molecules according to our learned design rules, they were therefore predicted to preferentially partition into CL membranes. These molecules both possess at least one ionizable group that will likely be positively charged at pH $\approx$ 7, hydrophobic character, and aromatic ring systems and partial charges that may participate in hydrogen bond formation. We experimentally measured the interaction of these two molecules with model membranes consisting of 100% 1,2-dioleoyl-sn-glycerol-3-phosphocholine (PC), 90% PC / 10% 1,2-dioleoyl-sn-glycerol-3-phosphoglycerol (PG), and 90% PC / 10% CL. We hypothesized that permeation of the molecules into the lipid membranes would affect lipid packing and order. This can be monitored using an experimental fluorescence anisotropy

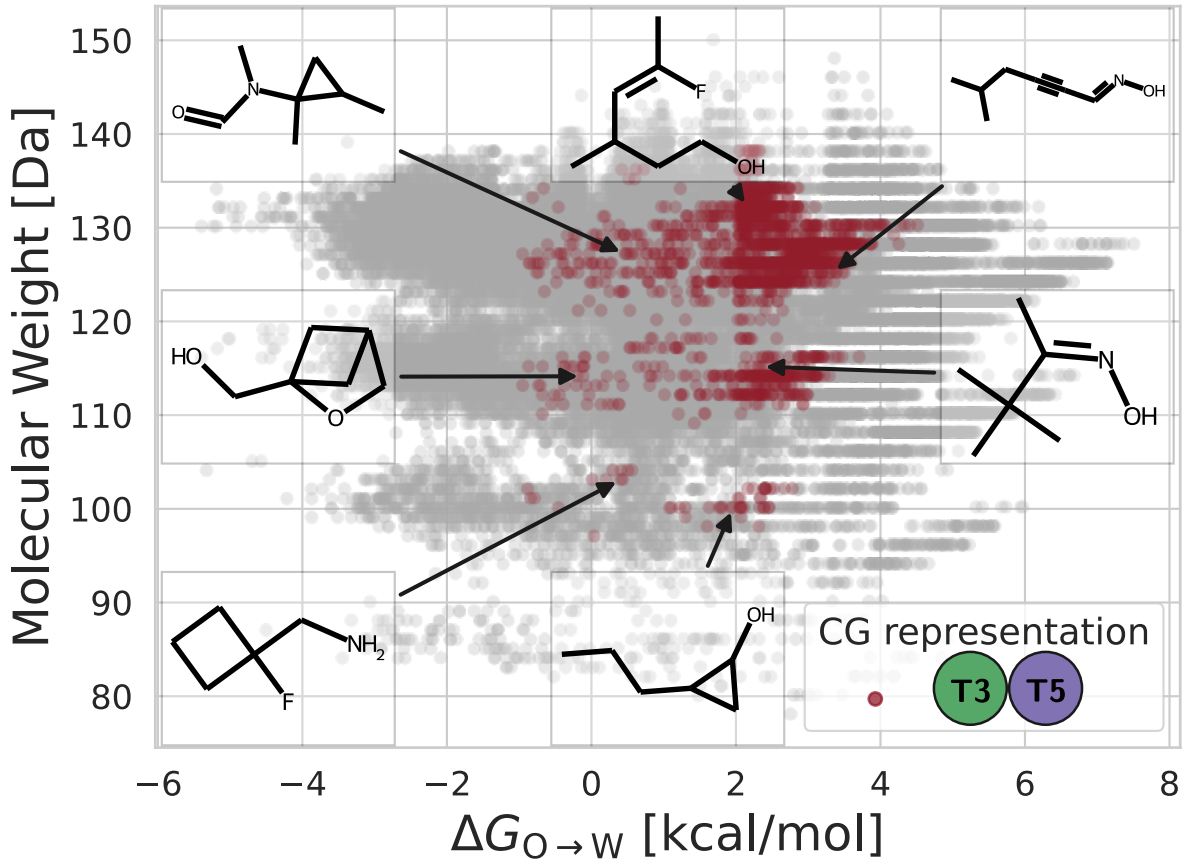


Figure 4.8: Backmapping of a coarse-grained T3-T5 dimer. The ensemble of 55639 GDB-derived small molecules consistent with any coarse-grained dimer are projected as grey points into the 2D plane spanned by the two order parameters of molecular weight and octane to water transfer free energy $\Delta G_{O\to W}$. We highlight in red the 2157 GDB-derived small molecules whose coarse-grained mapping corresponds to a T3-T5 dimer. The chemical structures of seven molecules are visualized. A large number of diverse chemical structures map to a particular CG representation.

assay to report the differential fluorescence anisotropy $\Delta A = A_s - A_\emptyset$ of a lipid membrane in presence of the candidate molecule A_s and in its absence $A_\emptyset$. Our hypothesis is that the two selected molecules should lead to a moderate increase in anisotropy ΔA if they incorporate into a membrane, and to a larger increase if they not only incorporate, but specifically interact with a lipid species in the membrane. The measurements of the pure PC membrane were added to differentiate between simple incorporation and specific interactions, since the latter were not to be expected with this particular phospholipid.

We present the results of these measurements in Fig. 4.9a. A negative control experiment confirms no change in ΔA in the absence of any molecular candidate. The addition of benzothiazolium led to slight increases in ΔA (i.e., decrease in membrane fluidity) for all three model membranes, but this increase was similar in all cases indicating no preferential interaction that would result in a differential response. Quinaldine red led to more significant increases in ΔA for all three membranes and a rank ordering in the magnitude of the response of PC/CL > PC/PG > PC, indicating a preferential partitioning into the CL-containing model membrane. These results indicate that our design rules successfully identified quinaldine red as a CL-selective molecule, whereas benzothiazolium was a false positive. Our learned design rules are a heuristic tool by which to identify candidate molecules, so we next sought to subject these two molecules to CG free energy calculations to determine if the ΔG and $\Delta\Delta G$ values recapitulated the experimental trends.

We constructed CG representations of quinaldine red and benzothiazolium under the *5+1* model and subjected these to alchemical free energy calculations. Our experimental measurements of ΔA for NAO were clouded by interferences between the fluorescence properties of the compound and the dye used for the anisotropy measurements (DPH). We therefore do not report the results in this work. We nevertheless also performed these calculations for NAO as a baseline comparison. As illustrated in Fig. 4.9b, all three compounds showed favorable partitioning from bulk water and octane into the interface regions of both the PG

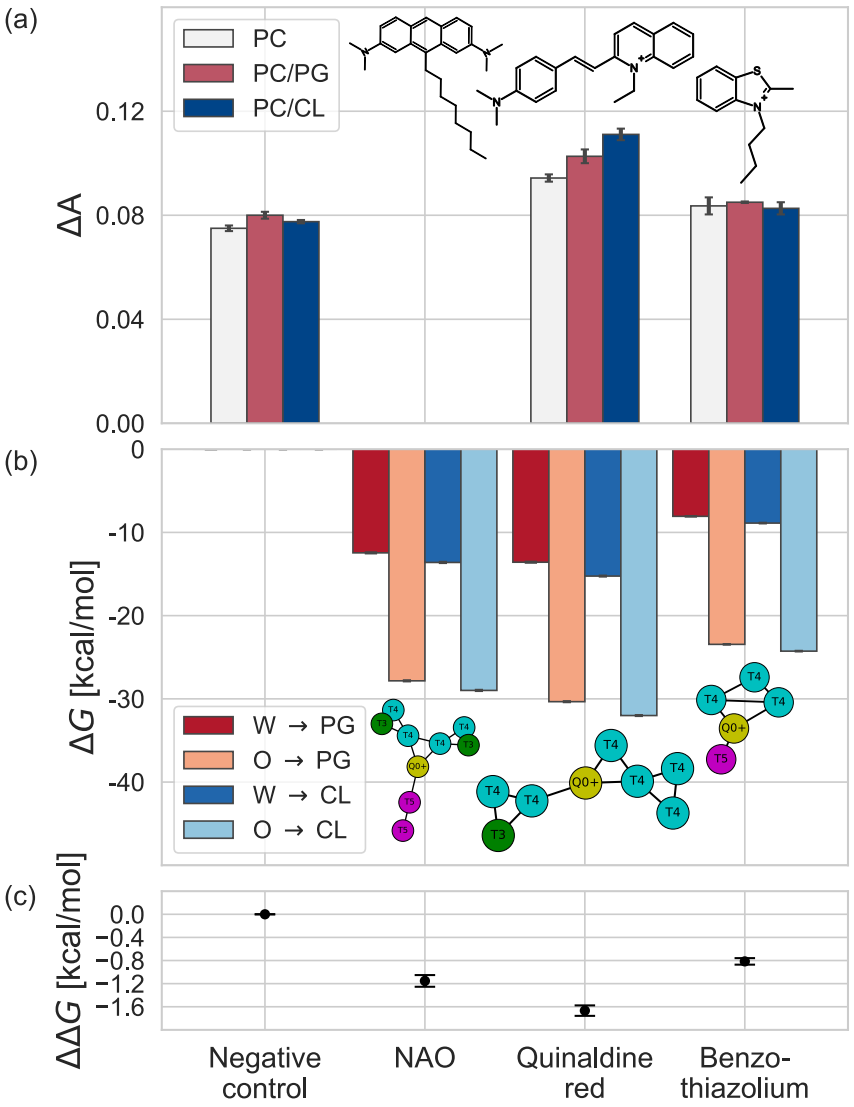


Figure 4.9: Comparison of experimental measurements and computational predictions of the selectivity of two candidate molecules quinaldine red and benzothiazolium selected according to our learned design rules within three model membranes consisting of 100% PC, 90% PC / 10% PG, and 90% PC / 10% CL. (a) Experimental measurements of the differential fluorescence anisotropy ΔA. The negative control confirms negligible differences between the different membranes in the measured value of ΔA. Measurements of ΔA for NOA were not reported due to technical issues. (b) Computational predictions of the transfer free energies $\Delta G_{\mathrm{W}\to\mathrm{I}}$ from water to the membrane interface and from octane (representative of the membrane midplane) to the interface $\Delta G_{\mathrm{O}\to\mathrm{I}}$ calculated using the *5+1* CG model and alchemical transformations. Calculated uncertainties of the computational results are smaller than the symbol size. (c) Computational predictions of the relative partitioning free energy $\Delta\Delta G = \Delta G^{\mathrm{CL}}_{\mathrm{W}\to\mathrm{I}} - \Delta G^{\mathrm{PG}}_{\mathrm{W}\to\mathrm{I}}$.

and the CL membranes (i.e., $\Delta G_{\mathrm{W}\to\mathrm{I}}$, $\Delta G_{\mathrm{O}\to\mathrm{I}} < 0$). These results are consistent with the experimental observations of elevated ΔA for all three model membranes that is indicative of spontaneous thermodynamic partitioning of the molecules into the membranes. As demonstrated in Fig. 4.9c, the calculated relative partitioning free energies $\Delta\Delta G$ follow the same trends as the experimental ΔA trends. In the case of benzothiazolium, the $\Delta\Delta G = -0.82$ kcal/mol is close to thermal energy $k_{\mathrm{B}}T \approx 0.6$ kcal/mol at room temperature, showing moderate CL selectivity. This is in reasonable agreement with the lack of difference in anisotropy change observed in the experiments. In contrast, the $\Delta\Delta G = -1.67$ kcal/mol of quinaldine red is more than twice as large as that for benzothiazolium and nearly three times larger than the thermal energy scale. This is consistent with the differential response in the experimental ΔA measurements showing significantly elevated selectivity of quinaldine red for the CL membrane environment. Finally, we observe that the $\Delta\Delta G = -1.15$ kcal/mol calculated for NAO is superior to that for benzothiazolium but inferior to that for quinaldine red, suggesting that the latter may offer superior CL selectivity.

4.6 Conclusions

The phospholipid cardiolipin (CL) is found exclusively within energy transducing membranes and constitutes approximately 20% of the composition of the inner mitochondrial membrane in eukaryotes.[76, 75] It plays a key metabolic role while having demonstrated capability to also behave as a biomarker for detecting a number of diverse pathologies. A standing challenge in developing reliable CL-based diagnostics lies in the lack of molecules that display selectivity in binding to CL compared to other phospholipid membranes. With the number of drug–like molecules being on the order of $\sim 10^{60}$, a major challenge is the efficient exploration and filtration of this enormous molecular design space. Using transferable CG models to systematically reduce the size of chemical compound space we mapped this all-atom space into a more manageable 124,327-member CG design space. Nevertheless, a

naïve Edisonian trial-and-error or random sampling within this design space remains prohibitively expensive. This motivated us to employ techniques in Bayesian optimization and deep representational learning to perform a directed data-driven search over the CG space of small organic molecules in an effort to minimize the data complexity, develop surrogate models to exploit the subtle structural differences in the CL and PG lipids to wield control over CL selectivity and efficiently discover the best performing candidates within an iterative active learning loop. After seven rounds of active learning and simulation of only 0.42% of the molecular design space, we identified candidate molecules with 184% better selectivity than the fluorescent dye NAO that is commonly used as a selective CL stain.

Analyzing our accumulated CG simulation data with interpretable linear models enabled automatic discovery of chemical motifs predictive of good CL selectivity. We were able to derive design rules correlating the presence/omission of functional groups to CL selectivity by probabilistically mapping each CG bead to possible all-atom functional groups and relating these to the automatically identified CG motifs. CL selectivity is linked to positively charged groups, hydrophobicity, and the ability to form hydrogen bonds. In contrast, negatively charged groups and overall polarity led to the molecule partitioning into the bulk water, and dominant apolarity led to a tendency of the candidate molecules to insert into the membrane midplane regions.

We tested our computational models against experimental measurements of CL selectivity for two candidate molecules selected using our learned design rules. The trends in our calculated transfer free energies and relative partitioning free energy differences are in good accord with experimental measurements providing support and validation of our computational approach.

In future work, we plan to conduct experimental testing of other candidate molecules identified by our learned design rules and CG screen and conduct all-atom free energy calculations with a second higher-resolution virtual screen. This work adopted thermodynamic

selectivity as the sole optimization objective for virtual screening without regard to the optical activity of the candidate molecules. Since the ultimate goal is to discover high-selectivity CL dyes, in subsequent work we plan to employ computational filtrations and/or multi-objective optimization strategies to discover molecules that are both highly selective for CL and possess either inherent fluorescent activity or can support the addition of fluorescent tags without compromising their thermodynamic behavior. Finally, we note that the coarse-grained chemical space embeddings and active learning search are generically transferable to other molecular design applications by replacing the thermodynamic selectivity prediction for a computational or experimental assay of the property of interest.

Data Availability

We present a full accounting of intermediate alchemical transfer free energies used to calculate $\Delta\Delta G$ CL selectivity; predicted $\Delta\Delta G$ from the terminal GPR model; neural network weights and training codes; GPR training codes; embeddings of the RAE; LASSO regression pipeline; demonstration of the active learning workflow; Python pipeline for generating Gromacs input files; Gromacs run parameter files; Python implementation of the Ertl algorithm with our extension; and previously unpublished datasets of small molecules containing five- or six-membered rings with CG representations.[334]

Supporting Information

Electronic supplementary information (ESI) available. See `https://doi.org/10.1039/D2SC00116K`

CHAPTER 5

COMPUTATIONAL DISCOVERY OF HIGH CHARGE MOBILITY SELF-ASSEMBLING π-CONJUGATED PEPTIDES

5.1 Abstract

Organic electronics offer a route toward electronically active biocompatible soft materials capable of interfacing with biological and living systems. One class of promising organic electronic materials are π-conjugated peptides, synthetic molecules comprising an aromatic core flanked by oligopeptides, that can be engineered to self-assemble into elongated nanostructures with emergent optoelectronic functionality. In this work, we combine molecular dynamics simulations with electronic structure and charge transport calculations to computationally screen for high charge mobility π-conjugated peptides and to elucidate design rules linking aromatic core character with charge mobility. We consider within our screening library variations in the aromatic core chemistry and length of the alkyl chains connecting the oligopeptide wings to the core. After completing our computational screen we identify particular π-conjugated peptides capable of producing self-assembled biocompatible nanoaggregates with predicted hole mobilities of 0.224 $cm^2/(Vs)$ and electron mobilities of 0.143 $cm^2/(Vs)$, and uncover design rules that enhance understanding of the molecular determinants of charge mobility within π-conjugated peptide assemblies.

This chapter is based on the work reported in: **Shmilovich, K.**, Yao, Y., Tovar, J. D., Katz, H. E., Schleife, A., & Ferguson, A. L. (2022). "Computational discovery of high charge mobility self-assembling π-conjugated peptides." *Molecular Systems Design & Engineering*, 7(5), 447-459.

5.2 Introduction

Organic semiconductors have demonstrated impressive performance in recent applications to bio-sensing [339] and in photovoltaics [340, 341]. Organic semiconductors maintain the advantage of being mechanically flexible, lightweight, and possessing inexpensive solution processability compared to traditional silicon-based semiconductors. One class of promising organic semiconductors are π-conjugated peptides composed of a central π-core flanked by oligopeptide wings representing biocompatible building blocks for the bottom-up construction of nanoaggregates with tunable optical and electronic properties [1, 3, 4, 9, 10, 11, 12]. These peptide based assemblies form elongated and hierarchical nanostructures driven by a combination of hydrogen bonding, van der Waals interactions, and intermolecular π-π stacking [13, 14, 15, 16]. Structural and optoelectronic characteristics of supramolecular π-conjugated assemblies are amenable to a variety of human controlled modulation such as salt concentration, temperature, pH, flow, and light [16, 17, 18, 19, 20, 21, 22]. Materials built from π-conjugated peptides are capable of long-range charge transport within aqueous and biologically relevant environments that can enable device construction of solar cells and field-effect transistors for a plethora of functional applications in photovoltaic energy harvesting and power generation [4, 28, 29, 30, 24, 31, 32, 33, 34]. Although crystalline electronic devices tend to support higher charge mobilities, organic electronics such as π-conjugated peptides benefit from biocompatibility and conformational flexibility that make them ideally suited to interface with biological and living systems such as photosynthetic machinery for energy harvesting or engineering interfaces with living cells. An important design challenge is to maximize charge mobility within π-conjugated peptide-based materials by controlling the molecular chemistry of these molecules.

Interchanging the different molecular building blocks that make up π-conjugated peptides enables precise tunablity and customizability over their emergent optical and electronic properties. For example, our previous work explored the influence that the variation of individual

amino acids within the peptide wing can have on the emergent supramolecular chirality and structure of self-assembled π-conjugated peptide nanoaggregates [26]. The role of alkyl spacers on molecular properties has previously been explored in the context of alkyl chains connecting the peptide wings to the core in π-conjugated peptides [27, 290], as well as for other systems such as zwitterionic polymer brushes [342, 343]. We also studied the impact of varying the length of the oligothiophene aromatic cores upon self-assembled π-conjugated peptide morphologies as a function of ionic strength, pH, temperature, and peptide concentration [22]. Computational work has employed all-atom and coarse-grained molecular dynamics simulations and electronic structure calculations to elucidate π-conjugated peptide self-assembly kinetics, thermodynamics, and electronic properties [33, 190, 18, 344]. Other works have engaged families of π-conjugated peptides using simulations and/or machine learning to establish design rules linking components of their chemical character to self-assembly and molecular packing behavior [131, 248, 345]. Recently, we developed an integrated computational-experimental framework to propose, synthesize, and characterize high-performing π-conjugated peptides with desirable electronic properties containing a fixed quaterthiophene core within the vast combinatorial space of varying length and composition peptide wings [346]. In this work, we build upon these foundations to explore the influence of π-core chemistry and the length of the alkyl spacer linking the core to the wings upon the self-assembled morphologies and emergent charge mobility within the resulting nanoaggregates.

A detailed understanding of the structure-(electronic) property relationships in π-conjugated peptides is a prerequisite toward systematic improvement and engineering of high charge mobility. Although theoretical developments in the transport properties of organic semiconductors and their dependence on intra- and inter-molecular morphology has proved challenging [347], a number of promising methods have been proposed in recent years to better understand the relationship between molecular structure and charge transport [348, 349]. For ex-

ample, in Ref. [48] the authors perform high-throughput virtual screening of novel crystalline dinaphtho[2,3-b:2′,3′-f]thieno[3,2-b]thiophene derivatives using Marcus theory [350, 351] to discover organic crystals with high levels of theoretical hole mobility. Here, we apply these methods for calculating charge mobility to π-conjugated peptides which are inherently disordered, biocompatible molecular systems. As π-conjugated peptide nanostructures lack a well-defined crystal structure, we employ molecular dynamics simulations to predict their molecular packing behavior and self-assembled morphologies. We then apply density functional theory to ascertain the capacity for charge mobility within these nanoaggregates. Chemical stability is assessed by comparing reduction/oxidation potential of the species to baseline water reduction/oxidation potentials to determine the p/n-type nature of each molecule, and charge mobility is quantified by adopting a Marcus hopping model [48] that predicts the transfer rate of holes/electrons between molecules. Combining density functional theory and molecular dynamics simulations in this way enables us to probe the electronic behavior of self-assembled π-conjugated peptide nanostructures and identify promising candidate molecules anticipated to display good charge mobility.

We employ our combined molecular dynamics simulations and density functional theory workflow to investigate charge transport properties within a library of π-conjugated peptides. In particular, we attempt to optimize both the aromatic chemistry among a number of synthetically available cores and the alkyl linker length between the aromatic core and the peptide wings of π-conjugated peptides (Fig. 5.1). These two molecular determinants are anticipated to affect both the structural arrangement and packing of the self-assembled nanoaggregate structures and the degree of electronic coupling between the cores and wings. Each π-conjugated peptide studied here is built with VEVAG peptide wings, which is a prototypical oligopeptide sequence commonly known to mediate self-assembly and promote intermolecular π-electron delocalization [49, 27, 28]. The oligopeptide wings are symmetric so that the molecule possesses two C-termini and the N-termini are covalently bonded to the

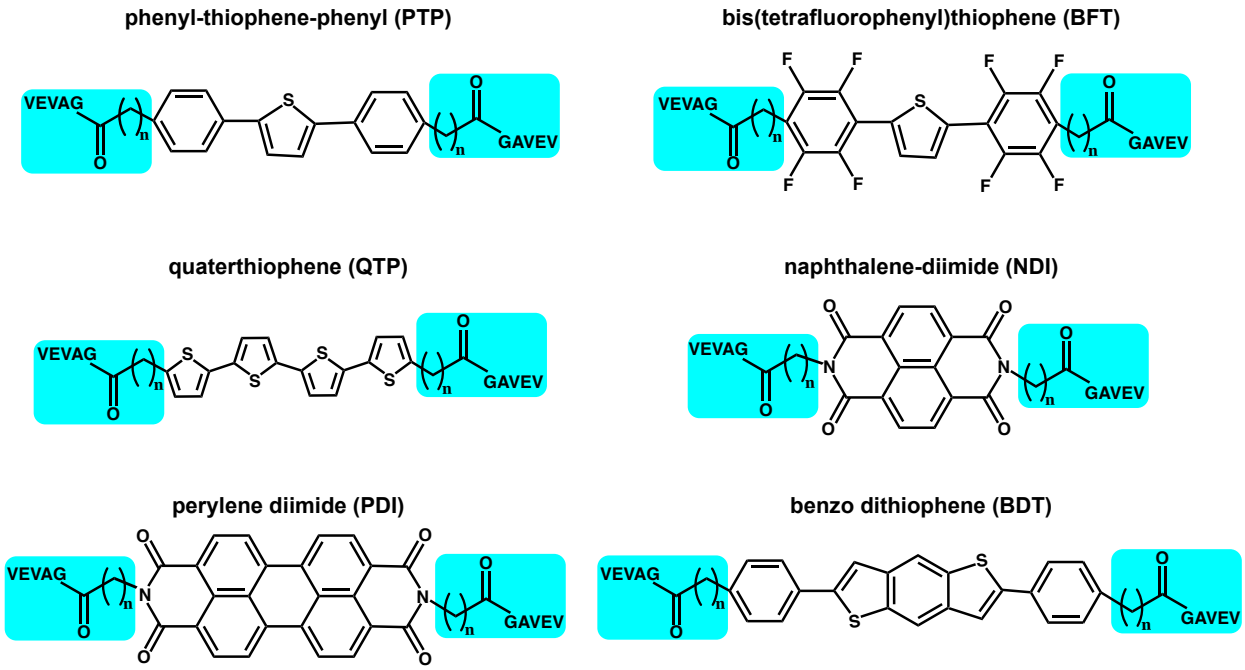


Figure 5.1: Chemical structure of the different π-conjugated peptide molecules built from six unique aromatic cores that we screen in this work. The mirror symmetric oligopeptide wings VEVAG-Π-GAVEV are highlighted in blue and which flank one of six different aromatic cores $\Pi \in \{\text{PTP, DFT, QTP, NDI, PDI, BDI}\}$ separated from the wings by alkyl spacers of length $n \in \{0, 1, 2, 3\}$. By considering all combinations of π-cores and alkyl chain lengths, our screening library comprises a total of $6 \times 4 = 24$ candidate π-conjugated peptide molecules.

π-core via an alkyl spacer. This arrangement enables pH control over assembly. At high pH the carboxyl groups at the C-termini and glutamic acid side chains are deprotonated leading to a formal (-4) charge on each molecule that prevents large-scale assembly by Coulombic repulsion. At low pH the molecules become charge neutral and self-assembly proceeds driven by hydrophobicity, hydrogen bonding, and π-π stacking [131, 346, 34].

The fixed VEVAG peptide wing flanks one of six aromatic cores – (i) phenyl-thiophene-phenyl (PTP), (ii) bis(tetrafluorophenyl)thiophene (BFT), (iii) quarterthiophene (QTP), (iv) naphthalene-diimide (NDI), (v) perylene diimide (PDI), and (vi) benzo dithiophene (BDT) – and is linked to the core by a variable length alkyl chain containing n=0, 1, 2, or 3 carbon atoms. The six aromatic cores were selected from a combination of cores commonly used within the π-conjugated peptide literature and standalone aromatic structures that are well-

established as performant organic semiconductor materials [50, 51, 52, 53, 28, 54, 23]. The range of lengths of the alkyl chains were selected based on the known negligible π-electron delocalization operative through saturated carbon chains which was anticipated to isolate and decouple the electronics of the central π-electron core from any electronic influence of the peptide carboxamide linkers. Each π-conjugated peptide is then constructed with VEVAG oligopeptides affixed to a π-core and separated from the peptide wing on each end with an alkyl spacer. By considering all combinations of π-cores and alkyl chain lengths, our library comprises a total of $6\times4=24$ candidate π-conjugated peptide molecules (Fig. 5.1). Throughout this work we abbreviate the collection of molecules with a fixed aromatic core and varying spacer length as 'Lib-X', where X is the three-letter abbreviation corresponding to each aromatic core (e.g., Lib-PTP corresponds to π-conjugated peptides with a fixed phenyl-thiophene-phenyl core and the set of all $n \in \{0, 1, 2, 3\}$ spacer lengths).

5.3 Methods

5.3.1 Overview of the computational pipeline

We establish a computational pipeline combining classical molecular dynamics (MD) simulations with density functional theory (DFT) electronic structure and Marcus theory charge mobility predictions to computationally screen π-conjugated peptides and discover candidate molecules predicted to self-assemble into nanoaggregates with high charge mobility. Unlike crystalline semiconductors with a well-defined crystal structure that can be directly probed with electronic structure calculations, π-conjugated peptide require us to first predict their molecular packing behavior and emergent nanoaggregate morphologies using MD simulations. We then perform DFT calculations to assess the stability and mobility of holes and electrons within these nanoaggregates. The predicted charge mobilities are a function of both the self-assembled morphology and the inherent electronic properties of each can-

didate molecule. Our integrated MD/DFT screen can be represented schematically by the five-step protocol in Fig. 5.2, in which we **(i)** perform MD simulations of the assembly in water, **(ii)** determine hole/electron stability, **(iii)** calculate reorganization energy, **(iv)** calculate the transfer integral, and **(v)** combine the reorganization energy and transfer integral calculations to compute the intermolecular charge mobility from Marcus theory.

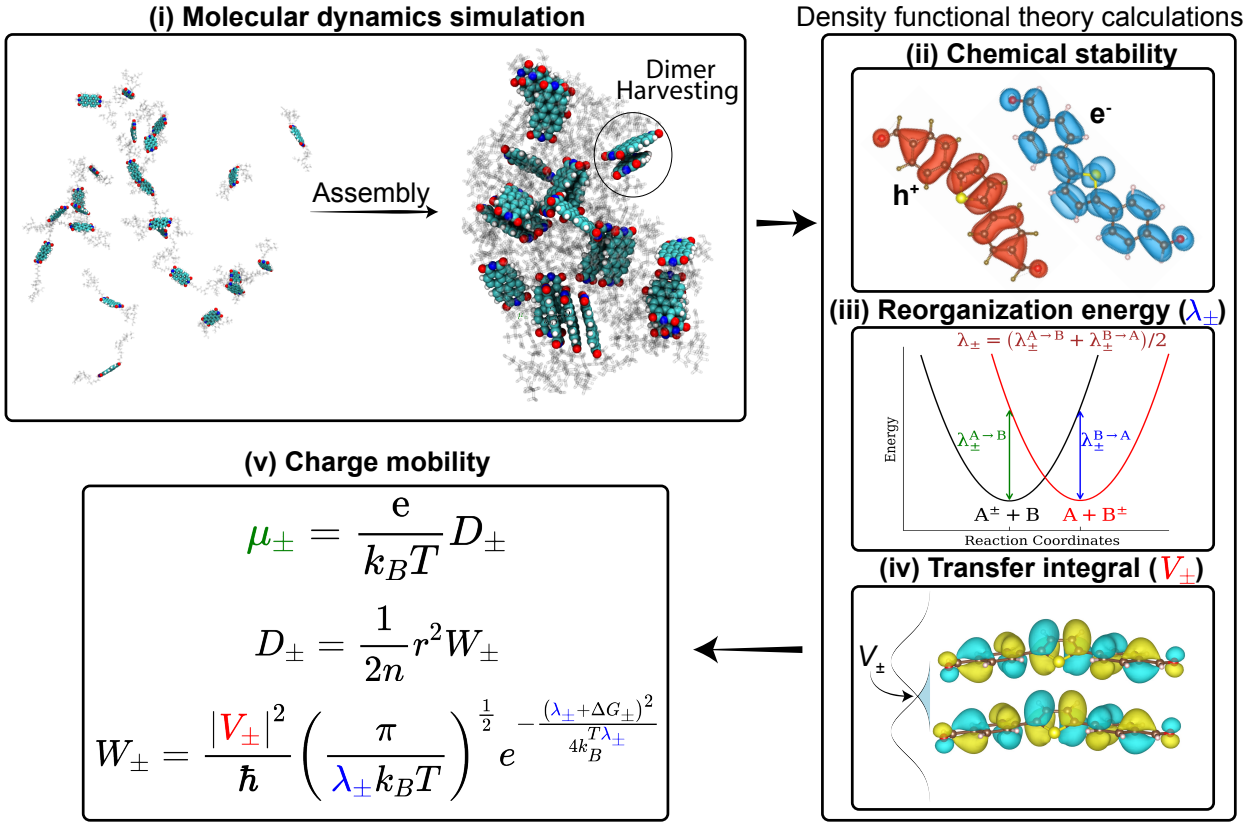


Figure 5.2: Schematic illustration of our computational pipeline for measuring π-conjugated peptide charge mobility. **(i)** All-atom classical molecular dynamics (MD) simulations are conducted to predict the morphology of the self-assembled aggregates formed by each π-conjugated peptide. Electronically interacting dimer pairs are harvested and passed to density functional theory (DFT) calculations. DFT calculations are performed to compute **(ii)** the chemical stability of molecules supporting a hole or electron within an aqueous environment, **(iii)** the reorganization energy $\lambda_\pm$, and **(iv)** the transfer integral $V_\pm$. **(v)** Marcus theory is then used to predict the intermolecular drift mobility $\mu_\pm$ of holes/electrons within the nanoaggregates as a figure of merit for the potential optoelectronic functionality of the candidate π-conjugated peptide. Molecular visualizations were constructed using Visual Molecular Dynamics (VMD) [352] and molecular orbitals visualized using VESTA [353].

5.3.2 MD calculations

The GROMACS 2019.2 simulation suite was used to perform all molecular dynamics simulations [257]. The Restrained Electrostatic Potential (RESP) method [354, 355] was used to derive partial charges for each molecule where calculations were performed at the B3LYP/6-31G(d) level of theory using the Gaussian 16 quantum chemistry software [356]. All-atom topologies for each molecule were then generated using the AnteChamber PYthon Parser interfaceE (ACPYPE) [357] wrapper for ANTECHAMBER (AmberTools17.0) [358] and modeled with the AMBER99SB-IDLN [259] forcefield. Each molecule was prepared in a charge neutral state with fully protonated glutamic acid residues corresponding to the low-pH conditions under which acid-triggered self-assembly proceeds [34]. Newton's equations of motion were solved and integrated forward with the leapfrog algorithm using a 2 fs timestep [204]. Covalent bonds involving hydrogen atoms were constrained using the LINCS algorithm [359]. Lennard-Jones interactions were shifted smoothly to zero at a distance of 1.0 nm. Coulombic interactions were treated using particle mesh Ewald (PME) with a 1.0 nm real space cutoff and 0.16 nm Fourier grid spacing, and these values optimized for performance during runtime [360]. Each system was initialized by placing 24 identical molecules randomly throughout a 12 x 12 x 12 nm^3 simulation box with three-dimensional periodic boundary conditions, and subsequently solvating them with TIP3P water [258]. Steepest descent energy minimization was then performed until forces larger than 1000 kJ/mol·nm were removed. Initial velocities were then sampled from a Maxwell-Boltzmann distribution at 300 K to perform an NVT equilibration at 300 K for 100 ps and subsequently an NPT equilibration at 300 K and 1 bar for 100 ps. Production runs were performed in the NPT ensemble for 500 ns at 300 K and 1 bar, where temperature was controlled using the velocity rescaling thermostat [260] and pressure controlled using the Parrinello-Rahman barostat [261]. All simulations were performed on a single NVIDIA RTX 2080 Ti GPU achieving execution speeds of ∼90 ns/day. These molecular dynamics simulation are carried out for all 4×6=24

molecules in our molecular design library. Input files and simulation trajectories for our MD calculations are provided as detailed in the Data Availability statement.

Over the course of the MD simulations, the initially monodisperse monomers self-assemble into nanoaggregates driven by van der Waals interactions, hydrogen bonding, and π-π stacking [131, 346]. Nanoaggregate snapshots of Lib-PDI spacer 2 are visualized in Fig. 2, with the terminal self-assembled nanoaggregates for each π-conjugated peptide simulated in this work shown in Fig. S1a in the ESI†. To predict the charge transport properties within these nanoaggregates, we consider nearest neighbor interactions between dimers within the multimeric assembly, a method that is commonly employed in predicting the transport properties in organic crystals and amorphous organic materials[349]. Rather than exhaustively considering all nearest neighbors as molecular dimers we focus our sampling to pairs of molecules most likely to interact electronically. Prior x-ray scattering and energy minimization calculations show that associated π-conjugated peptide molecules lie within a center of mass (COM) separation of $\sim$0.46 – 0.47 nm [361, 362]. Based on this observation, we consider a pair of molecules to be an electronically interacting dimer within a given simulation snapshot if the COM distance $d_{i,j}$ between neighboring aromatic cores of molecules i and j satisfies $d_{i,j} \leq 0.5$ nm. We have previously reported that the total number of dimers serves as a suitable metric to evaluate convergence of the structural evolution of the π-conjugated peptide self-assembled nanoaggregates [131, 346]. The total number of dimers meeting the $d_{i,j} \leq 0.5$ nm criterion to approximately stabilize after 375 ns (Fig. S1b in the ESI†), motivating us to extract dimers from frames harvested every 100 ps within the final 125 ns of each 500 ns simulation. We randomly select 50 dimer configurations to pass forward to the DFT calculations as a sufficiently large ensemble to gather reliable statistics while balancing the high computational cost of DFT simulations.

5.3.3 DFT calculations

As an initial preprocessing step for the dimer configurations extracted from the molecular dynamics simulations, the oligopeptide wings are removed from the π-conjugated peptide geometries and the electronic structure calculations are performed only on aromatic cores that include the alkyl spacer capped on each end with terminal aldehyde groups. The reasoning for this pre-processing step is two-fold. First, the various candidate molecules differ from one another in the identity of the π-core and the length of the alkyl spacer, however the identity of the oligopeptide wings is held constant for all candidates. The primary source of optoelectronic functionality in self-assembled π-conjugated peptide nanostructures stems from π-electron delocalization fostering charge transfer between neighboring aromatic cores [190, 48, 363]. The oligopeptide wings implicitly affect the charge mobility by mediating the relative intermolecular packing between molecules in the self-assembled nanoaggregates, but are understood to be quite electronically insulated from π-cores, particularly for long alkyl spacer lengths [27]. Secondly, the oligopeptide wings comprise a large number of atoms and therefore incur considerable expense within the electronic structure calculations. Assuming that intermolecular charge transfer is mainly governed by the π-cores, we truncate our dimer structures to only possess these most relevant components for charge mobility and enable a larger number of representative dimer configurations to be considered within our ensemble by greatly alleviating the computational expense of each single calculation.

We perform DFT calculations of the chemical stability, reorganization energy, and transfer integral. All calculations are performed on 50 dimer configurations extracted from MD simulations from which we estimate the mean and variability of the these quantities within the structural ensemble. For the transfer integral calculations, the orbital coefficients, orbital energies, and overlap matrices are calculated using DFT following the scheme described in Sec. 5.3.5. For the reorganization energy and chemical stability calculations, the geometries of each π-conjugated peptide monomer are relaxed separately from one another in the

gas phase with single point energy calculations independently performed on each monomer. Solvation effects are additionally included for the chemical stability calculations by incorporating the polarizable continuum model (PCM) [364] where water is used as the solvent. All DFT calculations employ the B3LYP[365, 366] hybrid functional and the 6-311+G(D,P) basis set and are performed using the Gaussian 16 quantum chemistry code [356]. Input files and post-processing scripts for our DFT calculations and tabulated results of our chemical stability, reorganization energy, transfer integral, and charge mobility calculations are provided as detailed in the Data Availability statement.

5.3.4 Chemical Stability

The electrochemical reaction between water and a charged state polymer is a primary charge trapping and degradation pathway [367, 368]. The chemical stability of the n-type organic semiconductor can be determined by evaluating the difference between the oxidation potential of the polymer anion and the reduction potential for the half reaction involving oxygen. Similarly, the chemical stability of the p-type semiconductor can be determined by the difference between the reduction potential of the polymer cation and the oxidation potential for the half reaction involving water [367]. Assuming a Standard Hydrogen Electrode (SHE) potential, the oxidation electrode potential of water is +1.23 V (SHE),

$$O_2 + 4H^+ + 4e^- \rightarrow 2H_2O, \tag{5.1}$$

and the reduction electrode potential is −0.83 V (SHE) [369],

$$2H_2O + 2e^- \rightarrow H_2 + 2OH^-. \tag{5.2}$$

For each monomer A within a dimer configuration the oxidation (+)/reduction (−) potential is given by,

$$\pm(E_{\pm}(\mathrm{A}^{\pm}) - E_0(\mathrm{A}^0)), \tag{5.3}$$

where E_+, E_-, and E_0 denote the total energy calculated in one of the cationic (+), anionic (−), or neutral (0) states, respectively. These total energies are computed for a monomer geometry given by the argument $E(\cdot)$ relaxed in either the cationic A^+, anionic A^-, or neutral A^0 charge state. We screen for candidate molecules likely to have a stable cationic state with an ionization potential smaller than +1.23 V (SHE), and a stable anionic state with an electron affinity greater than −0.83 V (SHE). Molecule libraries which meet this criterion for a majority of spacer lengths are subsequently passed to further computational screening of their charge transport properties to identify candidates with the most potential as functional organic semiconductors. By conducting this chemical stability screen, we assure that only molecules capable of stably supporting holes or electrons within an aqueous environment are passed on to the charge mobility calculations.

5.3.5 Charge Mobility

The predicted charge mobility within the self-assembled nanoaggregate is a function of both the morphology of the aggregate and the intrinsic chemical properties of the constituent π-conjugated peptide monomers, and serves as the figure of merit by which to rank the performance of each candidate molecule. To measure charge mobility we adopt the hopping model described by Marcus theory [350, 351]. We aim to maximize the drift mobility of holes μ_+ and electrons μ_- in molecular dimers given by the Einstein relation,

$$\mu_{\pm} = \frac{\mathrm{e}}{k_B T} D_{\pm}, \tag{5.4}$$

where e is the electron charge. $D_{\pm}$ is the hopping rate of holes/electrons and can be modeled under the dimer approximation as,

$$D_{\pm} = \frac{1}{2n} r^2 W_{\pm}, \tag{5.5}$$

where n=3 is the spatial dimension, r is the COM distance between two neighboring molecules, and $W_{\pm}$ is given by,

$$W_{\pm} = \frac{|V_{\pm}|^2}{\hbar} \left(\frac{\pi}{\lambda_{\pm} k_B T} \right)^{\frac{1}{2}} e^{-\frac{(\lambda_{\pm} + \Delta G_{\pm})^2}{4 k_B T \lambda_{\pm}}}, \tag{5.6}$$

where $V_{\pm}$ is the transfer integral, $\lambda_{\pm}$ is the reorganization energy, k_B is the Boltzmann constant, $\hbar$ is the Planck constant, $\Delta G_{\pm}$ is the change of Gibbs free energy of the charge exchange process, and T=300 K is the temperature. In the following sections we present a detailed explication of how the reorganization energy and transfer integral terms are calculated.

Reorganization Energy

The reorganization energy $\lambda_{\pm}$ measures the change in the free energy of the reactant state to its equilibrium configuration, which describes the energy required to relax the molecular geometry upon charge transfer. Smaller values for the reorganization energy imply less energetic hindrance upon charge transfer and result in higher charge mobility (Eqns. 5.4-5.6). Given a molecular dimer configuration, the reorganization energy $\lambda_{\pm}^{\mathrm{A}\to\mathrm{B}}$ for the charge transfer process $\mathrm{A}^{\pm} + \mathrm{B} \to \mathrm{A} + \mathrm{B}^{\pm}$ from, arbitrarily labeled, monomer A to monomer B, is given by a four-point energy calculation scheme [370, 371],

$$\lambda_{\pm}^{\mathrm{A}\to\mathrm{B}} = E_{\pm}(\mathrm{B}^0) - E_{\pm}(\mathrm{B}^{\pm}) + E_0(\mathrm{A}^{\pm}) - E_0(\mathrm{A}^0), \tag{5.7}$$

where each term is a single point energy calculation performed in the ionic state given by the subscript, one of cationic E_+, anionic E_-, or neutral E_0, and in the configuration determined by a geometry optimization that is performed in either the cationic A^+, anionic A^-, or neutral A^0 charge state given in the argument $E(\cdot)$. Calculating the reorganization energy for a molecular dimer is complicated by the fact that our configurations are taken from molecular dynamics simulation snapshots where the two monomers will be in different geometries, and therefore $\lambda_{\pm}^{A\to B} \neq \lambda_{\pm}^{B\to A}$. We estimate the reorganization energy as the average of these two charge transfer processes [370, 372],

$$\lambda_{\pm} = \frac{(\lambda_{\pm}^{A\to B} + \lambda_{\pm}^{B\to A})}{2}, \tag{5.8}$$

where $\lambda_{\pm}^{A\to B}$ and $\lambda_{\pm}^{B\to A}$ are given in Eqn. 5.7. By performing this average in Eqn. 5.8 the reorganization energy $\lambda_{\pm}$ for a dimer configuration remains permutationally invariant and implicitly considers both possible charge transfer processes.

Transfer Integral

The transfer integral $V_{\pm}$ can be understood as characterizing the strength of electronic coupling between the two π-conjugated peptide monomer orbitals. Larger values of the transfer integral imply more electronic overlap between neighboring π-conjugated peptide molecules and result in higher charge mobility (Eqns. 5.4-5.6). Under the assumption that only the frontier orbital of each monomer interacts, the transfer integral is defined as $V_{\pm} = \langle \phi_A | H | \phi_B \rangle$, where ϕ_A and ϕ_B represent the frontier orbital of (arbitrarily labeled) monomer A or B, and H is the Kohn-Sham Hamiltonian. We compute transfer integrals following the protocol reported in Ref. [373] and illustrated schematically in Fig. S2 in the ESI†. Under the approximation that only the frontier orbitals will interact, for hole transport, ϕ_A and ϕ_B are given by the HOMO orbitals, while for electron transport they are given by the LUMO

orbitals. The matrix elements of the Kohn-Sham Hamiltonian $H_{\mathrm{AB}} = \langle\phi_{\mathrm{A}}|H|\phi_{\mathrm{B}}\rangle$ and the overlap matrix $S_{\mathrm{AB}} = \langle\phi_{\mathrm{A}}|\phi_{\mathrm{B}}\rangle$ are calculated in the basis of the frontier orbitals of each monomer. Due to the non-orthonormality of the monomer frontier orbitals, the computed matrix element H_{AB} is then transformed into an orthonormalized basis by performing the Löwdin transformation, $\hat{H}_{\mathrm{AB}} = S_{\mathrm{AB}}^{-\frac{1}{2}} H_{\mathrm{AB}} S_{\mathrm{AB}}^{-\frac{1}{2}}$. The value for the transfer integral $V_{\pm}$ is then given by the off-diagonal terms of $\hat{H}_{\mathrm{AB}}$. We verify that the transfer integrals computed for isolated dimers are in good agreement with significantly more expensive calculations for the same dimer embedded within larger tetramer stacks (Fig. S3, Table S1 in the ESI†). This demonstrates that transfer integrals computed at the dimer level are good approximations for those within self-assembled nanoaggregates comprised of multiple associated molecules.

Approximating $\Delta G_{\pm}$=0

The change in Gibbs free energy $\Delta G_{\pm}^{\mathrm{A}\to\mathrm{B}}$ for a molecular dimer configuration measures the energetic gain or loss upon charge transfer from (arbitrarily labeled) monomer A to monomer B. This free energy difference is given by[372],

$$\Delta G_{\pm}^{\mathrm{A}\to\mathrm{B}} = G_{\pm}(\mathrm{B}^{\pm}) + G_0(\mathrm{A}^0) - \Big(G_{\pm}(\mathrm{A}^{\pm}) + G_0(\mathrm{B}^0)\Big)\,, \tag{5.9}$$

where each term calculates the Gibbs free energy in the ionic state given in the subscript for monomer configurations obtained from performing geometry optimization in the charge state provided in the argument $G_{\pm}(\cdot)$. We note that explicitly calculating $\Delta G_{\pm}$ in Eqn. 5.9 is nontrivial due to entropic contributions in the Gibbs free energy at finite temperatures. Practitioners applying Marcus theory to calculate charge transfer of organic semiconductors have commonly taken $\Delta G_{\pm} = 0$ when the individual monomers have well-defined crystal structures [48]. In crystalline systems the two charge carrying species A and B will have identical geometries resulting in the change in the Gibbs free to be precisely zero independent

of the direction of the charge transfer process $\Delta G_{\pm}^{\mathrm{A}\to\mathrm{B}} = \Delta G_{\pm}^{\mathrm{B}\to\mathrm{A}} = 0$. The self-assembled π-conjugated nanoaggregates lack a well-defined crystal structure such that the two monomers in each electronically interacting dimer pair have slightly different geometries. This symmetry breaking will, in principle, result in nonzero $\Delta G_{\pm}$ for any single configuration, but it is expected that $\Delta G_{\pm} \to 0$ when averaged over a sufficiently large ensemble. In Fig. S4 in the ESI†, we test this assumption by calculating the total energy difference $\Delta E = E_0(\mathrm{A}^0) - E_0(\mathrm{B}^0)$ within dimer pairs of the Lib-PTP system. We confirm that the average value for ΔE calculated over a progressively larger ensembles approaches zero $\Delta E \to 0$, as is expected by the indistinguishability of the two arbitrarily labeled monomers A and B. We adopt ΔE as a proxy for $\Delta G_{\pm}$ as the latter is computationally prohibitively expensive to calculate for large ensembles, but the observation of $\Delta E \to 0$ implies by indistinguishability arguments that $\Delta G_{\pm} \to 0$. As a result, although we appreciate that $\Delta G_{\pm} \neq 0$ for any individual configuration, the computational challenges in the finite temperature entropy calculation lead us to follow Ref. [48] and make the simplifying assumption that $\Delta G_{\pm} = 0$ that is valid in the limit of sufficiently large dimer ensembles.

5.4 Results and Discussion

We now present the results of our MD and DFT calculations and their integration with Marcus theory to predict charge mobilities. We first present the results of our chemical stability calculations to identify stable cationic/anionic species (Fig. 5.3). The reorganization energy (Fig. 5.4) and transfer integral (Fig. 5.5) calculations are then combined using Eqns. 5.4-5.6 to calculate the predicted charge mobility (Fig. 5.6).

5.4.1 Chemical Stability

A prerequisite to charge mobility within our molecular nanoaggregates in aqueous environments is their ability to stably support electronically charged states in water to prevent charge

trapping and degradation via electrochemical reactions with water molecules [367, 368]. By reference to the electrochemical water half reactions [367] and assuming a Standard Hydrogen Electrode (SHE) potential [369], the oxidation potential of water is +1.23 V and the reduction potential is −0.83 V. Oligopeptides with ionization potentials smaller than +1.23 V (SHE) are capable of stably supporting the cationic state and can potentially serve as p-type organic semiconductor; those with electron affinities greater than −0.83 V (SHE) are capable of stably supporting the anionic state and can potentially serve as n-type semiconductors. We report in Fig. 5.3a the ionization potential of each candidate molecule in our five libraries relative to the reduction potential of water, and in Fig. 5.3b their electron affinities relative to the oxidation potential of water.

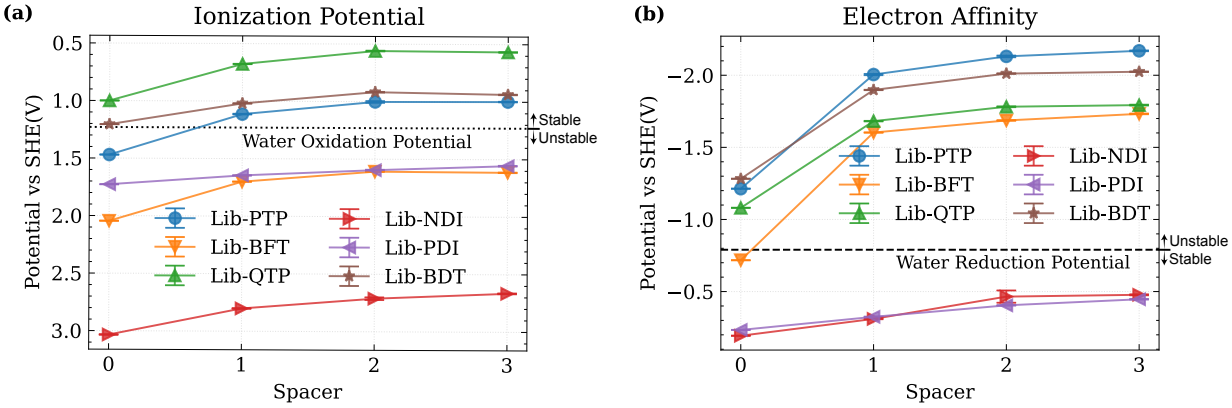


Figure 5.3: The chemical stability of the cationic and anionic states of each oligopeptide in water. (a) Calculated ionization potentials relative to the water oxidation potential of +1.23 V (SHE). Molecules residing above the dashed horizontal line are predicted to support stable cationic states and serve as putative p-type semiconductors. (b) Calculated electron affinities relative to the water reduction potential of −0.83 V (SHE). Molecules residing below the dashed line are predicted to support stable anionic states and serve as putative n-type semiconductors.

We observe from Fig. 5.3a that, with the exception of the Lib-PTP spacer 0 molecule, molecules with all spacer lengths within the Lib-PTP, Lib-QTP, and Lib-BDT libraries display stable cationic states indicating that they are potential candidates as p-type semiconductors. The remaining libraries, Lib-BFT, Lib-NDI, and Lib-PDI are not predicted to support electrochemically stable cations and we do not consider them further as potential

p-type semiconductors. We observe from Fig. 5.3b that all molecules within the Lib-NDI and Lib-PDI libraries have stable anionic states indicating that they are potential candidates for n-type semiconductors. With the exception of the Lib-BFT spacer 0 molecule that supports a marginally stable anion, none of the molecules in the Lib-PTP, Lib-BFT, Lib-QTP, and Lib-BDT libraries are predicted to support stable anionic states and are dropped from further consideration as potential n-type semiconductors. A recent review in Ref. [374] highlights NDI derivatives and PDI derivatives as stable n-type semiconductors, which is consistent with our predictions. Only Lib-BFT was predicted to support neither stable cationic not stable anionic states in water. This finding is consistent with previous work showing that both hole and electron injection into this pi system are likely to be unfavorable by more than ∼1 V relative to standard calomel or silver chloride reference electrodes, suggesting that neither holes nor electrons would be particularly stable [54].

5.4.2 Reorganization Energy

Having completed an initial screen for chemical stability, we proceed to interrogate the charge mobility of the candidate aromatic cores displaying stability in water solvent beginning by calculating the reorganization energy. The reorganization energy describes the energetic gain that results from the distortion of the equilibrium geometry between neutral state and charged state. A small reorganization energy suggests the capability of fast photo-induced charge transfer and is desirable for high charge mobility (c.f., Eqn. 5.6). We present the DFT-calculated reorganization energies for holes and electrons in all chemically stable species within Fig. 5.4. We note that the reorganization energy associated with hole transport is smallest for Lib-BDT (Fig. 5.4a), whereas that associated with electron transport is smallest for Lib-PDI (Fig. 5.4b). As a point of comparison, the reorganization energy of dinaphtho[2,3-b:2′,3′-f]thieno[3,2-b]thiophene (DNTT) and its derivatives, which are widely employed in organic electronic devices[375, 376], range from 75 meV to 188 meV[48], which

is about 55 meV (22%) lower than our best reported reorganization energy in Lib-BDT spacer 0 (243 meV). On the other hand, the reorganization energy for naphthodithiophene diimide (NDTI), a recently synthesized derivative of widely studied n-type semiconductors NDI [374], is around 286 meV[377], which is around 26% lower than our measured Lib-NDI spacer 1 (388 meV) and only 4% lower than Lib-PDI spacer 1 (301 meV).

Although our trends are non-monotonic, we find that reorganization energy tends to increase with increasing spacer size, degrading the charge mobility (c.f., Eqn. 5.6). To quantify this observation we measure the Pearson correlation coefficient ρ between spacer length and reorganization energy. The correlation coefficients are quite large and positive for the cationic systems: Lib-PTP ($\rho = 0.49$; p-value=2.4×10^{-13}), Lib-QTP ($\rho = 0.56$; p-value=3.3×10^{-18}), and Lib-BDT ($\rho = 0.6$; p-value=2.4×10^{-13}). They are weaker but still positive for the anionic systems: Lib-NDI ($\rho = 0.25$; p-value=3.4×10^{-4}) and Lib-PDI ($\rho = 0.25$; p-value=4.7×10^{-4}). We propose that the weaker trends for Lib-NDI and Lib-PDI may stem from these two libraries being composed of more rigid aromatic cores compared to Lib-PTP, Lib-QTP, or Lib-BDT. Lastly, as a general trend for both hole and electron transport, we notice that independent of spacer size the more rigid aromatic cores composed of larger fused ring systems tend to result in smaller reorganization energies across the board, evidenced by Lib-BDT for holes and Lib-PDI for electrons.

5.4.3 Transfer Integral

The transfer integral describes the strength of electronic coupling between neighboring molecules and is a strong function of the intermolecular distance and orientation. Large transfer integrals are desirable for good charge mobilities (c.f., Eqn. 5.6) and are associated with close structural alignment within the molecular pair leading to substantial overlap of electronic wavefunctions between the two molecules. Lower values for the transfer integral suggest weaker overlap of the individual molecular wavefunctions due to large intermolec-

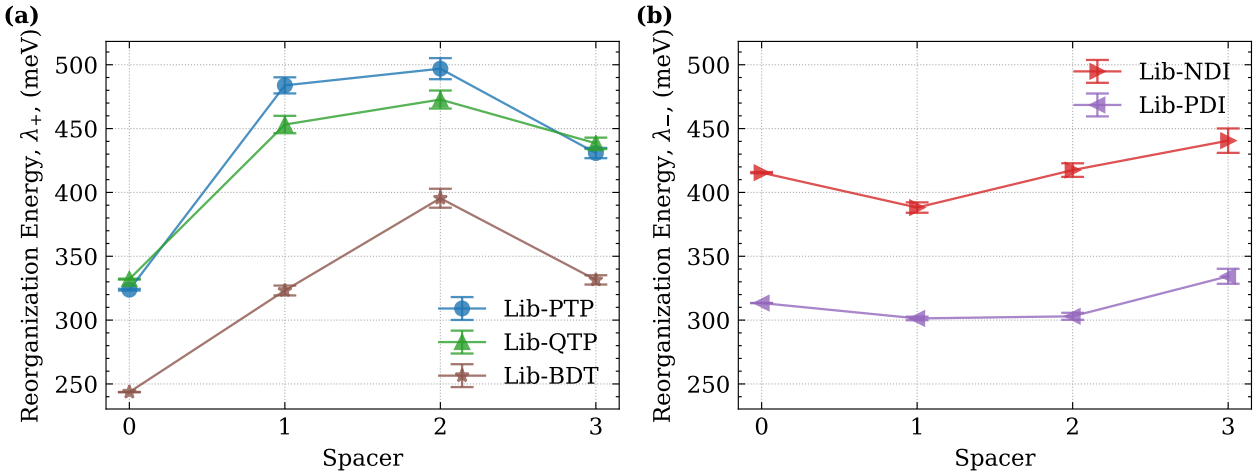


Figure 5.4: Reorganization energies of the chemically stable aromatic core libraries. Smaller reorganization energies are associated with higher charge mobility. (a) Reorganization energies of the stable cationic molecules associated with hole transport in p-type semiconductors. (b) Reorganization energies of the stable anionic molecules associated with electron transport in n-type semiconductors. Error bars represent the standard error in the mean.

ular spacing and/or a transverse offset of the monomers within the dimer geometry. We present in Fig. 5.5 the transfer integrals calculated for holes and electrons in all chemically stable species. In the case of hole transport, we observe the values of our calculated transfer integrals to lie within a range of 18-82 meV (Fig. 5.5a). Although the transfer integral heavily depends on the particular dimer geometries, previous literature has reported transfer integral values of ∼96.7 meV when the center of mass distance between pairs of DNTT is 4.374 Å [48]. This is the same order of magnitude that we observe for Lib-BDT spacer 0, which possesses transfer integral values of 82 meV at center of mass separations of ∼4 Å. In the case of electron transport, we find transfer integral values in the range 42 – 76 meV (Fig. 5.5b). In comparison, the transfer integral values for NDTI derivatives of 38-46 meV were reported at COM distances of 4.2 Å [377], which is the same order of magnitude as our Lib-PDI spacer 0 that possess transfer integral values of 76 meV at COM separations of ∼3.41 Å.

While the trends in the transfer integrals with core chemistry and spacer length are more subtle than in the reorganization energy calculations, we notice that in hole transport

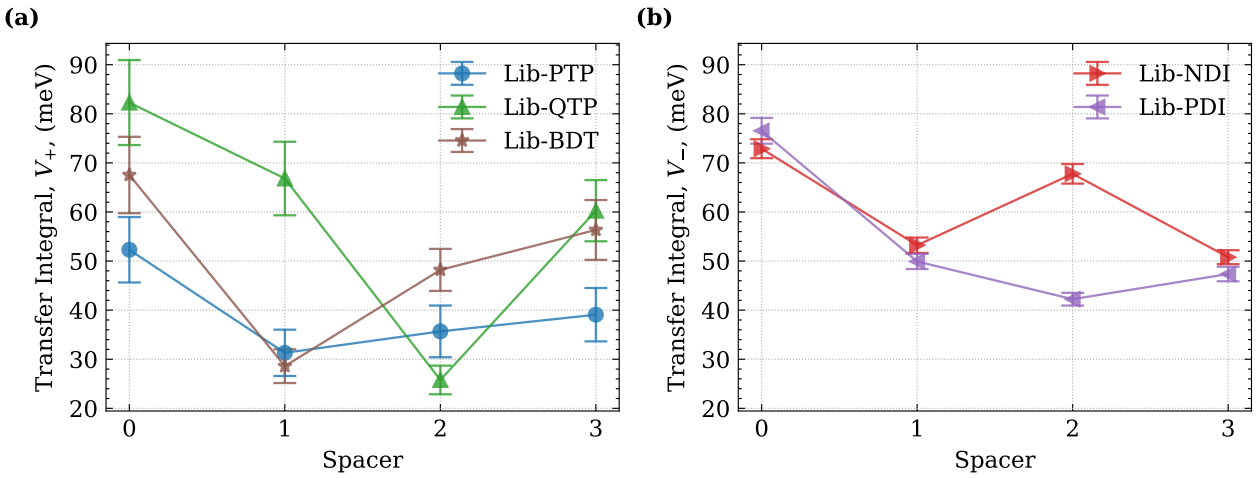


Figure 5.5: Transfer integrals for the chemically stable aromatic core libraries. Larger transfer integrals are associated with higher charge mobility. (a) Transfer integrals between stable cationic molecules associated with hole transport in p-type semiconductors. (b) Transfer integrals between stable anionic molecules associated with electron transport in n-type semiconductors. Error bars represent the standard error in the mean.

Lib-QTP typically performs better than Lib-PTP or Lib-BDT for smaller spacer lengths. For electron transport Lib-NDI and Lib-PDI perform similarly. Again, our trends are non-monotonic, but we observe relatively weak but negative correlations between transfer integral and spacer length: $\rho_{\text{Lib-PTP}} = -0.10$ (p-value=0.16), $\rho_{\text{Lib-QTP}} = -0.23$ (p-value=8.4×10^{-4}), $\rho_{\text{Lib-BDT}} = -0.04$ (p-value=0.6), $\rho_{\text{Lib-NDI}} = -0.05$ (p-value=0.48), $\rho_{\text{Lib-PDI}} = -0.24$ (p-value=5.8×10^{-4}). These trends indicate that increasing the spacer length tends to diminish the value of the transfer integral and signaling weaker intermolecular electronic coupling. Computing the Pearson correlation coefficient between the center-of-mass separation between pairs of molecules and the calculated transfer integral reveals uniformly negative, and in some cases quite strong, negative correlations: $\rho_{\text{Lib-PTP}} = -0.36$ (p-value=1.1×10^{-7}), $\rho_{\text{Lib-BDT}} = -0.31$ (p-value=7.9×10^{-6}), $\rho_{\text{Lib-PDI}} = -0.082$ (p-value=0.25), $\rho_{\text{Lib-NDI}} = -0.12$ (p-value=0.086), $\rho_{\text{Lib-QTP}} = -0.18$ (p-value=0.012) suggesting that higher transfer integral values correspond to π-conjugated peptide cores that are more tightly associated. These correlations also imply that, generally, smaller spacer lengths result in π-conjugated peptide nanoaggregates with closer intermolecular contacts.

5.4.4 Charge Mobility

Finally, we use Marcus theory (Eqns. 5.4, 5.5, and 5.6) to predict charge mobilities from our calculated reorganization energies and transfer integrals. We present in Fig. 5.6a the hole mobility and in Fig. 5.6b the electron mobility for all chemically stable species. We notice the hole mobility ranges between 0.003 and 0.224 $cm^2/(Vs)$, while the electron mobility ranges between 0.024 and 0.143 $cm^2/(Vs)$. In comparison, the hole mobility in DNTT and its derivatives reported in Ref. [48] range from 1.45 to 3.36 $cm^2/(Vs)$, which is at least ∼550% higher than our best reported mobility of 0.224 $cm^2/(Vs)$ in Lib-BDT spacer 0. On the other hand, the electron mobility of NDTI derivatives ranges from 0.24 to 0.39 $cm^2/(Vs)$ [377], which is only ∼71% higher than our highest electron mobility of 0.143 $cm^2/(Vs)$ in Lib-PDI spacer 0. We note that our predicted charge mobilities pertain to electron/hole transport along the π-cores in individual stacks of π-conjugated peptides and are commensurate with prior experimental reports of charge mobility within organic semiconductor nanowire structures [51, 378, 379]. We use these predictions as a consistent basis within which to compare the various π-conjugated peptides considered in this work, but appreciate that lower measured mobilities would be expected within films of these nanowires due to the presence of the insulating peptide side chains and possible attenuation effects due to the incorporation of these materials into a device [28].

Performing these charge mobility calculation serves as a comparative tool to help isolate the most promising π-conjugated peptides for optoelectronic applications. For hole mobility, we find that Lib-BDT with spacer 0 has the highest mobility at 0.224 $cm^2/(Vs)$ which is about 65% better than Lib-PTP spacer 0 and Lib-QTP spacer 0. This result is mainly due to the comparatively low reorganization energy for Lib-BDT which has an exponential dependence on the mobility, while the transfer integral only has a quadratic dependence based on Marcus theory (c.f., Eqn. 5.6). We also notice that Lib-QTP and Lib-PTP perform comparably to one another and similarly to Lib-BDT for spacer lengths of one and greater. For electron

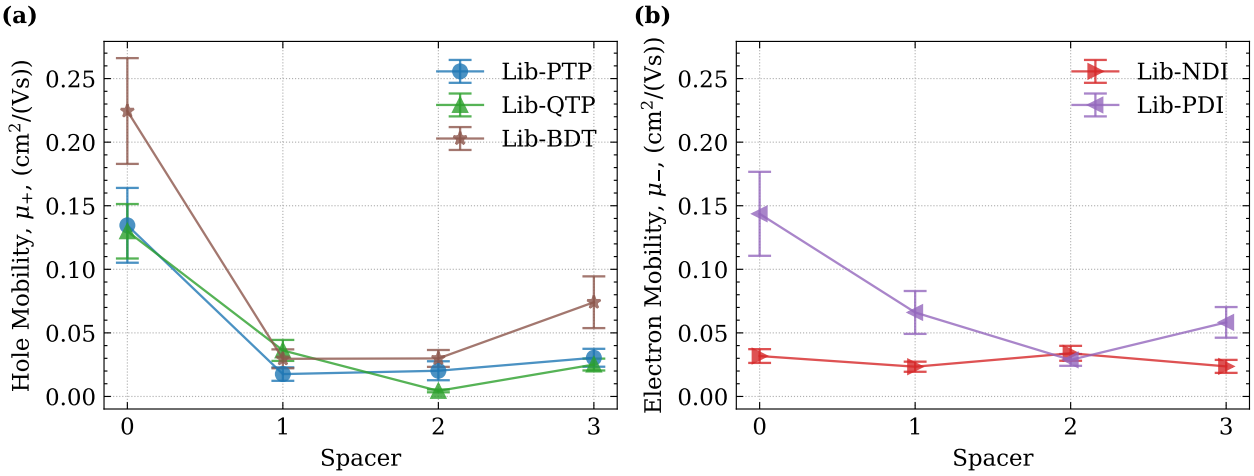


Figure 5.6: Charge mobility predicted from the calculated reorganization energies and transfer integrals using Marcus theory. (a) Predicted hole mobilities in stable cationic molecules forming putative p-type semiconductors. (b) Predicted electron mobilities in stable anionic molecules forming putative n-type semiconductors. Error bars represent the standard error in the mean.

mobility, Lib-PDI with spacer 0 is the highest performing with an electron mobility of 0.143 cm^2/(Vs) that is ∼320% better than Lib-NDI spacer 2, once again primarily due to the low calculated value of the reorganization energy.

In general, we notice charge mobility to peak at zero spacer length and degrade, within error, for increasing spacer length for all calculated libraries, except for Lib-NDI which we find performs similarly poorly at all spacer lengths. We can once again quantify this trend by observing overall negative, and sometimes relatively large, correlations between charge mobility and spacer length: $\rho_{\text{Lib-PTP}} = -0.29$ (p-value$=3.9 \times 10^{-5}$), $\rho_{\text{Lib-QTP}} = -0.41$ (p-value$=2.6 \times 10^{-9}$), $\rho_{\text{Lib-NDI}} = -0.043$ (p-value$=0.55$), $\rho_{\text{Lib-PDI}} = -0.23$ (p-value$=0.0012$), $\rho_{\text{Lib-BDT}} = -0.27$ (p-value$=8.4 \times 10^{-5}$). Based on these results, we suggest Lib-BDT with zero spacer and Lib-PDI with zero spacer as the promising candidate p- and n-type organic semiconductors, respectively. Furthermore, we identify reorganization energy as the primary determinant of charge mobility in our case and that superior search efficiencies may be achieved when performing larger computational screenings by pre-screening molecules based on reorganization energy prior to performing transfer integral calculations.

5.5 Conclusions

In this study, we combine molecular dynamics simulations with first-principles electronic-structure calculations to predict the chemical stability and charge transport properties of a 24-molecule library of π-conjugated peptides containing one of six π-conjugated cores and with alkyl chain spacers of length $n \in \{0, 1, 2, 3\}$ connecting the central aromatic cores to their flanking peptide wings (Fig. 5.1). For each library we perform electronic-structure calculations on a collection of dimer configurations harvested from molecular dynamics simulations of spontaneous self-assembly. Since these nanoaggregates lack a well-defined crystal structure, we are obliged to first perform molecular dynamics simulations of self-assembly to extract representative configurations to pass to the electronic structure calculations.

Chemical stability calculations serve as an initial filtration of our design space and are used to identify stable cationic and anionic candidates for putative hole and electron transport, respectively. Our stability calculations suggested that three of the candidate aromatic cores – phenyl-theophene-phenyl (Lib-PTP), quaterthiophene (Lib-QTP), and benzo-dithiophene (Lib-BDT) – have stable cationic states indicating that they are potential candidates for p-type semiconductors, whereas two of the candidate aromatic cores – naphthalene-diimide (Lib-NDI) and pyridine-diimide (Lib-PDI) – have stable anionic states indicating that they are potential candidates for n-type semiconductors. We then calculate the reorganization energy and transfer integral for these molecules as inputs to Marcus theory charge mobility calculations. Our results identify benzo-dithiophene (Lib-BDT) with no alkyl spacer as possessing the highest predicted hole mobility of 0.224 $cm^2/(Vs)$, and pyridine-diimide (Lib-PDI) with no alkyl spacer as possessing the highest predicted electron mobility of 0.143 $cm^2/(Vs)$. In general, we find that charge mobility typically degrades with increasing spacer length, suggesting the absence of alkyl spacers as a more promising chemical motif for future study.

By combining molecular dynamics simulation and first-principles electronic-structure

calculations we provide a detailed study of the chemical stability and charge mobility of π-conjugated peptide semiconductors. Although the calculated charge mobility values are lower than typical values that may be expected for crystalline semiconductors, π-conjugated peptide-based semiconductors possess advantages in terms of mechanical flexibility, solution processability, weight, aqueous stability, and biocompatibility, making them well-suited to charge transport and energy harvesting applications in biological and living systems. Our computational methodology is also expected to be generically extensible to screen other self-assembling and electronically active organic molecules.

Data Availability

Input files for the molecular dynamics simulations and electronic structure calculations along with MD trajectories, tabulated results of chemical stability, reorganization energy, transfer integral and charge mobility calculations, and post-processing scripts for the transfer integral calculation are provided via The Materials Data Facility (MDF) [240, 241] at DOI:10.18126/ef9q-qzod [380].

Supporting Information

Electronic supplementary information (ESI) available. See `https://doi.org/10.1039/D2ME00017B`

CHAPTER 6

TEMPORALLY COHERENT BACKMAPPING OF MOLECULAR TRAJECTORIES FROM COARSE-GRAINED TO ATOMISTIC RESOLUTION

6.1 Abstract

Coarse-graining offers a means to extend the achievable time and length scales of molecular dynamics simulations beyond what is practically possible in the atomistic regime. Sampling molecular configurations of interest can be done efficiently using coarse-grained simulations, from which meaningful physicochemical information can be inferred if the corresponding all-atom configurations are reconstructed. However, this procedure of backmapping to reintroduce the lost atomistic detail into coarse-grain structures has proven a challenging task due to the many feasible atomistic configurations that can be associated with one coarse-grain structure. Existing backmapping methods are strictly frame-based, relying on either heuristics to replace coarse-grain particles with atomic fragments and subsequent relaxation, or parameterized models to propose atomic coordinates separately and independently for each coarse-grain structure. These approaches neglect information from previous trajectory frames that is critical to ensuring temporal coherence of the backmapped trajectory, while also offering information potentially helpful to produce higher-fidelity atomic reconstructions. In this work we present a deep learning-enabled data-driven approach for temporally coherent backmapping that explicitly incorporates information from preceding trajectory structures. Our method trains a conditional variational autoencoder to non-deterministically reconstruct atomistic detail conditioned on both the target coarse-grain configuration and the previously reconstructed atomistic configuration. We demonstrate our backmapping approach on two exemplar biomolecular systems: alanine dipeptide and the miniprotein chignolin. We show that our backmapped trajectories accurately recover the structural, thermodynamic, and

kinetic properties of the atomistic trajectory data.

This chapter is based on the work reported in: **Shmilovich, K.**, Stieffenhofer, M., Charron, N. E., & Hoffmann, M. (2022). "Temporally Coherent Backmapping of Molecular Trajectories From Coarse-Grained to Atomistic Resolution." *The Journal of Physical Chemistry A*, 126(48), 9124-9139.

6.2 Introduction

A central limitation of modelling soft-matter systems with molecular dynamics (MD) simulations are the long characteristic timescales of interesting processes, such as protein folding, compared to the relatively short integration time steps required to accurately propagate the system forward in time. A plethora of strategies strive to overcome this time scale barrier, such as enhanced sampling techniques, [381, 382, 383, 101, 384] modern/specialized hardware, [385, 88, 89] and hierarchical multiscale modelling [113, 114, 115, 116, 117, 118]. Coarse-grained (CG) simulations are one such multiscale approach that enables access to spatiotemporal scales entirely out of reach of conventional atomistic molecular dynamics simulations. The process of coarse-graining typically aggregates groups of atoms into 'beads' or 'superatoms' intended to preserve important properties of the original atomistic system [119, 120, 121]. As such, CG simulations require monitoring fewer particles which allows for the study of larger and more complex systems typically untenable in the atomistic regime at comparable computational cost. A typical consequence of this reduction in resolution is an effective 'smoothing' of the underlying free energy surface, which helps expedite large-scale and slowly evolving conformational motions that might otherwise be frustrated or kinetically trapped in the more rugged atomistic landscape [122, 123, 116, 124, 125] These advantages have led to the growing popularization and use of CG models particularly for simulations of proteins, polymers, molecular self-assembly, membranes, and high-throughput screening [126, 127, 128, 129, 130, 131, 132, 133, 134, 135].

The primary concession of coarse-graining is the sacrifice of fine-grained, atomistic detail. Restoring this lost detail by converting a CG representation into a corresponding atomistic representation is commonly dubbed "backmapping" and is important for analyses requiring atomistic resolution, for example electronic structure calculations for determining NMR spectra or dipole moments [386, 387]. Traditional backmapping strategies rely on geometric heuristics to replace beads with their associated atomic fragments. These approaches typically produce quite poor initial structures that must be subsequently subjected to refinement using energy minimization and/or (restrained) molecular dynamics to equilibrate each backmapped frame [115, 136, 137, 138, 139]. However, significant computational cost is incurred with the required frame-by-frame intervention in these approaches which hinders the applicability of backmapping larger systems and/or longer trajectories. More recently, data-driven backmapping techniques have been proposed which deploy machine learning (ML) models that learn to reconstruct atomistic details from training examples [140, 141, 144, 143, 145, 142]. These approaches offer more scalability with higher throughput as they are typically trained to produce well-equilibrated structures that do not require frame-by-frame energy minimization or relaxation. Coarse-graining is an inherently many-to-one operation, with multiple atomic structures corresponding to each CG representation. A favorable feature of any backmapping procedure is the capacity to recapitulate the conformational diversity of atomic structures corresponding to a particular CG representation. A subset of these data-driven methods [140, 141, 142, 143] that possess this conformational expressibility are therefore capable of non-deterministic backmapping, where a variety of feasible and novel atomistic structures can be generated when backmapping any individual CG configuration.

A commonality between all existing approaches is backmapping each frame individually and separately from one-another. However, leveraging temporal information can improve reconstruction quality and enable the recovery of dynamic properties. In particular, some

important dynamic properties rely on time correlations of local atomistic details. For example, the calculation of diffusion constants is related to the integral of the velocity auto-correlation [388], infrared absorption spectra is related to the auto-correlation function of the total dipole moment [389, 390], and scattering functions that are related to Fourier transforms of the van Hove correlation function [391, 392, 393, 394]. Existing backmapping schemes are not temporally aware and correlations between consecutive frames are only maintained via large-scale characteristics. As a consequence, the reintroduced degrees of freedom between consecutive frames might decorrelate locally, and time correlations based on local, atomistic descriptors are typically not reliable for such backmapped trajectories. Therefore, presently absent from this suite of backmapping methods is a data-driven approach for generatively backmapping CG trajectories that also incorporates temporal information.

We present in this work a new method to perform temporally coherent backmapping of molecular simulation trajectories via a deep learning-based solution for all-atom reconstruction of CG simulation trajectories that aims at both generating well-equilibrated molecular structures for each frame and achieves temporal coherence between frames. To this end, we explicitly incorporate configurational information from previous simulation frames when generating reconstructions. This task is accomplished by training a conditional variational autoencoder (cVAE) that learns to up-scale CG configurations into full atomistic resolution by conditioning on the current coarse- and previous fine-grained structures. Our cVAE learns to model a variety of feasible atomistic structures associated with each CG configuration, which allows us to generatively produce novel backmapped trajectories that are not a simple carbon copy of the training data. We show for two exemplar biomolecular systems alanine dipeptide (ADP) and the miniprotein chignolin (CLN) that our approach generates atomic reconstructions which recover atomistic structural, thermodynamic, and kinetic properties. Furthermore, we show our backmapping model performs well on held out in distribution data and generalizes to CG data originating from unseen and approximate CG force fields.

6.3 Methods

Our approach utilizes a reference atomistic trajectory (or set of atomistic trajectories) for a molecule we intend to backmap. The trajectory contains N atoms and is composed of a collection of T frames $\mathbf{AA} = \{\mathrm{AA}^0, \mathrm{AA}^1, \ldots, \mathrm{AA}^{T-1}\}$, where $\mathrm{AA}^t \in \mathbb{R}^{N\times 3}$ are the atomistic coordinates for each frame. We assume there exists a coarse-graining function f_{cg} that maps all-atom coordinates to CG coordinates, such that $f_{cg}(\mathrm{AA}^t) = \mathrm{CG}^t \in \mathbb{R}^{n\times 3}$, where n is the number CG beads such that $n < N$. This function is then applied frame-by-frame to the atomistic frames $\mathbf{AA}$ yielding the corresponding CG trajectory $\mathbf{CG} = \{\mathrm{CG}^0, \mathrm{CG}^1, \ldots, \mathrm{CG}^{T-1}\}$. This pair of atomistic and CG trajectories $(\mathbf{AA}, \mathbf{CG})$ compose the data used to train our super-resolution model (Fig. 6.1a). From this data our model attempts to learn the conditional distribution $P(\mathrm{AA}^t|\mathrm{CG}^t, \mathrm{AA}^{t-1})$ implicitly, such that we can reconstruct an atomisitic configuration given the CG structure CG^t and a previous atomistic structure AA^{t-1}. Each training sample therefore consists of a sequential pair of atomistic configurations and the current CG configuration $(\mathrm{AA}^t, \mathrm{CG}^t, \mathrm{AA}^{t-1})$. While increasing the number of previous frames to incorporate $\mathrm{AA}^{t-2}, \mathrm{AA}^{t-3}, \ldots$ is straight-forward, our experiments show that this simpler Markovian posture already yields accurate temporally coherent reconstructions that reproduce structural, thermodynamic and kinetic properties with remarkable accuracy. A complete PyTorch [395] implementation of our model with all associated analyses is publically available at DOI:10.18126/tf0h-w0jz [396].

6.3.1 Representing molecules as spatially voxelized particle densities

Learning complex and high-order dependencies is a hallmark of computer vision. One of the most successful generative models are convolutional neural networks (CNNs), which have led to groundbreaking successes in image processing [147, 148, 149, 150, 151]. In order to take advantage of CNNs for our backmapping task we choose to represent our data as a set of 3D featurized images. To this end, a smooth density representation discretized on a 3D grid is

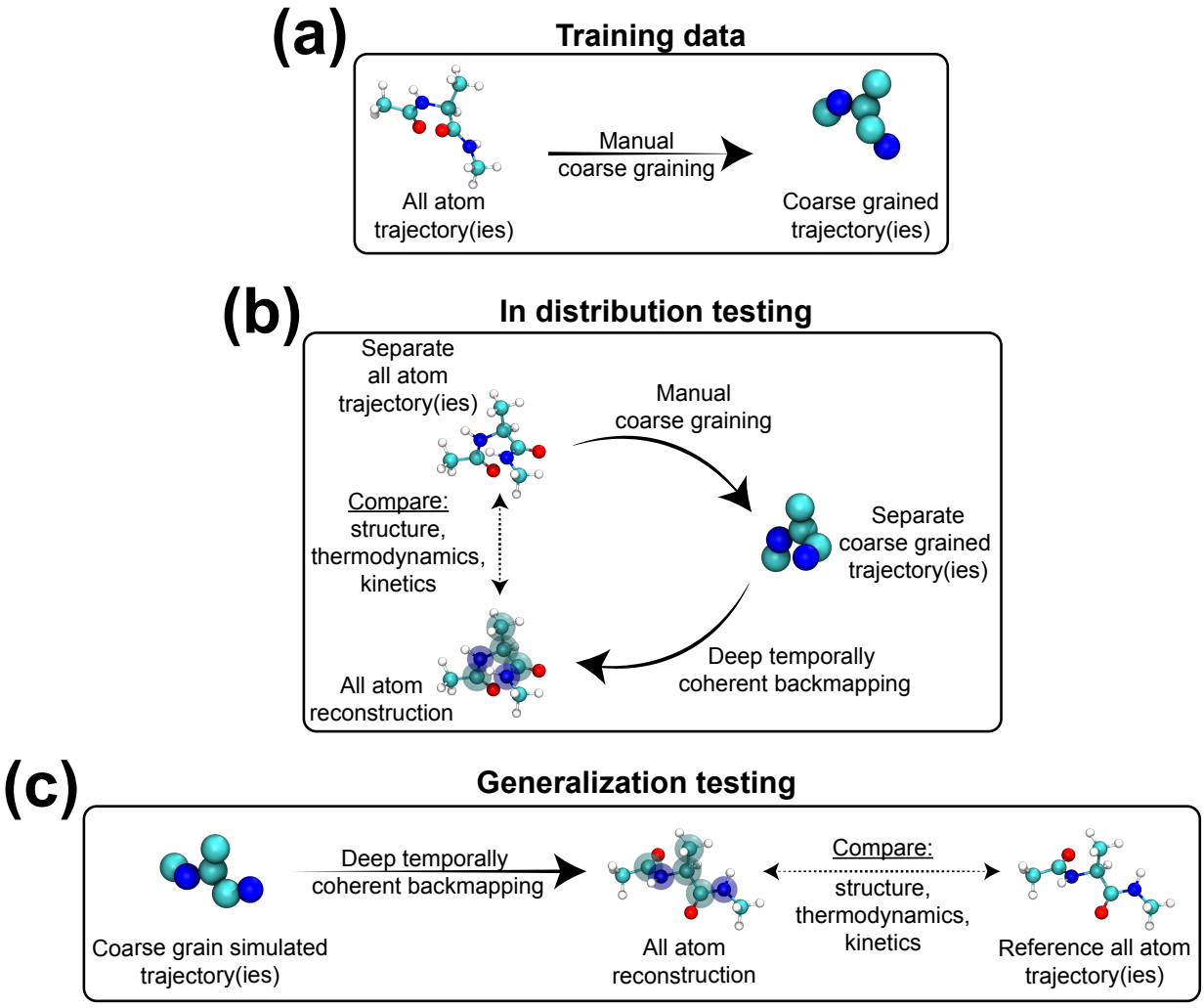


Figure 6.1: Illustration of training and testing setups.

used to encode the positions of atoms and beads. Each particle is placed in a separate tensor channel to avoid overlap of densities that would be difficult to disentangle. More concretely, an atomistic configuration $\mathrm{AA}^t \in \mathbb{R}^{N\times 3}$ is represented as a 4D tensor $\mathrm{VOX}^{\mathrm{AA}^t} \in \mathbb{R}^{d\times d\times d\times N}$, where the first three dimensions discretize the location of each particle in space and the final dimension represents a channel for each particle:

$$\mathrm{VOX}_i^{\mathrm{AA}_t}(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x}-\mathrm{AA}_i^t)^2}{2\sigma^2}\right). \tag{6.1}$$

Here, i is the particle index and $\mathbf{x} \in \mathbb{R}^3$ is the Cartesian location of each grid point taken from a regular Cartesian grid of width r_{grid} such that the maximal and minimal values of each spatial dimension are $-\frac{r_{grid}}{2}$ and $\frac{r_{grid}}{2}$. The density of each particle is therefore a 3D Gaussian centered about the position $\mathrm{AA}_i^t \in \mathbb{R}^3$ with the width σ. It is critical to ensure r_{grid} is large enough such that each particle density within the trajectory is fully enclosed by the voxelized grid. The parameter d then controls the resolution of our spatial discretization, such that a larger d leads to more spatial resolution at the cost of higher memory and processing requirements. Furthermore, σ, which controls the effective size of each particle density, will also impact the mass assigned to each voxel.

From a density profile $\mathrm{VOX}_i^{\mathrm{AA}^t}$ we can also recover a set of particle coordinates by performing a weighted average over the voxelized grid,

$$\tilde{\mathrm{AA}}_i^t = \frac{1}{z} \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{l=1}^{d} x_{j,k,l} \mathrm{VOX}_i^{\mathrm{AA}^t}(x_{j,k,l}) \tag{6.2}$$

where $z = \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{l=1}^{d} \mathrm{VOX}_i^{\mathrm{AA}^t}(x_{j,k,l})$ is a normalization constant, $x_{j,k,l}$ is a particular coordinate value within the three-dimensional spatially voxelized grid and the summations are carried over the three spatial dimensions. Given a predicted density profile $\mathrm{VOX}^{\hat{\mathrm{AA}}^t}$ we can perform this weighted average for each i particle channel $\mathrm{VOX}_i^{\hat{\mathrm{AA}}^t}$ to recover the complete set of atomistic coordinates $\hat{\mathrm{AA}}^t$. We note that transforming these densities into and from Cartesian coordinates constitutes a sequence of differentiable operations and therefore enables us to readily incorporate this particle averaging and voxelization within the computational graph of our model. CG configurations CG^t are treated in the same way as their atomistic counter-parts, yielding corresponding voxelized representations $\mathrm{VOX}^{\mathrm{CG}^t} \in \mathbb{R}^{d \times d \times d \times n}$ where n is the number of beads associated with each CG configuration.

6.3.2 Conditional Variational Autoencoder

To learn a temporally coherent probabilistic mapping from CG to atomistic configurations we train a conditional variational autoencoder (cVAE) [146]. A typical VAE [269] consists of an encoder that compresses high-dimensional inputs into a lower dimensional latent space that captures salient information characterizing the input data. This latent, compressed representation is then inputted to a decoder that aims to reconstruct the original high-dimensional input. The cVAE operates under a similar premise but the decoder is also provided with some partial information about the input along with the latent code when producing reconstructions. Fig. 6.2a shows a schematic illustration of our cVAE architecture. In our application, the cVAE encoder takes as input the triplet of configurations $(\mathrm{AA}^t, \mathrm{CG}^t, \mathrm{AA}^{t-1})$, where the subset $(\mathrm{CG}^t, \mathrm{AA}^{t-1})$ is interpreted as the conditional variable and AA^t the intended reconstruction target. The decoder is a function of both the latent code and the conditional variable $(\mathrm{CG}^t, \mathrm{AA}^{t-1})$ and learns to reconstruct atomistic configurations $\hat{\mathrm{AA}}^t$ to closely match the data AA^t. For a fixed condition $(\mathrm{CG}^t, \mathrm{AA}^{t-1})$ we can learn to meaningfully encode information about the target configuration AA^t into a low dimensional latent code, such that the decoder when presented with $(\mathrm{CG}^t, \mathrm{AA}^{t-1})$ yields slightly different, yet valid, reconstructions $\hat{\mathrm{AA}}^t$ for different instantiations of the latent code. This data driven approach enables our model to generatively produce reconstructions AA^t in a temporally aware way by conditioning on both the coarse grained configuration and the previous atomistic configuration $(\mathrm{CG}^t, \mathrm{AA}^{t-1})$ (Fig. 6.2b).

The Encoder

The purpose of the encoder in our cVAE is to distill information from the input configurations $(\mathrm{AA}^t, \mathrm{CG}^t, \mathrm{AA}^{t-1})$ into a low-dimensional latent space vector that is used in conjunction with the conditional variables $(\mathrm{CG}^t, \mathrm{AA}^{t-1})$ for the decoder to generate an atomistic reconstruction $\hat{\mathrm{AA}}^t$. These configurations are passed to the encoder as voxelized grids of par-

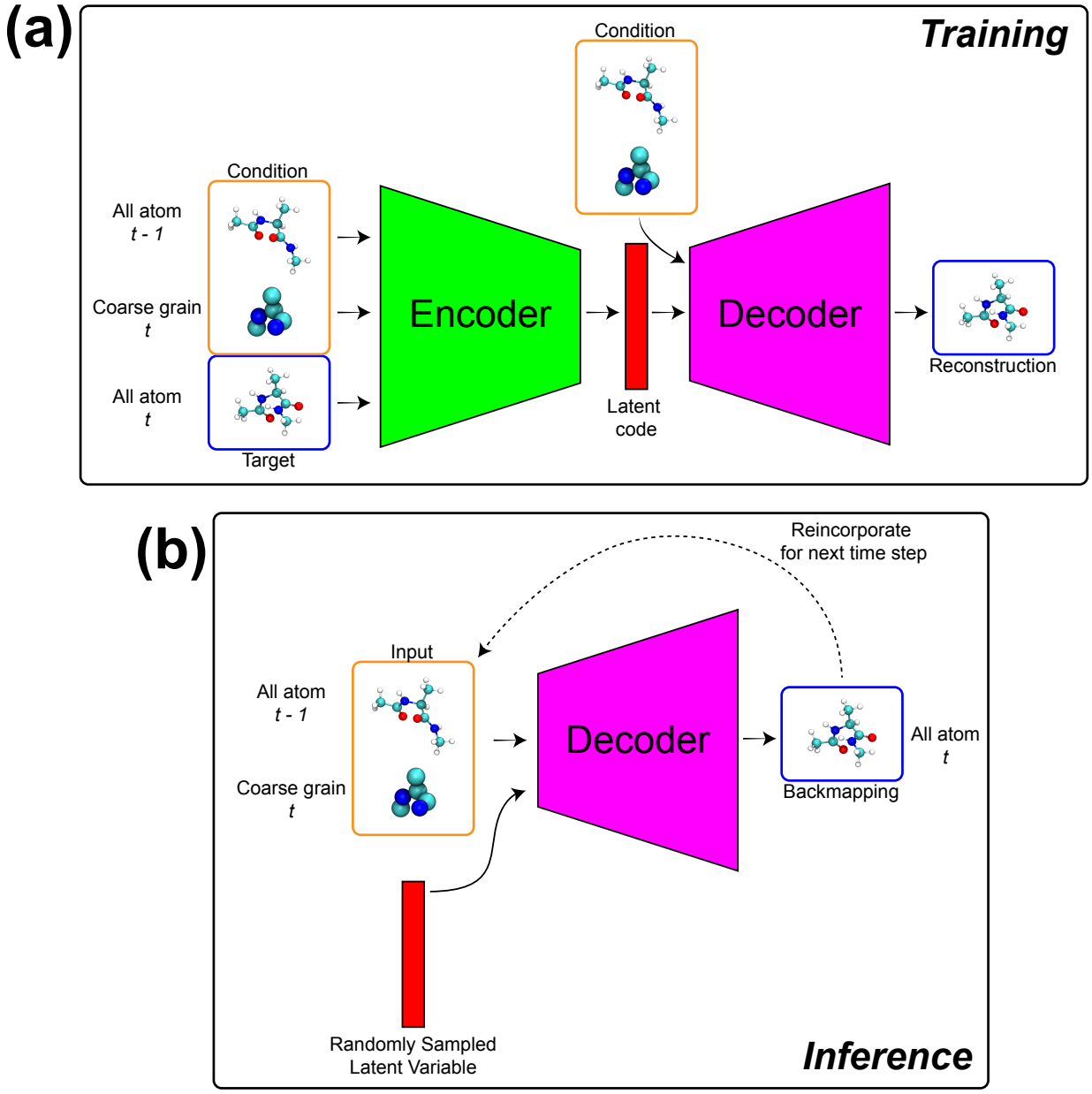


Figure 6.2: Schematic illustration of our model training and inference setups.

ticle densities ($\mathrm{VOX}^{\mathrm{AA}^t}, \mathrm{VOX}^{\mathrm{CG}^t}, \mathrm{VOX}^{\mathrm{AA}^{t-1}}$) which are constructed as a single monolithic tensor concatenated along the particle dimension $\mathbf{x} = [\mathrm{VOX}^{\mathrm{AA}^t} || \mathrm{VOX}^{\mathrm{CG}^t} || \mathrm{VOX}^{\mathrm{AA}^{t-1}}] \in \mathbb{R}^{d \times d \times d \times (2N+n)}$, where N and n are the number of atoms/beads in each atomistic and CG configuration, respectively. The encoder is composed of 3D residual CNNs with a terminal dense layer to extract a fixed dimensional latent vector. The locality of the CNN kernel provides a strong inductive bias by focusing on proximate spatial profiles from multiple high and low resolution scales and multiple time steps. Processing the input $\mathbf{x}$ using many consecutive CNN modules enables us to hierarchically incorporate progressively more distant features, and ultimately yielding a multiscale, spatially and temporally aware latent representation $\mathbf{z} \in d_{\text{latent}}$ for that slice of the trajectory. This latent representation $\mathbf{z}$ is then passed to the decoder in conjunction with the conditional variable $\mathbf{c} = [\mathrm{VOX}^{\mathrm{CG}^t} || \mathrm{VOX}^{\mathrm{AA}^{t-1}}]$ to predict the all atom configuration for the subsequent time step $\mathrm{Decoder}(\mathbf{c}, \mathbf{z}) = \hat{\mathrm{AA}}^t$.

The Decoder

The purpose of the decoder is to use information provided in the conditional variable $\mathbf{c} = [\mathrm{VOX}^{\mathrm{CG}^t} || \mathrm{VOX}^{\mathrm{AA}^{t-1}}]$ jointly with the latent code $\mathbf{z}$ produced by the encoder to reconstruct of the current atomistic configuration AA^t. Upon completing training, we can eschew the encoder and use the decoder as our generative model for backmapping. Using a randomly sampled latent variable $\mathbf{z}$ as a source of noise while specifying the conditional variable as the current CG configuration CG^t and the previous atomistic configuration AA^{t-1} we can backmap the atomistic reconstruction $\hat{\mathrm{AA}}^t$. Similar to the encoder, our decoder is primarily composed of residual 3D CNNs. Unlike the encoder, the neural network backbone of the decoder needs to only output a 4D tensor of equivalent dimensionality to the voxelized atomistic representation $\mathrm{VOX}^{\mathrm{AA}^t} \in \mathbb{R}^{d \times d \times d \times N}$. As the voxelized representation $\mathrm{VOX}^{\mathrm{AA}^t}$ contains separate channels for each atom, we can simply perform an average over of the spatial density profiles within each particle channel to independently localize each atom

coordinate (c.f. Sec. 6.3.1). We control the behavior of our model by defining a training loss that captures relevant aspects we ultimately want reflected in our reconstructions.

6.3.3 Training routines

Our model is trained end-to-end using the ADAM optimizer [217]. For a single sample, the complete loss $\mathcal{L}$ is given by

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}_{\text{VOX}} + \mathcal{L}_{\text{AA}} + \mathcal{L}_{\text{CG}} + \mathcal{L}_{\text{EDM}} + \lambda\mathcal{L}_{\text{ENERGY}} + \beta\mathcal{L}_{\text{KLD}} \qquad (6.3)\\
\mathcal{L}_{\text{VOX}} &= \frac{1}{d^3 N}\,||\text{VOX}^{\text{AA}^t} - \text{VOX}^{\hat{\text{AA}}^t}||_F^2\\
\mathcal{L}_{\text{AA}} &= \frac{1}{3N}\,||\text{AA}^t - \hat{\text{AA}}^t||_F^2\\
\mathcal{L}_{\text{CG}} &= \frac{1}{3n}\,||\text{CG}^t - \hat{\text{CG}}^t||_F^2\\
\mathcal{L}_{\text{EDM}} &= \frac{1}{2N^2}\,||EDM^{\text{AA}^t} - EDM^{\hat{\text{AA}}^t}||_F^2\\
\mathcal{L}_{\text{ENERGY}} &= (U^{\text{AA}^t} - U^{\hat{\text{AA}}^t})^2\\
\mathcal{L}_{\text{KLD}} &= \mathcal{D}_{KL}(\mathbf{z}||\mathcal{N}(0,\mathbf{I})).
\end{aligned}$$

The first five terms in Eqn. 8.1 are components of an effective reconstruction loss and $\mathcal{L}_{\text{KLD}}$ is the standard Kullback–Leibler divergence loss between the latent code $\mathbf{z}$ and a normal distribution typically used in VAE training [269]. While we ultimately strive to recover all atom coordinates AA^t, the model primarily operates on spatially voxelized particle densities representations VOX^{AA^t}. A critical component of learning will therefore involve reconstructing these density profiles to enable accurate and sharp atomic coordinate generation. The $\mathcal{L}_{\text{VOX}}$ term is a mean squared error (MSE) between the target VOX^{AA^t} and reconstructed $\text{VOX}^{\hat{\text{AA}}^t}$ atomistic voxels, while $\mathcal{L}_{\text{AA}}$ is an MSE between the ground truth AA^t and reconstructed $\hat{\text{AA}}^t$ atomic coordinates. $\mathcal{L}_{\text{VOX}}$ helps the model learn to reproduce the atomic densities of the intermediate voxelized representations, while $\mathcal{L}_{\text{AA}}$ helps ensure sharper co-

ordinate reconstruction when the voxels $\mathrm{VOX}^{\mathrm{AA}^t}$ are ultimately collapsed back into atomic coordinates AA^t. We also use the coarse graining function f_{cg} to determine the CG representation of an atomistic reconstruction $\hat{\mathrm{CG}}^t = f_{cg}(\hat{\mathrm{AA}}^t)$, which is used in $\mathcal{L}_{\mathrm{CG}}$ to calculate an MSE with respect to the input CG coordinates CG^t. The motivation to include $\mathcal{L}_{\mathrm{CG}}$ is that a coarse graining of the atomistic backmapping $\hat{\mathrm{AA}}^t$ should ultimately match the original CG structure CG^t from which the reconstruction is derived. In the $\mathcal{L}_{\mathrm{EDM}}$ term we calculate an MSE between the $N \times N$-dimensional Euclidean Distance Matrix (EDM) of the target EDM^{AA^t} and the reconstructed $EDM^{\hat{\mathrm{AA}}^t}$ atomic coordinates to help our backmapping better preserve bond lengths and other interatomic distances. Lastly, in the $\mathcal{L}_{\mathrm{ENERGY}}$ term we calculate an MSE between the scalar valued total potential energy of the target U^{AA^t} and reconstruction $U^{\hat{\mathrm{AA}}^t}$. $\mathcal{L}_{\mathrm{ENERGY}}$ serves as a regularizer to improve the quality of backmapped structures by penalizing reconstructed configurations that may have suitable geometric contributions but are otherwise energetically unfavorable. As such, it accelerates convergence and helps more precisely match the reconstructed energetics to the ground truth trajectory. During training, before configurations are voxelized, each training configuration is mean centered and we learn covariance with respect to rigid rotations by applying a random Euler rotation $\mathbf{R} \in \mathbb{R}^{3\times3}$ separately augmenting samples in each forward pass $(\mathrm{AA}^t, \mathrm{CG}^t, \mathrm{AA}^{t-1}) \rightarrow (\mathrm{AA}^t\mathbf{R}, \mathrm{CG}^t\mathbf{R}, \mathrm{AA}^{t-1}\mathbf{R})$.

A challenge of incorporating $\mathcal{L}_{\mathrm{ENERGY}}$ within the loss function is that the potential energy function U is sensitive to small perturbations of the atomic coordinates, which is most severe for the bonded and non-bonded Lennard-Jones interaction. As such, it can become dominatingly large during the early stages of training before the model learns to stably localize atomic coordinates. To alleviate this issue, we incorporate the prefactor λ in Eqn. 8.1 which we set to $\lambda = 0$ for a fixed number of initial training steps after which point λ is slowly annealed up to $\lambda = 1$ using an exponential annealing schedule. We also include the β prefactor alongside $\mathcal{L}_{\mathrm{KLD}}$ for more flexibility in balancing the impact of the

KL regularization against the reconstruction losses. For the ADP model we employ a cyclic annealing schedule for β to mitigate KL vanishing [397], while for the CLN model we simply maintain $\beta = 1$ throughout training. Complete training settings and hyperparameter details are presented in the Supporting Information.

6.3.4 Inference

At inference time, we omit the encoder and use the decoder as the primary tool for generatively backmapping a CG trajectory. The decoder thereby reconstructs the atomistic frames in an autoregressive manner, i.e., the previous reconstructed atomistic frame $\hat{\text{AA}}^{t-1}$ serves as the input for the reconstruction of the next frame $\hat{\text{AA}}^{t}$. More specifically, the input for the decoder consists of a fixed dimensional latent vector $\mathbf{z}$ and the conditional variable $\mathbf{c} = [\text{VOX}^{\text{CG}^t}||\text{VOX}^{\text{AA}^{t-1}}]$ (Fig. 6.2b). However, for the first trajectory frame $t = 0$ there is no preceding frame for us to ascertain AA^{-1}. In this case, an atomistic configuration from the training dataset is chosen as this initial seed configuration AA^{-1}. We select this configuration by first determining the trajectory frame t^* in the training dataset AA_{train} that minimizes the RMSD with respect to the first test set CG frame $t^* = \min_i \text{RMSD}(f_{cg}(\text{AA}^i_{\text{train}}), \text{CG}^0)$, and then simply set the initial seed configuration as the immediately preceding training set trajectory frame $\text{AA}^{-1} = \text{AA}^{t^*-1}_{\text{train}}$ and lastly apply a rigid rotation to align AA^{-1} with CG^0.

At inference time the decoder uses a fixed dimensional latent vector $\mathbf{z}$ to provide a source of variance when generating atomic reconstructions. For a fixed condition $\mathbf{c}$ our model learns to effectively produce valid yet slightly different atomic reconstructions $\hat{\text{AA}}^{t} = \text{Decoder}(\mathbf{c}, \mathbf{z})$ for different instantiations of the latent code $\mathbf{z}$. Typically, inference is performed for VAEs or cVAEs by simply sampling $\mathbf{z}$ for each inference pass from the prior distribution over $\mathbf{z}$ which is normally taken as an isotropic Gaussian $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ [269]. At training time, however, the decoder is exposed to latent codes produced by the encoder from only the training set $\mathbf{Z} = \{\mathbf{z} = \text{Encoder}(\mathbf{x}) \ : \ \mathbf{x} \in \mathbf{X}_{\text{train}}\}$. Therefore, any mismatch between

the true aggregated posterior $\mathbf{Z}$ and the assumed prior $p(\mathbf{z})$ can lead to poor generative performance by the decoder when samples are selected from $p(\mathbf{z})$ because this may lead to operating within regions of latent space previously unseen by the decoder during training. We remedy this issue, following Ref. [254], by performing *ex-post* density estimation fitting a 10-component Gaussian Mixture Model (GMM) over the true posterior $\mathbf{Z}$. We then at inference time randomly sample $\mathbf{z}$ from our fit GMM, instead of from the assumed prior $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This process ensures our decoder operates within densely sampled latent space regions, which leads to higher fidelity, and less error-prone, atomic reconstructions.

After completing training for both our ADP and CLN models, we can visualize our latent space posterior $\mathbf{Z}$ by projecting our full-dimensional latent space into the first two Principal Components (PCs) of $\mathbf{Z}$ (Fig. S5 in the Supporting Information). Color-coding latent codes by the potential energy of the corresponding target configuration U_{target} in each sample reveals a strong correlation between the leading PC and internal energy for both of our test systems ADP ($\rho_{\text{pearson}}(PC_0^{ADP}, U_{target}) \approx 0.84$) and CLN ($\rho_{\text{pearson}}(PC_0^{CLN}, U_{target}) \approx 0.81$). These strong correlation suggests our encoder effectively extracts features from the input data to encode physically meaningful features, reflected in the internal energy, of the target atomistic configuration within our latent space embedding. We can also interrogate the generative capabilities of our model by using our decoder to produce reconstructions of a fixed CG input under many different latent code instantiations (Fig. S6 in the Supporting Information). The diversity of valid generated structures which still adhere to the CG input suggest that the decoder is attentive to subtle variations in the latent code when reconstructing atomistic configurations. In all, these properties of our cVAE model and latent space embedding validate expected behavior: our encoder effectively distills information to the latent code from the input data, while our decoder purposes the latent code together with the condition to non-deterministically and generatively reconstruct atomistic structures.

6.3.5 Data curation

Our model learns to backmap a CG trajectory of a molecule by training on a reference atomistic trajectory of that molecule that we coarse-grain after the fact to yield exemplar pairs of atomistic and coarse grained frames (Fig. 6.1a). When backmapping to predict an atomistic structure AA^t we consider both the current CG structure CG^t along with the previous atomistic structure AA^{t-1}. An important consideration when obtaining training data is the temporal spacing between consecutive frames $\mathrm{AA}^{t-1}, \mathrm{AA}^t$, as it may be impossible to accurately recover molecular motions that occur faster than this time. We also find that it is critical to ensure the reference atomistic trajectory is sufficiently sampled and captures the relevant atomistic state transitions that are expected to be reflected in the backmapped atomistic trajectory. These aspects of appropriate conformational sampling are important considerations in our approach, and more broadly in data-driven backmapping schemes, to provide confidence when backmapping data originating from coarse-grained models as they are typically designed to improve sampling of rarely occurring atomic conformational states. For example, a data-driven backmapping model strictly trained on simulation data from a protein simulation in the folded ensemble is unlikely to reliably and effectively backmap coarse-grained configurations in the unfolded ensemble. While ensuring comprehensive coverage of the accessible conformation space in molecular simulations remains an open challenge, a continually growing library of enhanced and adaptive sampling techniques [398, 399, 400, 401, 402, 403, 404, 405], along with other classical techniques such as simulated annealing, can serve to provide landmark atomistic configurations to initialize unbiased simulations in the collection of a diverse and through training dataset.

We train separate models on reference atomistic trajectories of alanine dipeptide (ADP) and the mini protein chignolin (CLN). Separate held-out trajectories for ADP and CLN are used as test sets for evaluating the performance of our model on in distribution data (Fig. 6.1b). As a more challenging generalization test of our method, we also backmap CG

trajectories generated from a bespoke CG force-field, CGSchNet (Fig. 6.1c). We then evaluate structural, thermodynamic and kinetic statistics of our atomistic reconstructions against reference atomistic trajectories to evaluate performance of our method. In the following sections, we describe the details of the simulations methods used to generate the ADP and CLN trajectories we use to train and test our model.

Alanine dipeptide

Atomistic data

Atomistic trajectories of alanine dipeptide (ADP) used for training are collected by performing molecular dynamics simulations in explicit solvent using OpenMM [406]. We closely mimic the simulation procedures outlined Ref. [407]. Langevin dynamics simulations are performed with a 2 fs time-step in the NVE ensemble using the AMBER ff-99SB-ILDN force field [259] within a cubic box containing 651 TIP3P water molecules randomly placed within a volume of $(2.7273\text{ nm})^3$. Electrostatics are treated using the particle-mesh Ewald (PME) method [408] using a 1.0 nm cutoff for the direct space interactions. The length of all bonds involving hydrogen atoms are constrained. Steepest descent energy minimization is used to relax the initial system configuration to within an energy of 10 kJ/mol. We then assign initial velocities to the energy minimized configuration by sampling from a Maxwell-Boltzmann distribution at 300 K. A short 100 ps equilibration run is then performed followed by a 500 ns production run. This 500 ns production run comprises our training data. A separate 250 ns production run is performed to generate the in distribution data used for testing. In each case trajectory snapshots are saved every 1 ps, yielding 500,000 simulation frames for the training data and 250,000 simulation frames for the test data.

Coarse-grained data

For the CG representation of ADP, we choose to remove all solvent and represent the molecule using the five backbone carbon and nitrogen atoms (C, N, CA, C, N) and the

carbon beta (CB) of the alanine residue, resulting in a total of six coarse grain atoms. A CG trajectory is generated from the above-mentioned all atom trajectory by slicing the coordinates to retain only the specified coarse grain atoms. The same coarse grain mapping is also applied to the all-atom forces to produce an associated set of instantaneous coarse grain forces. Using both the coarse grain coordinates and forces, bespoke CG force fields are recovered using CGSchNet [152] neural network models. These ADP CG models are trained using the same data set from Refs.[409, 152], and are then used to generate out-of-sample data in the form of CG trajectories as a generalization test for our backmapping method to illustrate performance on real, noisy data. The trajectories used for backmapping consist of 100 separate simulations initalized at random configurations from the reference atomistic data set and containing a total of 4,000 frames each. Sequential CG simulation frames are temporally separated by the same 1 ps spacing as used in the training data trajectory. The training routines and hyperparameters of these CG force field models, as well as the simulation parameters for the out-of-sample CG simulations, are described in the Supporting Information.

Chignolin

Atomistic data

We use reference Chignolin (CLN) trajectories generated from atomistic simulations performed in Ref. [409]. Briefly recapping the protocols, these simulations are performed using the ACEMD program [95] on GPUgrid [410] at 350 K mimicking the setup originally used on the Anton supercomputer simulation [161]. To sufficiently sample folding/unfolding transitions in CLN the data is produced through an MSM-based adaptive sampling strategy, which expedites conformation sampling for this system up to an order of magnitude, [411] consisting of an aggregated $\sim$187 μs of molecular dynamics simulation split into 3,744 short trajectories. Simulation snapshots are spaced by 100 ps culminating in a total of 1,868,861

frames. As each of the 3,744 trajectories are independent we simply split-off 3,650 for training and the remaining 94 for testing to comprise our in distribution test set.

Coarse-grained data

For a CG representation of CLN, we choose to remove all solvent and represent the molecule using just the 10 sequential α-carbons (CA) atoms along the molecular backbone. Following the same procedure described above for ADP, a set of CG trajectories and associated forces are generated. The same CG coordinates and forces mapped from the atomistic data described in the preceding section are then used to train CGSchNet [152] neural network force fields, which in turn are used to generate out-of-sample data for generalization tests of our backmapping model. We produce 1,000 separate trajectories containing 4,000 frames each and, similar to ADP, are initialized at random configurations from the reference atomistic dataset. In this case for CLN each frame is temporally separated by 1 ps, which differs from the 100 ps frame spacing in the training data. While we appreciate this frame spacing represents a different regime than our model is trained to operate in, the inherently accelerated nature of CG dynamics makes direct comparison between CG and atomistic time steps difficult. We appeal to a smaller frame spacing in this work to account for this inherent acceleration of CG dynamics and enable better sampling of short-lived, transient states within these trajectories. As with ADP, the training routines, model hyperparameters, and CG simulation parameters for the CLN coarse grain force fields are detailed in the Supporting Information.

6.4 Results

We present a data-driven and temporally coherent approach for backmapping coarse grained trajectories into full atomistic resolution. Our approach is based on training a conditional variational autoencoder (cVAE) to generate atomistic coordinates of a coarse grain (CG) configuration while also incorporating information from the previous atomistic configura-

tion within the trajectory. The proposed method is applied to two biomolecular systems: alanine dipeptide (ADP) and the miniprotein Chignolin (CLN). The performance of our model is evaluated by measuring its ability to generate backmapped trajectories that preserve atomistic structural, thermodynamic and kinetic properties. Our model is trained on data consisting of pairs of atomistic and corresponding CG trajectories, that are obtained by mapping the atomistic data to CG resolution (Fig. 6.1a). For one evaluation test we apply our method to backmap CG trajectories generated from mapping separate held-out atomistic trajectories to CG resolution, which will be referred to as the *in-distribution* test set, or simply referred to as the *test set* (Fig. 6.1b). For a more challenging test of the generalization capabilities of our model, we apply our method to CG trajectories generated using a bespoke CG force-field CGSchNet [152], which will be referred to as *out-of distribution* or *generalization* set (Fig. 6.1c). For both molecules ADP and CLN the in-distribution and generalization tests show excellent agreement in measures of structural, thermodynamic and kinetic similarity between our backmapped and atomistic trajectory data.

6.4.1 Alanine dipeptide

Energetics

As a first example, we apply our backmapping method to the small molecule alanine dipeptide (ADP) composed of 22 atoms from which we consider a coarse-graining into six CG beads along the peptide backbone. We evaluate structural similarity between atomistic and our backmapped trajectories by comparing distributions of the internal potential energy. The internal energy aggregates contributions from bonded and non-bonded interactions. As such, agreement between atomistic and reconstructed energy distributions serves as a good indicator of overall structural similarity. The energy distribution for the in distribution test set shown in Fig. 6.3a reveals that our model nearly identically reproduces the energy distribution of the original atomistic data. Similarly, the energy distribution for the generalization

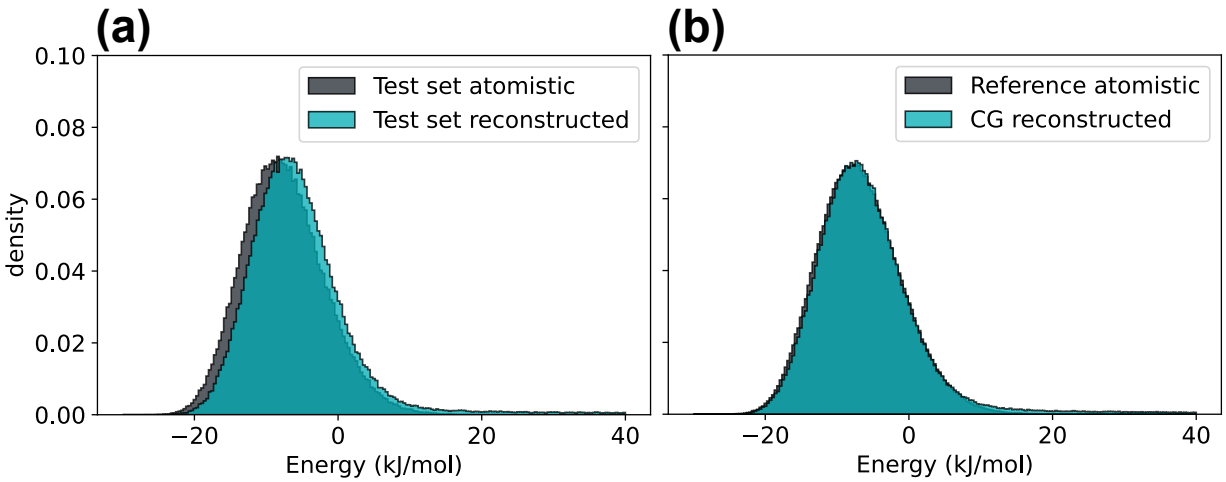


Figure 6.3: Comparison of the internal potential energy distributions for atomistic and backmapped ADP trajectories. **(a)** Internal energies for the held out test set trajectory and the atomistic reconstructed trajectory generated by backmapping a manual coarse graining of the original atomistic trajectory. **(b)** Generalization test comparing reference atomistic trajectories obtained from Ref. [412] to our model-generated backmappings of a real CG simulation conducted using CGSchNet [152].

test set obtained by backmapping a trajectory generated with CGSchNet [152] is displayed in Fig. 6.3b. This out-of distribution test represents a more difficult exercise for our model, as it must generalize to real CG simulated data generated by a different, approximate force field than our model was not exposed to during training. Our model yields backmapped structures that nearly identically match the energetics of the atomistic reference data. The only noticeable deviation are minor shifts in the energy distributions, which are a consequence of a slightly increased population of high-energy backmapped reconstructions (Fig. 6.3). While these rare high-energy configurations can be attributed to minor deviations in a few bond length and angle distributions, we find overall excellent agreement between distributions of local intramolecular features between backmapped and reference structures (Fig. S7-S10 in the Supporting Information). Our backmapping method in general reconstructs atomistic ADP trajectories with high structural and energetic similarity to reference atomistic data while also generalizing well to unseen and real CG force fields, and as a result generates visually identical backmaped structures (Fig. 6.4e and Fig. S6 in the Supporting Information).

Thermodynamics

We evaluate thermodynamic similarity by comparing free energy surfaces (FES) of our reconstructed trajectories to reference atomistic data. We construct our FES in the space of the backbone dihedral angles (ϕ, ψ) as they have proven to be good collective variables for characterizing the conformational states of ADP [413, 414]. In Fig. 6.4a we present the FES in (ϕ, ψ) for the in distribution test set atomistic trajectory compared to the CG backmapped trajectory FES in Fig. 6.4b. Our atomistic reconstruction here reproduces a nearly identical FES to the ground truth atomisitic data demonstrating that for in distribution data our model accurately captures the ground truth thermodynamics. For the generalization test we also present a FES generated from a reference atomistic trajectory (Fig. 6.4c) compared to the FES of a CG backmapped trajectory generated from a CG simulation performed with CGSchNet [152] (Fig. 6.4d). Once again our reconstruction is in excellent agreement with the atomistic reference, importantly correctly identifying the five major meta-stable states of ADP. In Fig. 6.4e we show a superimposed collection of configurations for each of these meta-stable states from the reference, test set reconstructed and CG reconstructed trajectories. For both in distribution and out of distribution data our model reconstructs visually identical configurations with remarkable similarity to the atomistic reference data. Note that the CG model yields configurations throughout the transition paths between meta-stable states, for example ($\phi \approx -2$,$\psi \approx -2$). While those configurations are underrepresented in the atomistic trajectory due to high-energy barriers, the smoothed energy landscape of the CG force field enables broader and more frequent exploration of these regions of phase space. We find our model generalizes well to those sparsely sampled areas and accordingly reconstructs these high-energy configurations (Fig. S11 in the Supporting Information). Overall, our backmapping scheme reproduces the FES for ADP that are in excellent thermodynamic agreement with reference atomistic data for both our in distribution and generalization tests.

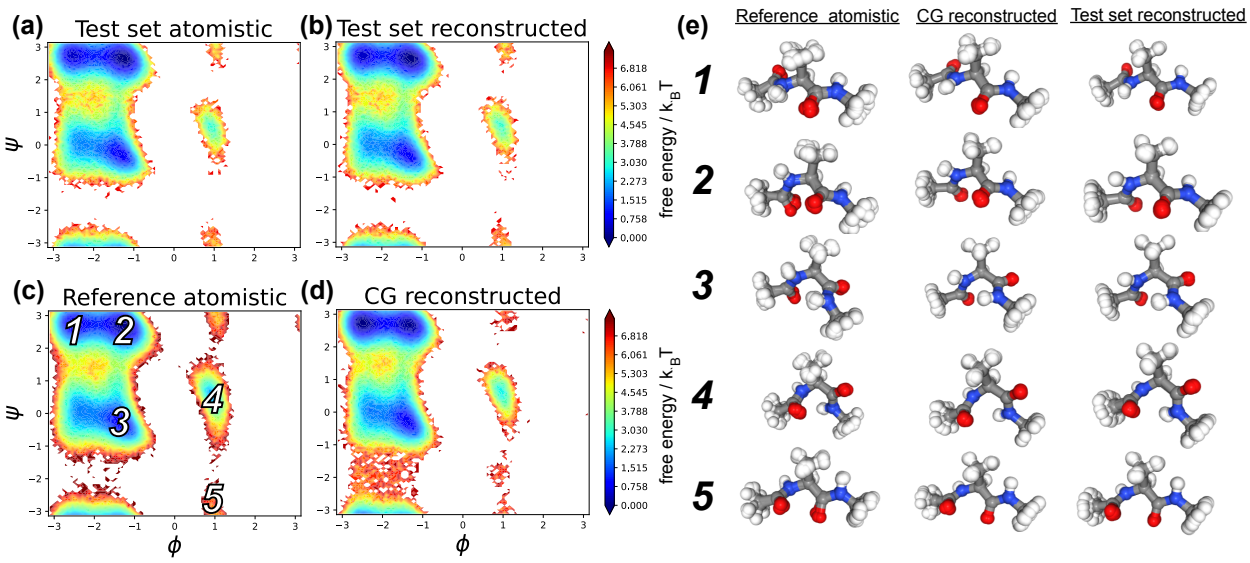


Figure 6.4: Comparison of atomistic and backmapped MSM-reweighted Free Energy Surfaces (FES) for ADP. Ramanchandran plots for the in distribution test set that includes **(a)** a held-out atomistic trajectory and **(b)** our model generated backmapping of the manually coarse grained atomistic trajectory. For a more challenging test of generalizability we compare the Ramachandran plots for **(c)** a reference atomistic trajectory taken from Ref. [412] and **(d)** a backmapping of a CG simulation performed using CGSchNet [152]. Labeled in **(c)** are phase space locations for the five meta-stable states of ADP. **(e)** Seven superimposed configurations near each of these five major meta-stable states from **(c)** the reference atomistic, **(d)** the CG reconstructed and **(b)** the test set reconstructed trajectories.

Kinetics

An important aspect of the proposed method is the incorporation of the previous trajectory configuration as a conditional input for our ML model. This provides temporal information required to achieve temporal coherence between consecutive frames, which is typically omitted in traditional backmapping strategies. Here, we test the temporal coherence of our backmapped trajectories by analyzing kinetics in terms of implied process timescales and velocity distributions.

We compare kinetic agreement between ground truth atomistic data and our backmapped reconstructions by building Markov State Models (MSMs) [415, 416, 417, 418, 419] to characterize and compare the recovered physical processes and timescales. Construction and estimation of MSMs are carried out using the Deeptime [419] software library. We construct

MSMs in ϕ,ψ space for ADP and perform state assignment by discretizing the phase space with 100 k-means centroids fit onto the atomistic data. These same 100 centroids are then also used to generate state assignments for the backmapped trajectory to which we compare. Complete validation of the MSM with associated clustering plots and implied timescale analysis for lag time selection is provided in the Supporting Information. Shown in Fig. 6.5a are the recovered timescales for the test set atomistic and test set backmapped data. Our backmapped trajectory reproduces the atomistic timescales within error for all recoverable processes. A result of using the same state assignments when building the original atomistic and backmapped MSMs is that the elements of the recovered eigenvectors characterize eigenfluxes between the same states. We can therefore measure the cosine similarity between these eigenvectors to quantitatively validate that these timescales correspond to the same physical processes between the original atomistic and the backmapped trajectories. Indeed, shown in Fig. S12a in the Supporting Information we confirm via this cosine similarity measure that the MSM eigenvectors are effectively identically recovered for the backmapped trajectory of in distribution test set.

Shown in Fig. 6.5b is a comparison of the implied timescales between reference atomistic data, backmapped trajectories generated by our model and the original CGSchNet CG simulation. For both the CG backmapped data and the original CG simulation we normalize the implied timescales such that the dominant (slowest) process matches the reference atomistic data. We perform such a normalization to correct for the fact that CG models typically sample configurational landscapes at an accelerated rate compared to atomistic force fields due to absence of explicit solvent and elimination of atomistic degrees of freedom that can cause the coarse-grained degrees of freedom to be accelerated by different scaling factors [122, 123, 420, 124, 125]. This normalization also serves as a visual convenience when comparing deviations of timescales between the accelerated CG dynamics and the atomistic data. An alternative method to this rescaling approach that provides the same

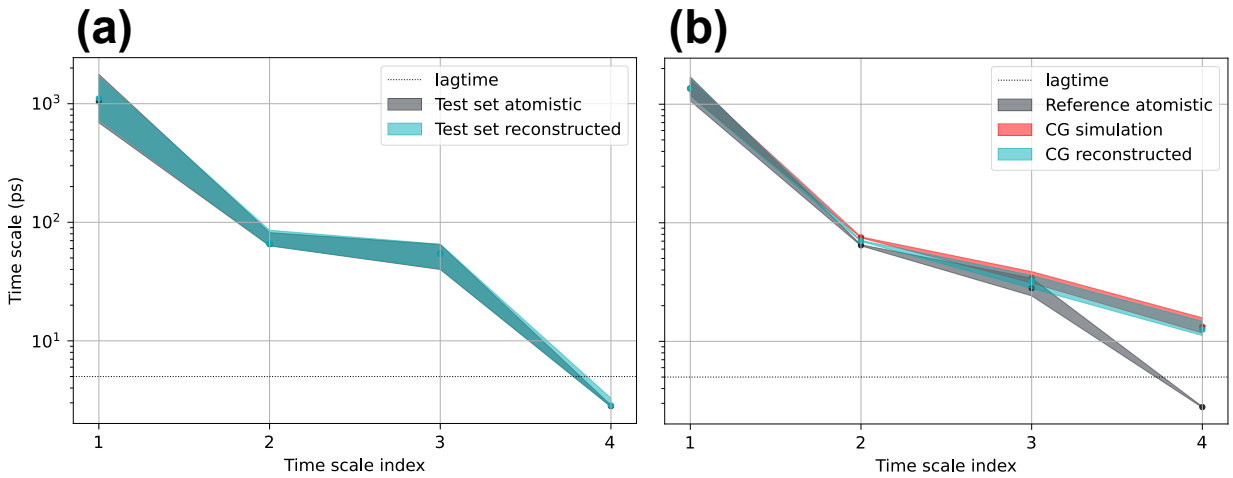


Figure 6.5: Implied timescales of atomistic and backmapped CG trajectories for ADP. **(a)** Comparisonon of timescales recovered for the test set atomistic and test set reconstructed trajectories. **(b)** Implied timescales calculated for the reference atomistic data taken from Ref. [412] compared to the backmapped CG simulation performed with CGSchNet [152] along with the original CG simulation. Timescales for the backampped and CG simulated data are normalized such that the dominant processes is the same as the reference atomistic trajectory. Errors are estimated with Bayesian sampling and represent a 95% confidence interval of the timescales of the sampled MSMs.

information would simply measure timescale ratios instead. We report once again overall excellent agreement of the implied timescales, within error, between our backmapped and reference atomistic trajectories in Fig. 6.5b, with the exception of the fourth, and subsequent, timescales that are faster than our lag time of 5 ps and therefore below the resolution limit of our MSM. We also validate again in Fig. S12b in the Supporting Information by comparing MSM eigenvector similarity that these first three timescales indeed correspond to the same physical processes between these two data sets. The backmapped CG simulated data ultimately reflects the kinetics produced by the coarse-grained model, evidenced by the nearly identical timescales of the CG simulation and backmapping in Fig. 6.5b. Nevertheless, upon normalization of the timescales our excellent agreement for these generalization results suggesting the ratio of timescales between processes is preserved upon backmapping with our model.

A result of conditioning our backmappings using previous atomistic configurations we

find our model is capable of successfully generating backmapped trajectories that reproduce intra-frame root mean square velocity distributions (Fig. 6.6). The distribution of velocities is an effective measure of the deviation of atomic coordinates between consecutive frames. Although our model is not explicitly trained to match intra-frame velocities, the temporal coherence built into our network architecture and training procedure enables the trained network to accurately reproduce these velocity distributions. The excellent agreement we observe for the in distribution data is enabled by the known and directly comparable spacing between frames. The temporal spacing between consecutive frames for the in distribution test set is well-defined by the frame spacing of the atomistic reference simulation, i.e., 1 ps. However, the specific temporal spacing is more obscure for the generalization set, as a direct consequence of the CG dynamics. While most CG models target thermodynamic consistency with a yet higher resolution model or experimentally observed properties, kinetic consistency is typically neglected. In general, CG force fields effectively accelerate simulation dynamics as a consequence of smoothing the energy landscape and lowering energy barriers. However, the timescales for transitions between meta-stable states are typically not rescaled uniformly [122, 123, 420, 124, 125]. To account for this inherent acceleration due to CG force fields, we rescale the velocity distribution of the backmapped generalization set data by a constant factor such that the mean of the velocity distribution matches the mean of the velocity distribution for the test set atomistic data. Applying this empirical correction we see reasonable agreement in the shape of the velocity distributions between the backmapped generalization set data and the native atomistic data, suggesting for ADP here our backmapping model is capable of reconstructing realistic atomic velocities up to a constant scaling factor for data originating from CG force fields.

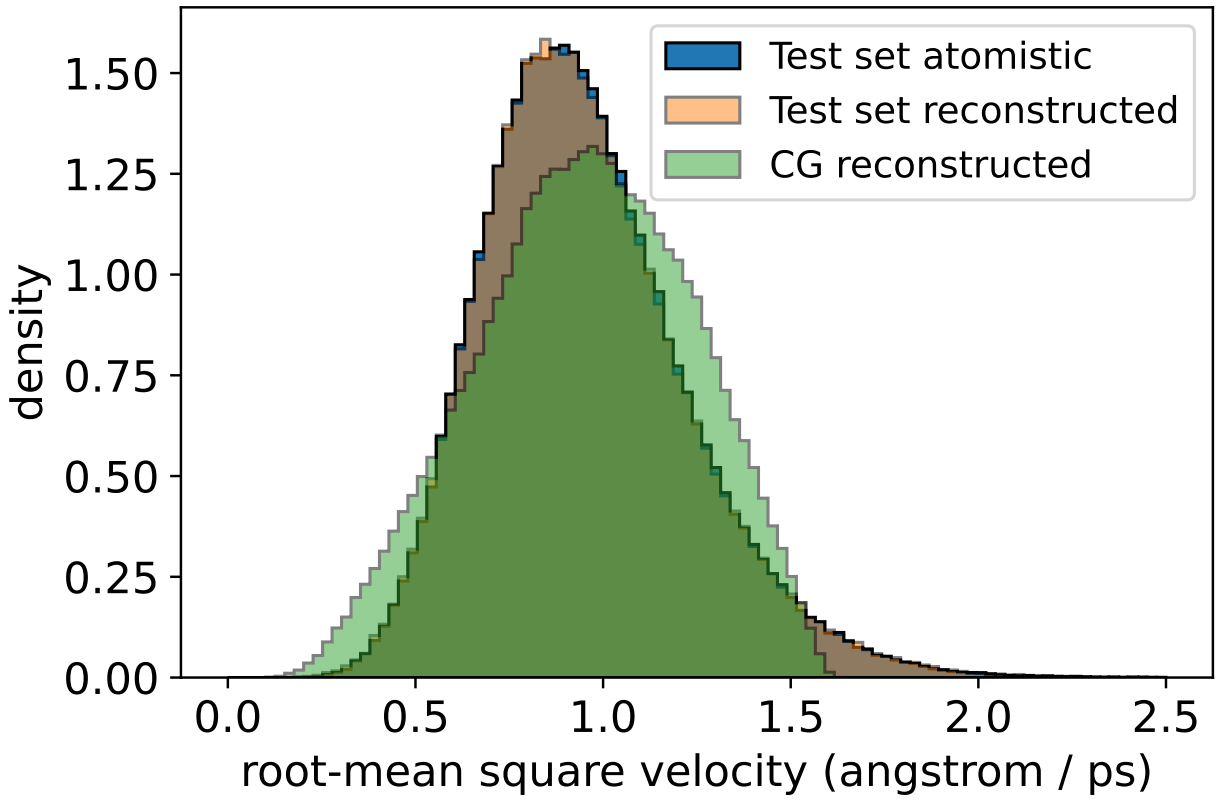


Figure 6.6: Comparison of atomistic and backmapped root mean square velocity distributions for ADP. Velocities are calculated atom-wise between sequential frames using a finite forward difference method and aggregated into a distribution over all atomic velocities. Velocities for the CG reconstructed data corresponding to the generalization set are rescaled by a constant factor ($\sim$3.25) such that the mean of the CG reconstructed velocities matches the mean of the test set atomistic velocities.

6.4.2 Chignolin

Energetics

The second molecular system we apply our backmapping method to is the miniprotin Chignolin (CLN), which is composed of 10 residues with 175 atoms from which we consider a coarse graining into the 10 α-carbons along the peptide backbone. Compared to ADP, the scale and complexity of CLN presents a more challenging test case for both our backmapping and accurate CG force field construction. We once again compare distributions of the internal energy as an indicator for structural similarity between reference atomistic and our backmapped data. Shown in Fig. 6.7a is a comparison of the internal energies for the in distribution test set atomistic and backmapped trajectories, while a comparison for the generalization set data is presented in Fig. 6.7b. These results show excellent energetic overlap between the reference atomistic and our backmapped data for both the in distribution and generalization sets. The internal energies recovered from the backmapped CGSchNet simulation in Fig. 6.7b from the generalization test show longer high-energy tails compared to the in distribution test in Fig. 6.7a. The origins of these higher-energy reconstructed configurations can be explained by an examination of the bond lengths and angles revealing the presence of very slightly contracted bond length distributions in the generalization set compared to the in distribution test set (Figs. S23-S26 in the Supporting Information). Our backmapped trajectories nevertheless show overall excellent agreement in these local intermolecular features and energetics, and therefore also produce visually convincing atomistic reconstructions (Fig. 6.8e and Fig. S6 in the Supporting Information).

Thermodynamics

We construct Free Energy Surfaces (FES) using the basis recovered from Time-lagged Independent Component Analysis (TICA) [421, 413, 422, 423] to compare thermodynamic

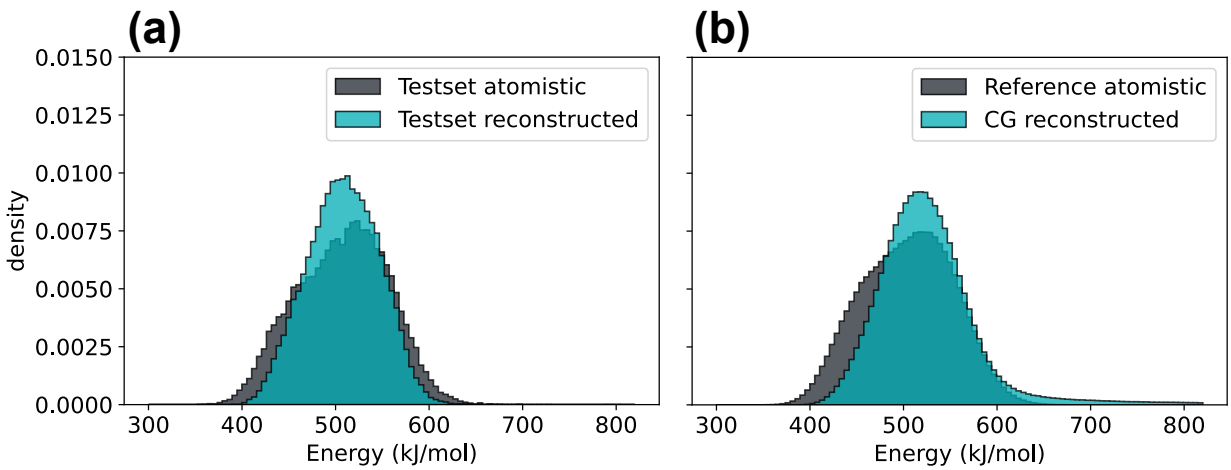


Figure 6.7: Potential energy distributions of atomistic and backmapped CLN trajectories for **(a)** the in distribution test set and **(b)** the generalization set.

similarity between reference atomistic data and our backmapped trajectories. Unlike ADP, there are no simple intuitive variables capable of compactly representing the CLN FES, so we instead use TICA as a dimensionality reduction technique to extract a low-dimensional representation for the CLN phase space. Using our entire reference atomistic dataset we first learn a TICA embedding into the first two non-trivial Independent Components (ICs) using all 45 pairwise α-carbon distances as features [409, 152, 424, 425]. This learned TICA model provides us with a fixed basis set we use for constructing an FES in these two leading ICs for the atomistic and backmapped data from both our in distribution and generalization data sets.

Shown in Fig. 6.8 is a comparison of these CLN FESs for in distribution test set atomistic (Fig. 6.8a) and backmapped (Fig. 6.8b) trajectories, alongside the generalization set reference atomistic FES (Fig. 6.8c) and the FES obtained from backmapping a CG simulation performed with CGSchNet [152] (Fig. 6.8d). In each case the backmapped FES recovers the presence of the three primary meta-stable states in CLN and produces visually identical atomic reconstructions compared to the atomistic reference (Fig. 6.8e). We notice our backmapped FES are contracted near regions corresponding to the folded state (labeled *1*)

and the misfolded state (labeled *2*) compared to the unfolded ensemble (labeled *3*). Correspondingly, backmapped structures extracted from the folded and misfolded states display slightly less configurational variability compared to atomistic reference than structures visualized from the unfolded ensemble (Fig. 6.8e). We suspect this is due to a loss-conserving strategy by the network: comparatively less loss is sacrificed by predicting a (nearly) fixed configuration for CG structures corresponding to folded and misfolded states compared to the benefit of capturing the diversity of possible structures in the unfolded ensemble. This results in more comprehensive coverage of the unfolded ensemble in our backmapped data compared to the folded and misfolded states. We also notice that the misfolded state is noticeably less stable in the generalization set backmapping (Fig. 6.8d) compared to the reference atomistic data (Fig. 6.8c). Indeed, we find this property to be a reflection of the misfolded state being inherently less stable in the original CG simulation as well (Fig. S27 in the Supporting Information). While our backmapping method does a reasonable job of reproducing the atomistic FES, we find that are ultimately limited by the underlying accuracy of the CG simulation that we backmap.

Kinetics

We evaluate kinetic similarity between reference atomistic data and our backmapped trajectories by comparing similarity of MSM-recovered implied timescales and processes. We perform state space discretization for our MSMs within the same two-dimensional TICA basis used to construct our FES. Data from the reference atomistic trajectories in this two-dimensional TICA space is fit to determine 150 k-means centroids identifying the state space decomposition. We then use these same 150 centroids to generate state assignments when building separate MSMs for other atomistic and backmapped data within the in distribution and generalization data sets. Using the same TICA projection and set of cluster centers between MSMs ensures direct comparability of the recovered eigenvectors and allows us to

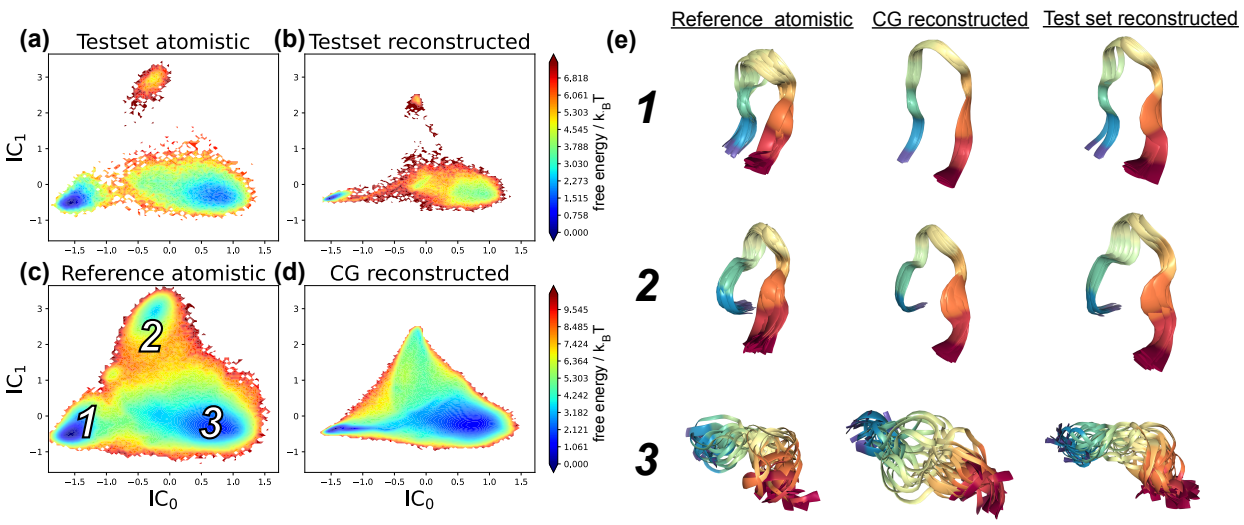


Figure 6.8: Comparison of the atomistic and backmapped MSM-reweighted Free Energy Surface (FES) for CLN. The FES for CLN is represented in the basis of the two leading non-trivial Independent Components (ICs) of a Time-lagged Independent Component Analysis (TICA) model fit to the reference atomic dataset. Shown are FES of the in distribution test set atomistic **(a)** and backmapped **(b)** trajectories, alongside the reference atomistic data **(c)** and the corresponding backmapped CGSchNet simulation **(d)** from the generalization set. **(e)** Visualization for a collection of 25 superposed structures from the three meta-stable states taken from the reference atomistic data **(c)**, the backmapped CG reconstructed data **(d)** and backmapped in distribution test set trajectory **(b)**.

determine similarity of the recovered processes. We use the same methodology as with ADP here for CLN for quantifying similarity of processes by measuring the cosine similarity of the MSM eigenvectors. In the cases where not all 150 states are utilized, we construct MSMs using only occupied states and compute cosine similarity using only this subset of mutually occupied states between the two MSMs being compared. Using this approach we can quantitatively ensure the that implied timescales in fact correspond to the same physical processes between the two data sets. As with ADP, complete details on MSM construction and validation are provided in the Supporting Information.

Presented in Fig. 6.9a is a comparison of implied timescales between atomistic and backmapped trajectories for the in distribution test set. For the in distribution test set our backmapped reconstruction precisely reproduces within error all implied timescales. The large gap between the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ timescales suggests the majority kinetic variance is cap-

tured in the first two processes, which we recover with >90% cosine similarity, and for the final remaining processes greater than our lag time – and therefore the only other resolvable process by our MSM – we also recover with ∼80% cosine similarity (Fig. S28a in the Supporting Information). Comparison of implied timescales for the generalization set between the reference atomistic data, backmapped CGSchNet [152] simulation and the original CG simulation is shown in Fig. 6.9b. We follow the same approach here as with ADP, where we normalize the implied timescales for data generated from CG simulations by a constant factor such that their slowest timescale matches the slowest timescale from the reference atomistic data. This normalization accounts for the inherently accelerated dynamics of CG simulations and enables us to effectively compare the ratio of implied timescales [122, 123, 420, 124, 125]. The ratio between the first two timescales is poorly recovered for both the backmapped data and the original CG simulation compared to the reference atomistic data. Comparison of the cosine similarity between these processes reveals the 1^{st} processes is recovered with ∼60% similarity for both the backmapped data and the CG simulation, while the 2^{nd} processes is recovered with also ∼66% similarity for the backmapped data and ∼90% similarity for the original CG simulation (Fig. S28b in the Supporting Information). While the ratio in the timescales for the faster 3^{rd} and 4^{th} processes seems to be better conserved in the original CG simulation, our backmapped data actually recovers these processes with slightly better similarity than the original CG simulation (Fig. S28b in the Supporting Information). The fact that our method precisely recovers kinetics for the in distribution data but can only approximately recovers kinetics when backmapping real CG simulated data confirms that our method produces backmapped trajectory data that is largely a reflection of the kinetics expressed in the underlying CG data.

Lastly, we compare velocity distributions as an indicator of temporal coherence between sequential frames in our backmapped reconstructions. Shown in Fig. 6.10 is a comparison of frame-by-frame root mean square velocity distributions for the reference and backmapped

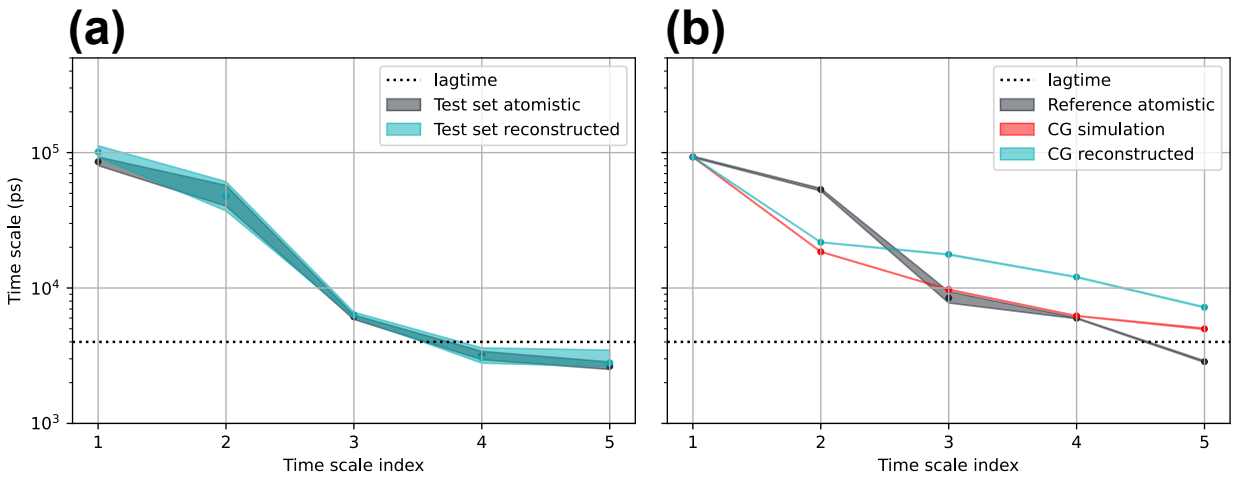


Figure 6.9: Implied timescales of atomistic and backmapped CG trajectories for CLN. **(a)** Comparison of implied timescales from atomistic and backmapped trajectories for the in distribution test set. **(b)** Implied timescales for the reference atomistic data, the original CGSchNet simulation and the backmapped CGSchNet simulation. Timescales for the original CG simulation and backmapped data are normalized such that the dominant (slowest) processes matches the slowest reference atomistic timescale.

in distribution test set data alongside the backmapped CGSchNet simulation data. For the in distribution test set the velocities are an excellent match between the reference and backmapped data. Mimicking our approach with ADP, to account for the inherently accelerated dynamics of the CGSchNet force field we use a constant factor to rescale the atomic velocities such that the mean of the backmapped generalization set velocities matches the mean of the test set atomistic velocities. After applying this constant scaling we notice excellent agreement in the shape of the velocity distributions between the backmapped generalization set data and the original atomistic data. We note that the scaling factor used to correct the CLN velocities ($\sim$17.81) is much larger than the scaling factor used for the ADP velocities ($\sim$3.25), suggesting a more substantial acceleration of backmapped CLN dynamics due to the CG force field compared to ADP. This relative speed-up of the backmapped CLN CG simulation could be attributed to the more severe coarse-graining of CLN from 175 atoms to 10 beads (17.5$\times$ reduction) compared to the ADP model from 22 atoms to 6 beads ($\sim$3.67$\times$ reduction). This more dramatic reduction in the degrees of freedom upon

coarse-graining for the CLN model could result in the CLN CG simulation operating in a comparatively 'smoother' free energy surface than ADP, and therefore leading to the relatively faster atomic motions identified by the velocity distributions.

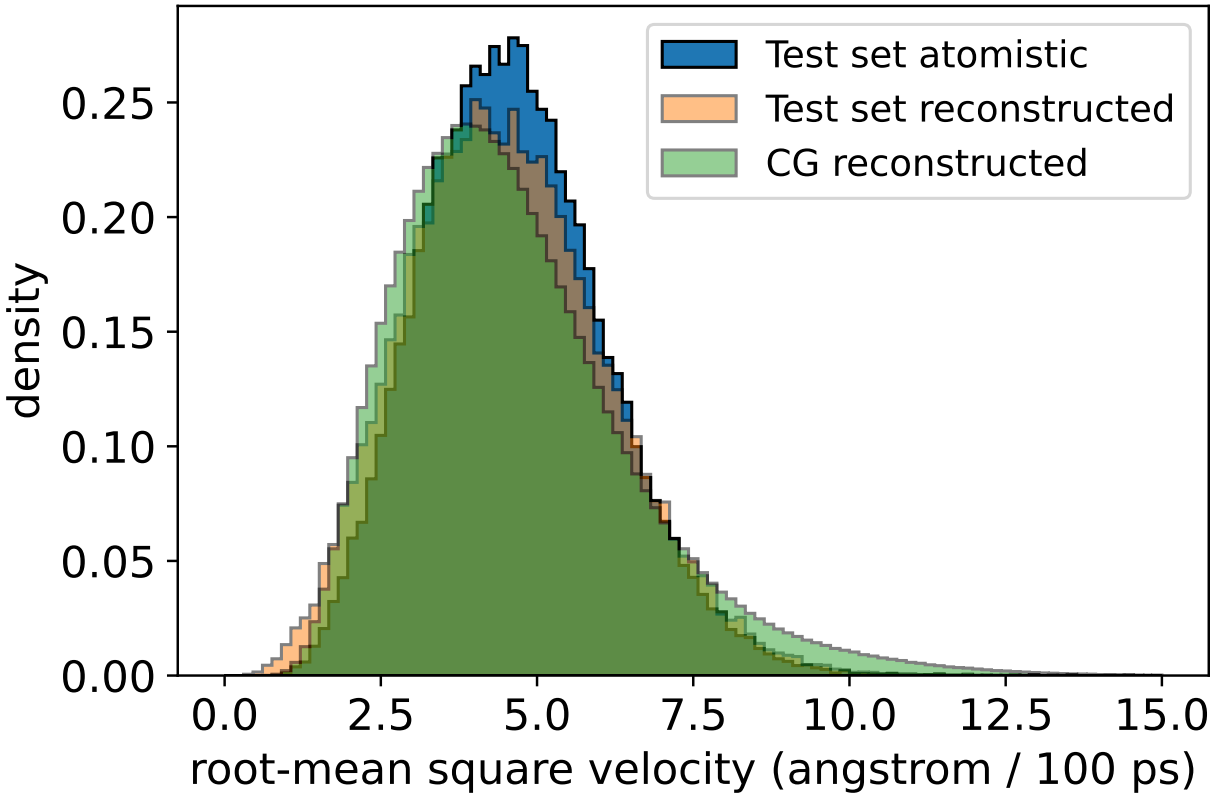


Figure 6.10: Comparison of atomistic and backmapped root mean square velocity distributions for CLN. Velocities are calculated frame-by-frame using a finite forward difference method and represented in angstrom / 100 ps, as 100 ps is the native temporal spacing between frames in the reference atomistic data [409]. The backmapped generalization set velocities are rescaled by a constant factor ($\sim$17.81) such that the mean of the CG reconstructed velocities equals the mean of the test set atomistic velocities.

6.5 Discussion & Conclusions

We present in this work a data-driven and temporally coherent scheme for backmapping CG trajectories into atomistic resolution. Our approach trains a conditional variational autoencoder (cVAE) to reconstruct atomistic detail given the target CG configuration and the

previous atomistic structure. Our method is showcased here to backmap two biomolecular systems: alanine dipeptide (ADP) and the miniprotein chignolin (CLN) – systems that are frequently used as the test bed in the demonstration of new methods in molecular dynamics simulation [152, 142, 426, 427, 428]. We train our model using a reference atomistic trajectory which we coarse-grain *post hoc* to produce exemplar pairs of atomistic and CG configurations (Fig. 6.1a). We test that our backmapping method on both in distribution data generated from backmapping a CG trajectory produced by coarse-graining held-out atomistic data (Fig. 6.1b), and out of distribution data generated from a real CG simulation performed using CGSchNet [152]. We evaluate the performance of our model in terms of capability to reproduce structural, thermodynamic and kinetic properties of reference atomistic systems. To this end, structural similarity is probed by comparing distributions of potential energies and local structural features, such as bond lengths and angles. Thermodynamic similarity is tested by analyzing free energy surfaces that are constructed in terms of collective variables. Kinetic agreement is tested by comparing implied timescales of processes identified by MSMs, while temporal coherence between consecutive frames is analyzed in terms of intra-frame velocity distributions. Our model yields backmapped trajectories for the in distribution test set that are in good agreement with atomistic data. Moreover, our model generalizes well producing convincing atomic reconstructions for out of distribution data obtained from CGSchNet simulations. While we notice slightly better generalizability to CGSchNet simulations of ADP compared to the more complex CLN molecule, we find our method generates backmapped trajectories that largely maintain thermodynamic and kinetic properties reflected in the original CG simulation while also reconstructing atomistic velocities up to a constant scaling factor.

As coarse-grained models are typically designed to accelerate conformational space sampling, this can lead to legitimate coarse-grained configurations which correspond to under represented atomistic training structures. Data-driven backmapping should therefore strive

to effectively extrapolate on the atomistic training data and generate reasonable backmapped atomistic reconstructions. Our method in general performs well when backmapping coarse-grained models that express good phase space overlap between coarse-grained and atomistic resolutions, but if this phase space deviation becomes too large due to poor coarse-grained models we may expect our backmapping to help heal these errors to some extent through training, but there is a limit to this remediation and at some point it must fail. As a result, errors in our backmapped reconstructions on the out of distribution generalization data can be attributed to both the performance and expressivity of our backmapping cVAE model compounded with biases and inaccuracies of the CGSchNet force field used to generate the CG trajectories. In the case of ADP, the highly accurate CGSchNet model leads to overall good backmapping generalization performance. For CLN, the CGSchNet model is comparatively lower-fidelity which we find results in slightly poorer generalization by our backmapping model in comparison to ADP.

Future work will strive to improve upon transferability to different coarse-grained mappings, data efficiency, and training/inference routines of our method. As a possible approach to avoid training separate bespoke models for different coarse-grained mappings of the same atomistic structure, a single set of atomistic simulation data could be used to train a model capable of backmapping a number of different coarse-grained representations. One idea in this direction is a hierarchical backmapping approach which involves an autoregressive component where separate prediction heads are tasked with reconstructing intermediate coarse-grained representations at progressively more detailed resolutions, conditioned on the preceedingly generated configurations until the final atomistic structure is produced. Transferability in this aspect would eliminate the need for retraining separate large models for each system through utilizing only a fixed set of atomistic simulation data for training. Currently, the backbone of our model primarily uses convolutional neutral networks (CNNs) operating on voxelized representations that are converted to and from Cartesian coordinates. Using

explicitly covariant network architectures [429, 430, 431], such as those employed in the backmapping scheme by Wang et al. [142], can lead to superior data efficiencies without the need to train with random rotations and the potential to massively reduce the network size and memory requirements compared to voxelizations improving scalability. This approach specifically uses graph neural networks with *E(3)* equivariant operations to perform generative frame-by-frame backmapping. Temporal coherence could possibly be incorporated within this framework by including preceding trajectory configurations as conditioning variables for the structure generation processes. Our approach is also currently designed for configurational backmapping of superatoms into high resolution atomistic detail, but our scheme for achieving temporal coherence could also be extended to different definitions of coarse-graining such as the dynamical coarse-graining employed in molecular latent space simulators by Ferguson and coworkers [432]. Training routines could also be augmented to incorporate more inductive biases that may benefit backmapping, such as: (i) better emphasizing sparsely populated regions of configurational space, which could be accomplished by accompanying training samples with thermodynamic or dynamical path weights; (ii) an autoregressive training protocol could be employed to improve the temporal coherence by using a recurrent approach to predict multiple consecutive frames each forward pass; (iii) further encourage the model to utilize knowledge of preceding trajectory frames by augmenting the training loss to incorporate information that is explicitly based on velocities or higher order time derivatives.

Data Availability

A complete PyTorch implementation of our model with associated training routines, data, and pretrained model weights for ADP and CLN along with Jupyter Notebooks demonstrating the energetic, thermodynamic, and kinetic analyses performed in this work is provided via the Materials Data Facility (MDF) [240, 241] at DOI:10.18126/tf0h-w0jz [396].

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.2c07716

CHAPTER 7

GIRSANOV REWEIGHTING ENHANCED SAMPLING TECHNIQUE (GREST): ON-THE-FLY DATA-DRIVEN DISCOVERY OF AND ENHANCED SAMPLING IN SLOW COLLECTIVE VARIABLES

7.1 Abstract

Molecular dynamics simulations of microscopic phenomena are limited by the short integration time steps required for numerical stability but which limit the practically achievable simulation time scales. Collective variable (CV) enhanced sampling techniques apply biases to pre-defined collective coordinates to promote barrier crossing, phase space exploration, and sampling of rare events. The efficacy of these techniques is contingent on the selection of good CVs correlated with the molecular motions governing the long-time dynamical evolution of the system. In this work, we introduce Girsanov Reweighting Enhanced Sampling Technique (GREST) as an adaptive sampling scheme that interleaves rounds of data-driven slow CV discovery and enhanced sampling along these coordinates. Since slow CVs are inherently dynamical quantities, a key ingredient in our approach is the use of both thermodynamic and dynamical Girsanov reweighting corrections for rigorous estimation of slow CVs from biased simulation data. We demonstrate our approach on a toy 1D 4-well potential, a simple biomolecular system alanine dipeptide, and the Trp-Leu-Ala-Leu-Leu (WLALL) pentapeptide. In each case GREST learns appropriate slow CVs and drives sampling of all thermally accessible metastable states starting from zero prior knowledge of the system. We make GREST accessible to the community via a publicly available open source Python package.

This chapter is based on the work reported in: **Shmilovich, K.**, & Ferguson, A. L.

(2023). "Girsanov Reweighting Enhanced Sampling Technique (GREST): On-the-Fly Data-Driven Discovery of and Enhanced Sampling in Slow Collective Variables." *The Journal of Physical Chemistry A,* 127(15), 3497-3517.

7.2 Introduction

Molecular dynamics (MD) simulations serve as a virtual microscope to provide microscopic understanding of the atomic-scale structure, thermodynamics, and dynamics of molecular and condensed-matter systems. By numerically integrating the classical equations of motion, MD simulations can accurately describe the evolution of molecular systems [433]. Advances in computational hardware [88, 89] and simulation software [90, 91, 92] have enabled simulations of systems containing up to trillions of atoms [93]. While the accessible size of systems continues to grow, a fundamental limitation of MD is the need for short integration time steps commensurate with the fastest atomic motions – typically on the order of femtoseconds – that limits practically achievable time scales to approximately microseconds to commodity hardware [94, 95] and milliseconds on specialized hardware [89]. Many interesting processes such protein folding [97], protein-protein association [98], ligand binding [99] and molecular self-assembly [100] occur near or far beyond these attainable time scales, which has motivated the development of computational strategies to facilitate the study of slow molecular motions and rare events.

Techniques like the string method [434], transition path sampling [435], and forward flux sampling [436] enable the study of rare events using astute initialization of simulation trajectories to explore the configuration space along paths connecting pre-defined "reactant" and "product" states (e.g., folded and unfolded states of a protein). Given these two end states, these techniques present efficient means to sample the pathways and associated dynamical properties of reactive transitions. Another class of popular methods are enhanced sampling techniques that seek to accelerate sampling not of a single reactive transition, but to promote

global exploration of configurational space by applying biasing potentials along predefined collective variables (CVs) [101]. Well chosen CVs correlated with transitions of interest and approximately orthogonal to separating free energy barriers can result in an increased frequency of barrier crossing to drive rare transitions and improve global sampling of configurational phase space. By exploiting this feature and modifying the system Hamiltonian to sample configurational space more efficiently, thermodynamic reweighting can be employed to estimate thermodynamic averages under the unbiased Hamiltonian. A plethora of methods have been developed for constructing biasing potentials along predefined CVs including, for example, metadynamics [398], hyperdynamics [437], umbrella sampling [399], thermodynamic integration [438], and variational enhanced sampling [400]. While these methods are demonstratively powerful tools for practitioners, their success is contingent upon the availability of good CVs in which to accelerate sampling and surmount the free energy barriers separating the thermally relevant metastable states. Accelerating poorly chosen CVs typically leads to little to no improvement in sampling over unbiased MD simulations.

The manual definition of good CVs is challenging for large, complex molecular systems where intuition is frequently limited. To help ameliorate this issue of proper CV selection, data-driven techniques can deploy machine learning (ML) to automatically discover optimal CVs for biasing. The central premise behind these approaches is an assumption of emergent low-dimensionality, and it is a generic feature of molecular systems that cooperative couplings between the atomic degrees of freedom lead to emergent low-dimensionality on sufficiently long time scales [102, 103, 104, 105, 106, 107, 104]. Geometrically, once can view the dynamical evolution within the $3N$-dimensional configurational phase space comprising the Cartesian coordinates of the N constituent atoms to be approximately restrained to a low-dimensional manifold supporting a dynamical attractor parameterized by the leading slow modes of the system (e.g., large-scale folding/unfolding of a protein) and to which the fast degrees of freedom (e.g., side chain rotations) are annealed as effective noise. In this

sense, CV discovery can be viewed as a form of dimensionality reduction for identifying variables from data that best characterize the slowest evolving (i.e., maximally autocorrelated) degrees of freedom. A central challenge in data-driven approaches for CV discovery is a "chicken-and-egg problem": performing good configuration space sampling requires high-quality CVs to bias, but discovery of high-quality CVs requires sufficient data obtained from good configuration space sampling [107]. Consequently, data-driven methods typically strive to overcome this conundrum by iterating between rounds of enhanced sampling and CV learning to incrementally improve configuration space coverage. To this end, a variety of methods performing on-the-fly discovery of CVs and enhanced sampling have been proposed, including: REinforcement learning based Adaptive samPling (REAP) [439], Reweighted Autoencoded Variational Bayes for Enhanced sampling (RAVE) [401], Diffusion-Map-directed MD (DM-d-MD) [402], intrinsic Map Dynamics (iMapD) [403], and Molecular Enhanced Sampling with Autoencoders (MESA) [404]. A commonality between many of these adaptive sampling strategies is that they typically seek to identify high-variance CVs correlated with the elongated directions in configurational phase space containing the highest variance in the simulation trajectory. One may typically expect the high-variance directions to also be slowly evolving since they generally correspond to large-scale delocalized motions, but for the purposes of accelerating sampling along the slow collective motions, one would typically prefer an approach capable of identifying them directly.

Data-driven learning of slow CVs can be accomplished by appealing to the Variational Approach to Conformational dynamics (VAC) [440]. The VAC formalism presents a variational approach to numerically estimate the maximally autocorrelated CVs by empirical analysis of time series data. The VAC underpins a number of popular techniques for slow CV discovery, including Time-Lagged Independent Component analysis (TICA) [441, 421, 442], kernel TICA (kTICA) [443], and State-free Reversible VAMPNets (SRVs) [444], also known as Deep-TICA [426]. In particular, SRVs have been demonstrated as a powerful deep learning-

enabled means to estimate slow variables from simulation data that have been applied to produce high-resolution models of Trp-cage miniprotein dynamics [445] and describe DNA (de)hybridization [446]. Crucially, since slow CVs are inherently dynamic observables, their recovery from biased simulation data requires the use of dynamical, in addition to thermodynamic, corrections which are sufficient for the reweighting of purely configurational quantities [111, 447, 112]. Methods to rigorously apply these dynamical corrections are necessary to solve the chicken and egg problem by iterative rounds of slow CV discovery and enhanced sampling exploration of configurational phase space.

A theoretical basis for dynamical reweighting is provided by path reweighting methods rooted in the Girsanov theorem [108, 109, 110, 111, 112]. These approaches are similar to dynamical reweighting schemes that reweight discrete transition probabilities [448, 449, 450, 451, 452], but more powerful and general in that it is applied directly to the transition density elements and so does not require the assumption of local equilibrium or a predefined discretization of configurational phase space [109, 110, 111]. Conceptually, Girsanov reweighting is a technique to correct the probability weights of dynamical pathways collected under a perturbed system Hamiltonian to those under an unperturbed Hamiltonian by applying dynamical reweighting of the transition density elements along the dynamical path through phase space [110]. Expectations of dynamical observables can be formulated as expectations in a path ensemble, and Girsanov reweighting provides a means to estimate unbiased averages from trajectories collected under an applied bias. Appealing to precisely the same rationale for enhanced sampling techniques in the estimation of configurational observables, applied biases that improve sampling of the phase space by accelerating rare transitions can lead to higher accuracy and lower variance estimates of dynamical observables.

Path reweighting approaches present a powerful tool for the estimation of dynamical observables from biased trajectories, but do have two important limitations. First, analytical expressions for path reweighting factors are only available for particular choices of stochas-

tic integrators, specifically the Euler-Maruyama (E-M) scheme to integrate the overdamped Langevin (i.e., Brownian dynamics) equation [109, 110, 111] and the Izaguirre, Sweet, and Pande (ISP) scheme – a variant of the Langevin leapfrog scheme – to integrate the (underdamped) Langevin equation [111, 453]. Elegant recent work by Kieninger and Keller developed a prescription to derive path weights for other Langevin integration schemes and presented evidence that approximate E-M reweighting expressions may offer high accuracy approximations for other Langevin integrators, thereby generalizing the applicability of the methodology to popular numerical schemes employed in the molecular simulation community [111]. Second, the path reweighting factors can become extremely large or extremely small for large applied biases risking numerical instability and overflow/underflow. This means that there typically exists a "Goldilocks region" for the intensity of the applied bias that is sufficiently high to promote good sampling but sufficiently low for stable convergence and low-variance estimation of the path ensemble expectations [109, 110, 112].

There have been relatively few applications of Girsanov reweighting to molecular systems. Viewing temperature as a simple modification of the system Hamiltonian, Chodera, Swope, Noé, Prinz, Shirts, Smith, and Pande performed path reweighting of simulation trajectories at various temperatures within a parallel tempering / replica exchange framework to estimate dynamical expectations from simulations conducted different temperatures under an Anderson thermostat, overdamped Langevin dynamics, and Langevin dynamics [454, 447]. Weber and Pande employed the Girsanov approach to construct Markov state models (MSMs) for the 20-residue Trp-cage fast-folding mini-protein using biased simulation trajectories in which the dynamical evolution of a key dihedral angle was accelerated using artificial biasing forces [112]. Donati and Keller presented an elegant theoretical basis and devised efficient numerical schemes that were applied to construct Markov state models of alanine dipeptide, valine dipaptide, VGVAPG hexapeptide, and a β-hairpin peptide by on-the-fly Girsanov reweighting of metadynamics enhanced sampling trajectories in which

accelerating biases were applied to pre-selected backbone or side-chain dihedral angles, key hydrogen bonds, or the molecular end-to-end distance [109, 110]. In all studies to date, the CVs to accelerate, or more generally the variables/parameters in which the Hamiltonian was modified, were pre-defined. The success of the CV enhanced sampling is predicated on knowledge of CVs coincident with the slow motions of the system, and has limited applications to systems where this knowledge is *a priori* available.

In the present work, we present Girsanov Reweighting Enhanced Sampling Technique (GREST) as an adaptive sampling scheme that performs interleaved rounds of data-driven slow CV discovery and enhanced sampling along those coordinates. Our approach is analogous to our previously introduced Molecular Enhanced Sampling with Autoencoders (MESA) method for interleaved data-driven discovery and enhanced sampling in *high variance* CVs [404], but by appealing to Girsanov reweighting GREST extends this approach to *slow* (i.e., maximally autocorrelated) CVs. We build upon the elegant theoretical and numerical work by Keller and co-workers [109, 110] to perform nonlinear CV discovery using SRVs as a deep learning-enabled version of TICA that employs both thermodynamic corrections and Girsanov path weights to recover slow CVs from the biased simulation trajectories [444]. This modification of the SRV objective function enables the formally correct numerical estimation of slow collective variables from short, biased, and discontinuous trajectories. This estimation process is similar to that proposed by Bonati et al. [426], but we explicitly include the Girsanov path weights in addition to the thermodynamic corrections to rigorously account for the influence of the bias upon the dynamics, and, rather than a one-shot discovery paradigm, we perform discovery within a virtuous iterative loop to simultaneously converge sampling of configurational phase space and the learned slow CVs. We demonstrate GREST in applications to two biomolecular systems – alanine dipeptide and the WLALL pentapeptide – and make the approach freely available to the community via publicly available open source Python packages accompanied with example notebooks. We anticipate that GREST may be

a useful tool for rapid exploration of the configurational phase space of molecular systems without the requirement for prior knowledge of the important underlying slow dynamical modes, and a means to advance fundamental understanding of the slow molecular modes, guide rational molecular engineering, and furnish slow subspaces for the parameterization of efficient reduced-dimensional dynamical simulators [432, 455].

7.3 Methods

7.3.1 Transfer operator theory

The transfer operator $\mathcal{T}_\tau$ of a dynamical system possessing configurational microstates $\mathbf{x} \in \mathbb{R}^n$ propagates probability distributions over configurational microstates $q_t(\mathbf{x})$ normalized with respect to the equilibrium probability distribution $u_t(\mathbf{x}) = q_t(\mathbf{x})/\pi(\mathbf{x})$ under transition densities $p(\mathbf{x}_{t+\tau}|\mathbf{x}_t)$ at a lag time τ [456, 457],

$$u_{t+\tau}(\mathbf{x}) = \mathcal{T}_\tau \circ u_t(\mathbf{x}) = \frac{1}{\pi(\mathbf{x})} \int p_\tau(\mathbf{x}|\mathbf{x}')u_t(\mathbf{x}')\pi(\mathbf{x}')d\mathbf{x}'. \tag{7.1}$$

Here $p_\tau(\mathbf{x}|\mathbf{x}') = \mathbb{P}(\mathbf{x} = \mathbf{x}_{t+\tau}|\mathbf{x}' = \mathbf{x}_t)$ is the (normalized) conditional probability density that a system in microstate $\mathbf{x}' = \mathbf{x}_t$ at time t will advance to state $\mathbf{x} = \mathbf{x}_{t+\tau}$ at time $(t+\tau)$. Crucially, Eqn. 7.1 assumes Markovianity, which is an increasingly good approximation at large lag times τ, and time homogeneity of the transition densities $p_\tau(\mathbf{x}|\mathbf{x}')$. For equilibrium systems satisfying detailed balance, we have the additional condition that,

$$\pi(\mathbf{x})p_\tau(\mathbf{x}'|\mathbf{x}) = \pi(\mathbf{x}')p_\tau(\mathbf{x}|\mathbf{x}'), \tag{7.2}$$

which implies the transfer operator $\mathcal{T}_\tau$, also known as the Perron-Frobenius operator describing the evolution of densities with respect to the equilibrium density, becomes identical to the Koopman operator describing the evolution of observables, and self-adjoint with respect

to the equilibrium distribution [456],

$$\begin{aligned}
\langle \mathcal{T}_\tau \circ g, h \rangle_\pi &= \int \mathcal{T}_\tau \circ g(\mathbf{x}) h(\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} \\
&= \int \left[\frac{1}{\pi(\mathbf{x})} \int p_\tau(\mathbf{x}|\mathbf{x}') g(\mathbf{x}') \pi(\mathbf{x}') d\mathbf{x}' \right] h(\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} \\
&= \int \int \frac{\pi(\mathbf{x})}{\pi(\mathbf{x}')} p_\tau(\mathbf{x}'|\mathbf{x}) h(\mathbf{x}) g(\mathbf{x}') \pi(\mathbf{x}') d\mathbf{x}' d\mathbf{x} \\
&= \int \left[\frac{1}{\pi(\mathbf{x}')} \int p_\tau(\mathbf{x}'|\mathbf{x}) h(\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} \right] g(\mathbf{x}') \pi(\mathbf{x}') d\mathbf{x}' \\
&= \int \mathcal{T}_\tau \circ h(\mathbf{x}') g(\mathbf{x}') \pi(\mathbf{x}') d\mathbf{x}' \\
&= \langle g, \mathcal{T}_\tau \circ h \rangle_\pi. \qquad (7.3)
\end{aligned}$$

The self-adjoint nature of $\mathcal{T}_\tau$ under the detailed balance condition implies that it admits diagonalization into a set of eigenfunctions $\{\psi_i(\mathbf{x})\}$ forming a complete orthonormal basis, and that the associated eigenvalues $\{\lambda_i\}$ are real and bounded from above by $1 = \lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \ldots$ [456, 458, 413, 422],

$$\mathcal{T}_\tau \circ \psi_i(\mathbf{x}) = \lambda_i \psi_i(\mathbf{x}), \qquad \langle \psi_i(\mathbf{x}) | \psi_j(\mathbf{x}) \rangle_\pi = \delta_{ij}. \qquad (7.4)$$

As a result, any state function $\xi_t(\mathbf{x})$ can be represented within the eigenfunction basis $\xi_t(\mathbf{x}) = \sum_i \langle \psi_i(\mathbf{x}) | \xi_t(\mathbf{x}) \rangle_\pi \psi_i(\mathbf{x})$, and we can describe the dynamical evolution of ξ_t in increments of

τ via consecutive applications of $\mathcal{T}_\tau$ [444, 432],

$$\begin{aligned}
\xi_{t+k\tau}(\mathbf{x}) &= \underbrace{\mathcal{T}_\tau \circ \mathcal{T}_\tau \circ \ldots \circ \mathcal{T}_\tau}_{k} \circ \xi_t(\mathbf{x}) \\
&= \sum_i \langle \psi_i(\mathbf{x}) | \xi_t(\mathbf{x}) \rangle_\pi \mathcal{T}_\tau^k \circ \psi_i(\mathbf{x}) \\
&= \sum_i \langle \psi_i(\mathbf{x}) | \xi_t(\mathbf{x}) \rangle_\pi \lambda_i^k \psi_i(\mathbf{x}) \\
&= \sum_i \langle \psi_i(\mathbf{x}) | \xi_t(\mathbf{x}) \rangle_\pi e^{\left(-\frac{k\tau}{t_i}\right)} \psi_i(\mathbf{x}),
\end{aligned} \tag{7.5}$$

where we have transformed the eigenvalues into the implied time scales $t_i = -\tau / \ln(\lambda_i)$ associated with the relaxation of each of the orthonormal eigenfunctions in the expansion. These time scales t_i characterize the exponentially decaying contributions for the eigenfunctions $\{\psi_{i>0}(\mathbf{x})\}$, such that at sufficiently large lag times τ the contributions from basis associated with faster relaxation times can be neglected and ξ accurately approximated by keeping only the leading slowest terms. The leading pair ($\psi_0 = \mathbf{1}, \lambda_0 = 1$) therefore corresponds to the equilibrium distribution $\pi(\mathbf{x}) = \lim_{t\to\infty} q_t(\mathbf{x})$ and the remaining pairs to a hierarchy of increasingly quickly relaxing modes.

7.3.2 Approximating slow modes from data

The Variational Approach to Conformation dynamics (VAC) formalism provides a means to approximate the eigenfunctions $\{\psi_i(\mathbf{x})\}$ of the transfer operator $\mathcal{T}_\tau$. A variational principle underpins the VAC approach [443, 413, 444]. This can be straightforwardly understood by starting from an identity that for the i^{th} eigenfunction $\psi_i(\mathbf{x})$ that follows directly from Eqn. 7.4,

$$\frac{\langle \psi_i(\mathbf{x}) | \mathcal{T}_\tau \circ \psi_i(\mathbf{x}) \rangle_\pi}{\langle \psi_i(\mathbf{x}) | \psi_i(\mathbf{x}) \rangle_\pi} = \lambda_i. \tag{7.6}$$

Consider now a trial function $\nu(\mathbf{x})$ that is orthogonal to the preceding $(i-1)$ eigenfunctions,

$$\langle\psi_0(\mathbf{x})|\nu(\mathbf{x})\rangle_\pi = 0, \langle\psi_1(\mathbf{x})|\nu(\mathbf{x})\rangle_\pi = 0, \ldots, \langle\psi_{i-1}(\mathbf{x})|\nu(\mathbf{x})\rangle_\pi = 0, \tag{7.7}$$

and may therefore be expanded within the eigenfunction basis as,

$$\nu(\mathbf{x}) = \sum_{j\geq i}\langle\psi_j(\mathbf{x})|\nu(\mathbf{x})\rangle_\pi\psi_j(\mathbf{x}) = \sum_{j\geq i}\kappa_j\psi_j(\mathbf{x}), \tag{7.8}$$

where $\kappa_j = \langle\psi_j(\mathbf{x})|\nu(\mathbf{x})\rangle_\pi$ are the linear expansion coefficients. Combining Eqns. 7.4, 7.6, and 7.8, we obtain the following inequality,

$$\begin{aligned}
\tilde{\lambda}_i &= \frac{\langle\nu(\mathbf{x})|\mathcal{T}_\tau \circ \nu(\mathbf{x})\rangle_\pi}{\langle\nu(\mathbf{x})|\nu(\mathbf{x})\rangle_\pi} \\
&= \frac{\langle\sum_{j\geq i}\kappa_j\psi_j(\mathbf{x})|\mathcal{T}_\tau \circ \sum_{j\geq i}\kappa_j\psi_j(\mathbf{x})\rangle_\pi}{\langle\sum_{j\geq i}\kappa_j\psi_j(\mathbf{x})|\sum_{j\geq i}\kappa_j\psi_j(\mathbf{x})\rangle_\pi} \\
&= \frac{\sum_{j\geq i}\kappa_j^2\lambda_j}{\sum_{j\geq i}\kappa_j^2} \\
&\leq \frac{\sum_{j\geq i}\kappa_j^2\lambda_i}{\sum_{j\geq i}\kappa_j^2} \\
&= \lambda_i,
\end{aligned} \tag{7.9}$$

such that $\tilde{\lambda}_i$ is bounded from above by the true eigenvalue λ_i and the equality $\tilde{\lambda}_i = \lambda_i$ holds if and only if the trial function $\nu(\mathbf{x})$ is equal to the true eigenfunction $\nu(\mathbf{x}) = \psi_i(\mathbf{x})$. As such, we are guaranteed to approach λ_i from below by variational optimization of a trial function $\nu(\mathbf{x})$ to maximize $\tilde{\lambda}_i$ in Eqn. 7.9 subject to the orthogonality conditions in Eqn. 7.7.

It is convenient to construct trial functions as a linear expansion of arbitrary (possibly non-orthogonal) basis functions $\{\chi_i(\mathbf{x})\}$ and optimize the expansion coefficients within this

basis to develop optimal approximations $\{\phi_i(\mathbf{x})\}$ to the true eigenfunctions $\{\psi_i(\mathbf{x})\}$,

$$\phi_i(\mathbf{x}) = \sum_j \alpha_{ij}\chi_j(\mathbf{x}), \tag{7.10}$$

where α_{ij} are the linear expansion coefficients for basis function j within our approximation for eigenfunction i. Under this linear formulation, the VAC procedure can be shown to yield the following generalized eigenvalue problem [413],

$$C^\tau \boldsymbol{\alpha}_i = \tilde{\lambda}_i S \boldsymbol{\alpha}_i, \tag{7.11}$$

where $\boldsymbol{\alpha}_i = \{\alpha_{ij}\} = [\alpha_{i0}, \alpha_{i1}, \alpha_{i2}, \ldots]$ are the linear expansion coefficients for assembling the approximate eigenfunction $\phi_i(\mathbf{x})$ with associated approximate eigenvalue $\tilde{\lambda}_i$. This generalized eigenvalue problem is isomorphic to the Roothaan-Hall equations in quantum mechanics that are used to represent the Hartree-Fock equation in a non-orthonormal basis set [459]. The matrix elements of C^τ and S are defined as,

$$C^\tau_{jk} = \langle \chi_j(\mathbf{x}) | \mathcal{T}_\tau \circ \chi_k(\mathbf{x}) \rangle_\pi, \tag{7.12}$$

$$S_{jk} = \langle \chi_j(\mathbf{x}) | \chi_k(\mathbf{x}) \rangle_\pi, \tag{7.13}$$

where C^τ is the time-lagged correlation matrix – analogous to the Fock matrix – and S the overlap matrix. Standard techniques [460] can be employed to solve the generalized eigenvalue problem in Eqn. 7.11, which will yield the linear expansion coefficients for the optimal approximate eigenfunctions $\{\phi_i(\mathbf{x})\}$ within the basis $\{\chi_i(\mathbf{x})\}$.

Using this VAC formalism we can numerically estimate $\{\phi_i(\mathbf{x})\}$ from trajectory data by

replacing the exact inner products in Eqns. 7.12 and 7.13 with empirical estimates,

$$\begin{aligned} C_{jk}^{\tau} = \langle \chi_j(\mathbf{x}) | \mathcal{T}_\tau \circ \chi_k(\mathbf{x}) \rangle_\pi &\approx \mathbb{E}[\chi_j(\mathbf{x_t}) \chi_k(\mathbf{x_{t+\tau}})] \\ &\approx \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \chi_j(\mathbf{x}_t) \chi_k(\mathbf{x}_{t+\tau}) \\ &\approx \frac{1}{L} \sum_{i=1}^{L} \chi_j(\mathbf{x}_t^{(i)}) \chi_k(\mathbf{x}_{t+\tau}^{(i)}), \end{aligned} \tag{7.14}$$

$$\begin{aligned} S_{jk} = \langle \chi_j(\mathbf{x}) | \chi_k(\mathbf{x}) \rangle_\pi &\approx \mathbb{E}[\chi_j(\mathbf{x_t}) \chi_k(\mathbf{x_t})] \\ &\approx \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \chi_j(\mathbf{x}_t) \chi_k(\mathbf{x}_t) \\ &\approx \frac{1}{L} \sum_{i=1}^{L} \chi_j(\mathbf{x}_t^{(i)}) \chi_k(\mathbf{x}_t^{(i)}), \end{aligned} \tag{7.15}$$

where the terminal summations defining the estimators in Eqns. 7.14 and 7.15 are carried out over $i = 1 \dots L$ training samples harvested from the trajectory data assembled as pairs of states $(\mathbf{x}_t^{(i)}, \mathbf{x}_{t+\tau}^{(i)})$ separated by a lag time τ.

7.3.3 State-free Reversible VAMPnets (SRVs)

The quality of the eigenfunction approximations under the VAC is contingent on the choice of a sufficiently expressive and powerful basis. The Markov state model (MSM) formalism [461, 462] follows as a special case of the VAC under a choice of basis corresponding to indicator functions partitioning of the configurational phase space into n_c disjoint sets $\{S_j\}_{j=1}^{n_c}$ such that $\chi_j(\mathbf{x}) = 1$ for $\mathbf{x} \in S_j$ and $\chi_j(\mathbf{x}) = 0$ for $\mathbf{x} \notin S_j$ [440]. In this case estimation of the matrix elements C_{jk}^{τ} simply amounts to counting transitions from trajectory data that carry states $\mathbf{x}_t \in S_j$ to $\mathbf{x}_{t+\tau} \in S_k$ in a lag time τ. The inherent orthonormality and binary nature of the indicator function basis $\{\chi_i\}$ implies that the matrix $\mathbf{S}$ is diagonal

with matrix elements $S_{jj} = \pi_j$ corresponding to the stationary distribution of normalized state counts $\pi_j = \frac{\sum_{t=1}^{T-\tau} \chi_j(\mathbf{x}_t)}{T-\tau}$ [413]. The elements of the approximate MSM eigenfunctions $\{\phi_i(\mathbf{x})\}$ can then be interpreted as fluxes into or out of sets $\{S_j\}$ with each eigenfunction $\phi_i(\mathbf{x})$ characterizing slow modes as traversals along the discretized state-space $\{S_j\}$ [463]. The quality of state-space discretization is a critical determinant of MSM precision, and is typically most amenable to metastable dynamics that can be effectively discretized into long-lived metastable states [464, 413]. State-space clustering to determine disjoints sets that define the basis $\{\chi_i\}$ is commonly performed within low-dimensional embeddings of the full-dimensional state-space using techniques such as TICA [441, 421, 442] combined with k-means clustering. The different interacting components within MSM workflows typically suffer from requiring manual trial-and-error and human expertise to handle selecting the appropriate dimensionality reduction technique, the number of cluster states, choice of clustering method, and optimizing location of cluster centers [413]. Mardt, Pasquali, Wu, and Noé developed VAMPnets as an end-to-end replacement for MSM construction pipeline using deep neural networks to integrate the featuruzation, dimensionality reduction, clustering, and kinetic modeling steps [465]. Indicator functions represent one rather restrictive choice of basis and the quality of the approximate eigenfunction expansions depends sensitively upon the choice of phase space discretization. More general classes of basis functions offer the possibility of superior eigenfunction approximations, but it is challenging to intuit good choices of a basis set.

We previously introduced State free Reversible VAMPNets (SRVs) as a means to perform data-driven learning of good basis functions for the VAC by performing simultaneous learning of the basis functions and the linear expansion coefficients for approximation of the transfer operator eigenfunctions [444]. In brief, we learn $\{\chi_i\}$ from training trajectories using simple feedforward neural networks, pass these basis functions through the VAC to learn approximations for the leading transfer operator eigenfunctions $\{\phi_i(\mathbf{x})\}$, and quantify

the quality of the learned eigenfunctions via a loss function based on the slowness of the associated implied time scales. We then iteratively update the neural networks to learn superior basis functions by backpropagating the gradient of the loss function through the entire computational graph.

Mathematically, the SRV neural networks $f^{\theta}(\mathbf{x}) \in \mathbb{R}^m$ are parameterized with model weights θ and tasked with learning a map from microstate configurations $\mathbf{x}$ into a m-dimensional output space, where m is the number of desired approximate eigenfunctions of the transfer operator (i.e., slow CVs) to learn. The neural network outputs $\chi_j(\mathbf{x}) = f_j^{\theta}(\mathbf{x})$ represent the learned basis functions, which we pass to the VAC and solve for the optimal linear expansion approximations of the transfer operator eigenfunctions $\phi_i(\mathbf{x}) = \sum_j \alpha_{ij}\chi_j(\mathbf{x})$ using Eqns. 7.11-7.15. The model weights θ are optimized via minimization of the VAMP-r loss function,

$$\mathcal{L}_{SRV} = -\sum_{j=0}^{m-1} \tilde{\lambda}_j^r, \tag{7.16}$$

with a typical choice of $r = 2$ corresponding to the VAMP-2 loss that can be interpreted as the cumulative kinetic variance [463, 423]. The loss function attains a minimum when the approximate eigenfunctions equal the true eigenfunctions and the approximation becomes exact [465].

Learning iterates between solving for the expansion coefficients $\{\alpha_{ij}\}$ and eigenvalues $\{\tilde{\lambda}_j\}$ by solving the generalized eigenvalue problem in Eqn. 7.11 and updating the network weights θ by gradient descent steps of $\mathcal{L}_{SRV}$ in Eqn. 7.16. The generalized eigenvalue problem is formulated by estimating the matrices C^{τ} and S from the neural network outputs in batches according to Eqns. 7.14 and 7.15, and solved using Cholesky decomposition to provide stable gradients for the learning problem [444]. The neural network weights are updated by mini-batch gradient descent with the Adam [217] optimizer. Neural network convergence is assessed by monitoring and ensuring plateauing of the loss $\mathcal{L}_{SRV}$ on a held-out validation set, which together with providing randomly shuffled training samples, also

serves as a regularizer to also help prevent overfitting.

The trained SRV may then be used for out of sample inference on unseen configurations $\mathbf{x}^*$ to furnish the mappings into the approximate eigenfunctions $\mathbf{x}^* \rightarrow \{\phi_i(\mathbf{x}^*)\}_{i=0}^{m-1}$. This learned projection represents a data-driven embedding into a slow CV subspace spanned by our approximations to the leading eigenfunctions of the transfer operator. An appropriate number m of CVs to retain is typically informed by a gap in the eigenvalue spectrum separating the slowly relaxing modes from a rapidly relaxing continuum of fast modes that are effectively annealed to the slow subspace on sufficiently long time scales. An appropriate lag time τ is determined by convergence of the implied time scales of the learned slow modes and satisfaction of the Chapman-Kolmogorov test [444]. Full details of the architecture, training, and hyperparameter tuning of the SRV models employed in this work are provided in the Supporting Information.

7.3.4 Dynamical reweighting

Dynamical reweighting provides a formalism to estimate equilibrium dynamical properties from biased simulation data. Dynamical reweighting corrections for biased data have been analytically derived by Chodera and co-workers for the special case of temperature reweighting [447, 82, 466], and Keller and co-workers for arbitrary biases [109, 110, 111]. The premise of dynamical reweighting lies in path reweighting methods, based mathematically on the Girsanov theorem [108], in which phase space trajectories
$\boldsymbol{\omega}_{0\rightarrow\tau} = ((\mathbf{x}_0, \mathbf{v}_0), (\mathbf{x}_{\Delta t}, \mathbf{v}_{\Delta t}), \ldots, (\mathbf{x}_{\tau-\Delta t}, \mathbf{v}_{\tau-\Delta t}), (\mathbf{x}_\tau, \mathbf{v}_\tau))$ generated under some simulation potential $V_{sim}(\mathbf{x})$ are associated with a path probability density $\mu^{sim}_{\Omega_{\tau,\mathbf{x}_0}}(\boldsymbol{\omega}_{0\rightarrow\tau})$, where $\Omega_{\tau,\mathbf{x}_0}$ is the path space comprising the set of all paths of length τ that start at $\mathbf{x}_0$. Positing that the simulation potential $V_{sim}(\mathbf{x}) = V_{target}(\mathbf{x}) - U_{bias}(\mathbf{x})$ is a superposition of the underlying target potential $V_{target}(\mathbf{x})$ and a biasing potential $U_{bias}(\mathbf{x})$, path weights $M(\boldsymbol{\omega}_{0\rightarrow\tau})$ can be used to recover the probabilities for paths collected under the simulation potential

V_{sim} by reweighting to the target potential V_{target},

$$\mu^{target}_{\Omega_{\tau,\mathbf{x}_0}}(\boldsymbol{\omega}_{0\to\tau}) = M(\boldsymbol{\omega}_{0\to\tau})\mu^{sim}_{\Omega_{\tau,\mathbf{x}_0}}(\boldsymbol{\omega}_{0\to\tau}). \tag{7.17}$$

Until recently, analytical expressions of the path weights $M(\boldsymbol{\omega}_{0\to\tau})$ for arbitrary biasing potentials were only available for the Euler-Maruyama scheme for overdamped Langevin dynamics (i.e., Brownian dynamics) [467, 468, 469, 470, 111]. More recently, analytical path weights have been derived for (underdamped) Langevin dynamics schemes [111] which are known to better reproduce the true dynamics for large molecular systems and are widely implemented within existing molecular dynamics frameworks [471, 472, 473, 453, 474, 475, 476]. The molecular systems in this work are simulated in implicit solvent under (underdamped) Langevin dynamics that are numerically integrated using a variant of the Langevin leapfrog scheme developed by Izaguirre, Sweet, and Pande [453], hereafter referred to as the ISP scheme, for which analytical path weights have been derived by Kieninger and Keller [111]. These authors have also implemented this integrator within OpenMM [406] along with numerically efficient implementations for on-the-fly calculation of the path weights $M(\boldsymbol{\omega}_{0\to\tau})$ during runtime.

Co-opting this dynamical reweighting formalism within the SRV framework enables estimation of equilibrium slow modes from biased simulation data [109, 110]. Following the development of Donati and Keller [110], we observe that the matrix elements C^{τ}_{jk} of the time-lagged correlation matrix of basis functions $\{\chi_i(\mathbf{x})\}$ under the target potential $V_{target}(\mathbf{x})$ can be viewed as a nested integral of a path ensemble average over paths $\boldsymbol{\omega}_{t\to t+\tau} \in \Omega_{\tau,\mathbf{x}_t}$ inside a phase space ensemble average over path starting configurations $\mathbf{x}_t \in \Gamma$,

$$C^{\tau}_{jk} = \int_{\Gamma} \mu^{target}_{\Gamma}(\mathbf{x}_t)\chi_j(\mathbf{x}_t) \int_{\Omega_{\tau,\mathbf{x}_t}} \mu^{target}_{\Omega_{\tau,\mathbf{x}_t}}(\boldsymbol{\omega}_{t\to t+\tau})\chi_k(\mathbf{x}_{t+\tau}) d\boldsymbol{\omega}_{t\to t+\tau} d\mathbf{x}_t, \tag{7.18}$$

where $\mu^{target}_{\Gamma}(\mathbf{x}_t)$ and $\mu^{target}_{\Omega_{\tau,\mathbf{x}_t}}(\boldsymbol{\omega}_{t\to t+\tau})$ are, respectively, the phase-space and path-space

probability densities under $V_{target}(\mathbf{x})$. For simulation data collected under a potential $V_{sim}(\mathbf{x}) = V_{target}(\mathbf{x}) - U_{bias}(\mathbf{x})$, we are compelled to apply thermodynamic and dynamical corrections to estimate C_{jk}^{τ} under $V_{target}(\mathbf{x})$,

$$C_{jk}^{\tau} = \int_{\Gamma} h(\mathbf{x}_t)\mu_{\Gamma}^{sim}(\mathbf{x}_t)\chi_j(\mathbf{x}_t) \int_{\Omega_{\tau,\mathbf{x}_t}} M(\boldsymbol{\omega}_{t\to t+\tau})\mu_{\Omega_{\tau,\mathbf{x}_t}}^{sim}(\boldsymbol{\omega}_{t\to t+\tau})\chi_k(\mathbf{x}_{t+\tau})d\boldsymbol{\omega}_{t\to t+\tau}d\mathbf{x}_t. \tag{7.19}$$

The thermodynamic reweighting correction accounts for the change in the configurational probability of the initial configuration of the path $\mathbf{x}_t$ under V_{target} relative to V_{sim},

$$\begin{aligned} h(\mathbf{x}_t) &= \frac{\mu_{\Gamma}^{target}(\mathbf{x}_t)}{\mu_{\Gamma}^{sim}(\mathbf{x}_t)} \\ &= \frac{\exp\left(-\beta V_{target}(\mathbf{x}_t)\right)}{Z_{target}} \Big/ \frac{\exp\left(-\beta V_{sim}(\mathbf{x}_t)\right)}{Z_{sim}} \\ &= \frac{Z_{sim}}{Z_{target}} \exp\left(-\beta U_{bias}(\mathbf{x}_t)\right), \end{aligned} \tag{7.20}$$

where we operate in the canonical ensemble such that $Z = \int_{\Gamma} \exp\left(-\beta V(\mathbf{x})\right) d\mathbf{x}$ is the configurational part of the canonical partition function associated with a potential $V(\mathbf{x})$, and $\beta{=}(k_B T)^{-1}$, where k_B is the Boltzmann constant, and T is the temperature. The dynamical reweighting correction accounts for the change in the probability of the path under V_{target} relative to V_{sim},

$$M(\omega_{t\to t+\tau}) = \frac{\mu_{\Omega_{\tau,\mathbf{x}_t}}^{target}(\boldsymbol{\omega}_{t\to t+\tau})}{\mu_{\Omega_{\tau,\mathbf{x}_t}}^{sim}(\boldsymbol{\omega}_{t\to t+\tau})}. \tag{7.21}$$

For underdamped Langevin dynamics numerically integrated using the ISP scheme [453], an explicit expression for this dynamical reweighting factor is known (c.f., Eqns. 23, 57, 58, 59

and 60 in Ref. [111]) [109, 110, 111],

$$
\begin{aligned}
&M(\boldsymbol{\omega}_{t\rightarrow t+\tau}) = \\
&\quad \exp\left(-\sum_{\alpha=1}^{N} \sum_{\gamma=\{x,y,z\}} \sum_{k=0}^{\tau/\Delta t-1} \right. \\
&\quad \left. \frac{\left(x_{t+(k+1)\Delta t}^{[\alpha,\gamma]} - x_{t+k\Delta t}^{[\alpha,\gamma]}\right)\left(\nabla_{[\alpha,\gamma]} V_{target}\left(\mathbf{x}_{t+k\Delta t}\right) - \nabla_{[\alpha,\gamma]} V_{sim}\left(\mathbf{x}_{t+k\Delta t}\right)\right)}{k_B T \xi (1+\exp(-\xi\Delta t))\Delta t} \right) \\
&\quad \cdot \exp\left(\sum_{\alpha=1}^{N} \sum_{\gamma=\{x,y,z\}} \sum_{k=0}^{\tau/\Delta t-1} \frac{v_{t+k\Delta t}^{[\alpha,\gamma]}\left(\nabla_{[\alpha,\gamma]} V_{target}\left(\mathbf{x}_{t+k\Delta t}\right) - \nabla_{[\alpha,\gamma]} V_{sim}\left(\mathbf{x}_{t+k\Delta t}\right)\right)}{k_B T \xi (1+\exp(\xi\Delta t))} \right) \\
&\quad \cdot \exp\left(-\frac{\exp(\xi\Delta t)-1}{\exp(\xi\Delta t)+1} \cdot \sum_{\alpha=1}^{N} \sum_{\gamma=\{x,y,z\}} \sum_{k=0}^{\tau/\Delta t-1} \right. \\
&\quad \left. \frac{\left(\left(\nabla_{[\alpha,\gamma]} V_{target}\left(\mathbf{x}_{t+k\Delta t}\right)\right)^2 - \left(\nabla_{[\alpha,\gamma]} V_{sim}\left(\mathbf{x}_{t+k\Delta t}\right)\right)^2\right)}{2 k_B T \xi^2 m_{[\alpha]}} \right),
\end{aligned}
\tag{7.22}
$$

where the summation $\alpha = 1 \ldots N$ is carried over the N particles in the system, the summation $\gamma = \{x, y, z\}$ over the three spatial dimensions, $x_t^{[\alpha,\gamma]}$ and $v_t^{[\alpha,\gamma]}$ represent the γ components of the position and velocity of particle α, $\nabla_{[\alpha,\gamma]}$ is the partial derivative taken with respect to the γ spatial component of particle α, $m_{[\alpha]}$ is the mass of particle α, ξ is the Langevin friction coefficient, Δt the integration time step, and the lag time τ is restricted to be an integer multiple of Δt. This equation may be equivalently rewritten in terms of the bias potential U_{bias} and the random number sequence that generated the Langevin integration path $\{\eta_{t+k\Delta t}\}_{k=0}^{\tau/\Delta t-1}$ to present a simpler and potentially less computationally expensive

expression to evaluate (c.f., Eqn. 27 in Ref. [111]) [109, 110, 111],

$$M(\boldsymbol{\omega}_{t\rightarrow t+\tau}) = \exp\left(-\frac{1-\exp(-\xi\Delta t)}{\sqrt{1-\exp(-2\xi\Delta t)}}\cdot\sum_{\alpha=1}^{N}\sum_{\gamma=\{x,y,z\}}\sum_{k=0}^{\tau/\Delta t-1}\frac{\nabla_{[\alpha,\gamma]}U_{bias}\left(\mathbf{x}_{t+k\Delta t}\right)\eta_{t+k\Delta t}}{\sqrt{k_BT\xi^2m_{[\alpha]}}}\right) \cdot\exp\left(-\frac{(1-\exp(-\xi\Delta t))^2}{1-\exp(-2\xi\Delta t)}\cdot\sum_{\alpha=1}^{N}\sum_{\gamma=\{x,y,z\}}\sum_{k=0}^{\tau/\Delta t-1}\frac{\left(\nabla_{[\alpha,\gamma]}U_{bias}\left(\mathbf{x}_{t+k\Delta t}\right)\right)^2}{2k_BT\xi^2m_{[\alpha]}}\right). \quad (7.23)$$

Having defined the reweighting factors, we return to Eqn. 7.19 and estimate it empirically from trajectories collected under the simulation potential $V_{sim} = V_{target} - U_{bias}$ from which we compile a batch of $i = 1 \ldots L$ training samples comprising pairs of states $(\mathbf{x}_t^{(i)}, \mathbf{x}_{t+\tau}^{(i)})$ separated within a simulation trajectory by a lag time τ and connected by paths $\boldsymbol{\omega}_{t\rightarrow t+\tau}^{(i)} = \Big((\mathbf{x}_t^{(i)}, \mathbf{v}_t^{(i)}), (\mathbf{x}_{t+\Delta t}^{(i)}, \mathbf{v}_{t+\Delta t}^{(i)}), \ldots, (\mathbf{x}_{t+\tau-\Delta t}^{(i)}, \mathbf{v}_{t+\tau-\Delta t}^{(i)}), (\mathbf{x}_{t+\tau}^{(i)}, \mathbf{v}_{t+\tau}^{(i)})\Big)$,

$$C_{jk}^{\tau} \approx \frac{1}{L}\sum_{i=1}^{L} h(\mathbf{x}_t^{(i)}) M(\omega_{t\rightarrow t+\tau}^{(i)}) \chi_j(\mathbf{x}_t^{(i)}) \chi_k(\mathbf{x}_{t+\tau}^{(i)}). \quad (7.24)$$

As a corollary, the matrix elements S_{jk} of the instantaneous correlation matrix can be estimated over the same data as,

$$S_{jk} \approx \frac{1}{L}\sum_{i=1}^{L} h(\mathbf{x}_t^{(i)}) M(\omega_{t\rightarrow t+\tau}^{(i)}) \chi_j(\mathbf{x}_t^{(i)}) \chi_k(\mathbf{x}_t^{(i)}). \quad (7.25)$$

The matrices C^{τ} and S are subsequently used to solve for the basis function expansion coefficients via the generalized eigenvalue problem defined in Eqn. 7.11. The thermodynamic reweighting factor $h(\mathbf{x}_t)$ (Eqn. 7.20) possesses a term $\left(\frac{Z_{sim}}{Z_{target}}\right)$ that serves as a multiplicative constant of each element of C^{τ} and S that is independent of state $\mathbf{x}_t$ and path $\omega_{t\rightarrow t+\tau}^{(i)}$ and may therefore be canceled out on each side of the generalized eigenvalue problem. As such,

this partition function ratio never need be computed and can be dropped. Defining,

$$g(\mathbf{x}_t) = \exp\left(-\beta U_{bias}(\mathbf{x}_t)\right), \tag{7.26}$$

we can rewrite $h(\mathbf{x}_t) = \left(\frac{Z_{sim}}{Z_{target}}\right) g(\mathbf{x}_t)$ and formally drop the partition function ratio by replacing $h(\mathbf{x}_t)$ with $g(\mathbf{x}_t)$ in Eqns. 7.24 and 7.25 to yield,

$$C_{jk}^{\tau} \approx \frac{1}{L} \sum_{i=1}^{L} g(\mathbf{x}_t^{(i)}) M(\omega_{t\to t+\tau}^{(i)}) \chi_j(\mathbf{x}_t^{(i)}) \chi_k(\mathbf{x}_{t+\tau}^{(i)}), \tag{7.27}$$

$$S_{jk} \approx \frac{1}{L} \sum_{i=1}^{L} g(\mathbf{x}_t^{(i)}) M(\omega_{t\to t+\tau}^{(i)}) \chi_j(\mathbf{x}_t^{(i)}) \chi_k(\mathbf{x}_t^{(i)}). \tag{7.28}$$

Finally, we define a combined thermodynamic and dynamic reweighting factor,

$$W(\boldsymbol{\omega}_{t\to t+\tau}^{(i)}) = g(\mathbf{x}_t^{(i)}) \times M(\boldsymbol{\omega}_{t\to t+\tau}^{(i)}), \tag{7.29}$$

and explicitly replace the basis functions $\{\chi_i(\mathbf{x})\}$ with the SRV neural network outputs $\chi_j(\mathbf{x}) = f_j^{\theta}(\mathbf{x})$ to obtain our final expressions for C_{jk}^{τ} and S_{jk} that we use in practice,

$$C_{jk}^{\tau} \approx \frac{1}{L} \sum_{i=1}^{L} W(\boldsymbol{\omega}_{t\to t+\tau}^{(i)}) f_j^{\theta}(\mathbf{x}_t^{(i)}) f_k^{\theta}(\mathbf{x}_{t+\tau}^{(i)}), \tag{7.30}$$

$$S_{jk} \approx \frac{1}{L} \sum_{i=1}^{L} W(\boldsymbol{\omega}_{t\to t+\tau}^{(i)}) f_j^{\theta}(\mathbf{x}_t^{(i)}) f_k^{\theta}(\mathbf{x}_t^{(i)}). \tag{7.31}$$

In the absence of bias (i.e., $U_{bias}(\mathbf{x}) = 0 \Rightarrow V_{sim}(\mathbf{x}) = V_{target}(\mathbf{x})$) the reweighting factors in Eqns. 7.26 and 7.23 both reduce to unity (i.e., $g(\mathbf{x}_t) = M(\boldsymbol{\omega}_{t\to t+\tau}) = 1 \Rightarrow W(\boldsymbol{\omega}_{t\to t+\tau}) = 1$) and the expressions for C_{jk}^{τ} and S_{jk} for biased simulation data (Eqns. 7.30 and 7.31) reduce to those for unbiased data (Eqns. 7.14 and 7.15) as they should.

Having adapted the SRV framework incorporating dynamical reweighting, training can now proceed to approximate the slow modes of a molecular system from short, discon-

tinuous, and biased trajectories. Since the learned modes will correspond only to the slowest modes present within the training data, the true slow modes for systems containing high free energy barriers and rugged free energy landscapes will not be well approximated within unbiased simulation trajectories that have a propensity to be kinetically trapped. This is the primary motivation for the GREST approach presented herein that performs interleaved rounds of slow CV discovery and enhanced sampling until convergence is achieved in the explored phase space and the learned CVs. A Python package containing a PyTorch [276] implementation of SRVs with capabilities for training with biased simulation data incorporating these reweighting factors is publicly available via GitHub at `https://github.com/andrewlferguson/snrv`.

7.3.5 GREST adaptive sampling workflow

The Girsanov Reweighting Enhanced Sampling Technique (GREST) is an adaptive sampling workflow combines slow CV discovery with SRVs and Girsanov path reweighting to perform simultaneous learning of the slow modes of a molecular system and efficient exploration of configurational space. This is accomplished by interleaving rounds of slow CV estimation with path reweighted SRVs and enhanced sampling along the discovered slow modes. Our approach is procedurally similar to MESA [404] but is designed to discover and accelerate slow rather than high variance CVs. It employs the path reweighting developed by Keller and co-workers to estimate dynamical observables from biased simulation trajectories [109, 110], but uses this approach to iteratively discover slow CVs as opposed to construct MSMs from biased simulations conducted in pre-defined CVs that was the focus of work by both Keller and co-workers [109, 110] and Weber and Pande [112]. The first iterative round of GREST shares methodological similarities with the approach reported in Ref. [426] but we apply both dynamical and thermodynamic corrections to rigorously reweight our path ensemble estimators of the slow CVs and perform successive rounds of discovery and biasing to converge

the CVs and exploration of configurational phase space. A schematic illustration of the four-step iterative workflow is presented in Fig. 7.1.

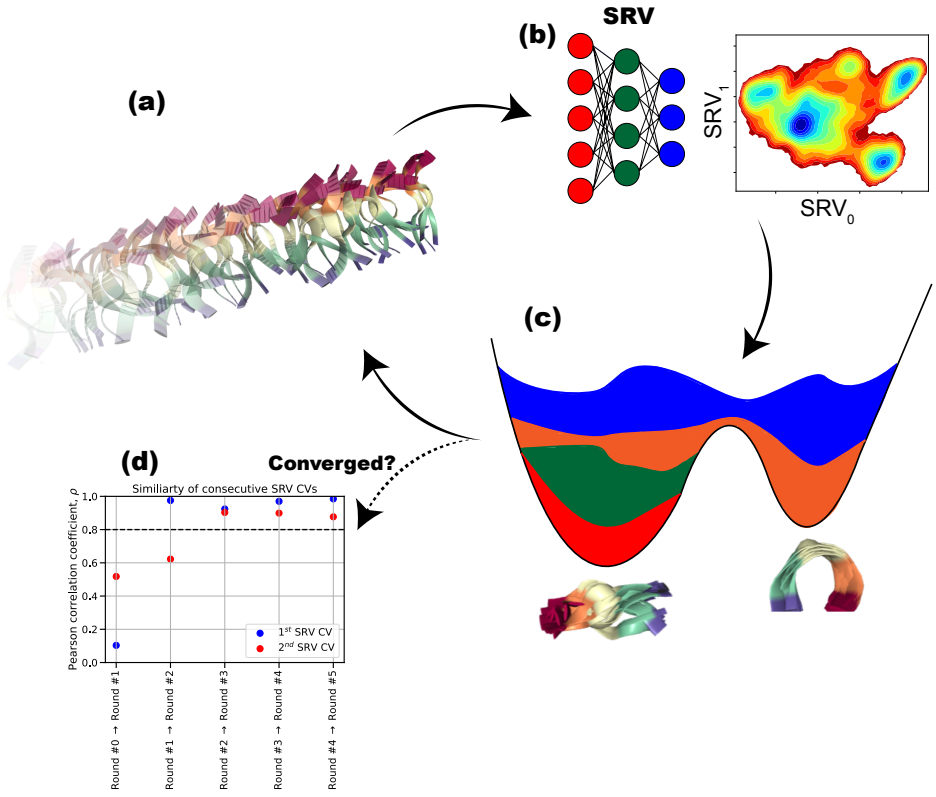


Figure 7.1: Schematic illustration of four-step iterative Girsanov Reweighting Enhanced Sampling Technique (GREST) workflow. **(a)** Initial training data is generated by unbiased or biased simulation data accelerated along intuitive CVs. **(b)** Slow CVs are estimated from the training data using Girsanov path reweighted State-free Reversible VAMPnets (SRVs). **(c)** Enhanced sampling calculations are performed within the learned slow CVs using a CV biasing technique such as metadynamics. **(d)** Convergence of the iterative cycle is assessed by evaluating the stability of the learned CVs in successive iterations.

Step 0: Generation of initial training data. Conduct an initial simulation to gather trajectory data for the first round of slow CV estimation. Slow CV estimation with SRVs is ultimately a data-driven technique, and therefore generally benefits from training on larger datasets that explore larger volumes of the configuration space. As this initial simulation serves as only a starting point for the adaptive sampling protocol, the length of the trajectory can judiciously be chosen by considering available computational resources. For larger,

complex systems with rugged free energy surfaces susceptible to kinetic trapping, an unbiased simulation may be insufficient to provide adequate exploration of local configurational space and a biased simulation in intuited CVs based on expert knowledge may be conducted to provide more extensive configurational space coverage within this initial training dataset.

Step 1: Slow CV estimation using SRVs. Train path reweighted SRV models using unbiased or biased simulation data to estimate slow CVs. The SRV learns a mapping from configuration space coordinates $\mathbf{x}$ to approximate slow eigenfunctions of the transfer operator $\mathbf{x} \rightarrow \{\phi_i(\mathbf{x})\}$. Since the dimensionality of the slow subspace may change over the course of the GREST iterative cycles as a result of the region and volume of configurational phase space that has been explored by the simulation trajectories, we recommend first training SRVs with a relatively high number of CVs (e.g., $n \sim 10$) to attempt to discern a spectral gap in the slow mode eigenvalue spectrum and make a principled inference of the dimensionality of the slow subspace. The only exception to this recommendation is for the first round of sampling where exploration may be poor and we have found that it can be useful to simply choose a 2D latent space for the purposes of driving the first round of biased sampling.

Step 2: Enhanced sampling along slow CVs. Perform enhanced sampling within the SRV-determined CVs. The slow CVs learned by the SRV can be passed to any CV enhanced sampling technique and used to drive accelerated sampling in an off-the-shelf manner. For the molecular systems in this work, we choose to employ well-tempered metadynamics (WTMetaD) [398, 477]. Converging WTMetaD can become computationally burdensome when sampling in more than two CVs, and in this case we recommend well-tempered parallel bias metadynamics (PBMetaD) as an approach that scales well to large numbers of CVs [478, 479]. The accumulation of bias throughout the metadynamics run can lead to large variances in the estimator of the dynamical path weights [110] and produce numerical instabilities during SRV training. As suggested in Refs. [110, 112], we instead converge the metadynamics bias U_{meta} and construct a potential $V_{sim} = V_{target} - U_{bias} = V_{target} - \gamma U_{meta}$

under which we conduct an additional round of metadynamics enhanced sampling in which we do not update the terminal bias potential and modulate its strength by a factor $\gamma = [0, 1]$. This constitutes enhanced sampling under an attenuated bias potential that eliminates the need to track the time varying nature of the bias and also enables us to judiciously tune the aggressiveness of the applied bias via the attenuation factor γ. Consistent with previous observations [110], we have found that selecting an attenuation factor of $\gamma \approx 0.05 - 0.2$ tends to provide a good balance between enhancing sampling of configurational space while moderating the magnitude of the path weights to stabilize low-variance numerical estimation of path ensemble averages. In this work, we adopt a relatively conservative attenuation factor of $\gamma = 0.05$. This attenuation of the applied bias does limit the barrier crossing accelerations achievable by the technique. The exponential nature of barrier crossing dynamics, however, means that significant accelerations can be achieved even by relatively modest reductions in barrier heights. For example, assuming a simple Arrhenius dependence $k = A \exp\left(-\frac{E_a}{k_B T}\right)$, where k is the rate constant, A is the pre-exponential factor, E_a is the activation energy, k_B is the Boltzmann constant, and T is the temperature, a modest 5% decrease of a E_a=10 $k_B T$ barrier leads to a 65% elevation of the rate constant.

Step 3: Evaluating convergence. Assess the convergence of our adaptive sampling workflow by determining if the learned SRV CVs have stabilized. Using all accumulated enhanced sampling data, we evaluate the Pearson correlation coefficient of the learned SRV CVs between consecutive enhanced sampling rounds. Plateauing of this metric suggests the slow modes identified with SRVs trained on our enhanced sampling data have approximately converged. Based on this metric we can evaluate if CV learning has halted and terminate our workflow. Otherwise, we perform another iterative round of GREST.

7.3.6 Simulation settings and parameters

One-dimensional four-well potential

Simulations under (underdamped) Langevin dynamics are performed for a single particle moving on the 1D 4-well potential $V_{4-well}(x) = 2x^8 + 1.6\exp(-80x^2) + 0.4\exp(-80(x - 0.5)^2) + \exp(-40(x+0.5)^2)$ introduced by Schwantes and Pande [443]. The ISP scheme [111] – a full-step adaptation of the Langevin leapfrog algorithm developed by Izaguirre, Sweet, and Pande [453] – is used to integrate the dynamics forward in time. Simulations are performed using a time step of 0.01 s for a 1 kg particle with a friction coefficient of $\xi = 1$ s^{-1} at a temperature of 60 K. Configurations are saved for analysis at a period of 0.05 s.

Alanine dipeptide

Simulations of alanine dipeptide (ADP) are carried out at a temperature of 300 K in implicit solvent using OpenMM [406] with the AMBER99SB-ILDN [259] force field and the GBSA-OBC [480] solvation model. Nonbonded interactions were cutoff at distances beyond 1 nm. The system is numerically integrated under (underdamped) Langevin dynamics using the ISP scheme [453] for which Keller and co-workers developed analytical expressions and efficient numerical calculation schemes for the Girsanov path weights [109, 110, 111]. A simulation time step of 2 fs is used with a friction coefficient of $\xi = 10$ ps^{-1}. Configurations are saved for analysis at a period of 0.1 ps.

WLALL

Simulations of the Trp-Leu-Ala-Leu-Leu (WLALL) pentapeptide are conducted at a temperature of 300 K using OpenMM [406] and the CHARMM22 [481, 482] force field with the GBSA-OBC [480] implicit solvent model. All non-bonded interactions are cutoff at 1 nm. The system is numerically integrated under (underdamped) Langevin dynamics using the

ISP scheme [453] and path weights calculated using the approach developed by Keller and co-workers [109, 110, 111]. A simulation time step of 2 fs is used with a friction coefficient of $\xi = 10$ ps^{-1}. Configurations are saved for analysis at a period of 0.1 ps.

7.4 Results

7.4.1 One-dimensional four-well potential

As a first illustrative toy example, we consider a single particle moving on the 1D 4-well potential (Fig. 8.1a) [443]. We model single particle diffusion over this landscape under (underdamped) Langevin dynamics. This system is sufficiently simple that the four wells and three barrier hopping transitions can be comprehensively explored under long unbiased simulations to provide a gold standard baseline against which to benchmark GREST [444, 443]. We represent the system state of this simple 1D system to our path reweighted SRVs by simply furnishing the 1D coordinate $\mathbf{x} = x \in \mathbb{R}^1$. In contrast to linear algorithms such as TICA [441, 421, 442], the SRV deep learning architecture can learn nonlinear transformations of the $\mathbf{x}$ coordinate to learn a hierarchy of slowly relaxing CVs. Full details of the simulations are provided in the Supporting Information.

Pre-defined bias. We first demonstrate the capacity of the dynamically reweighted SRVs to learn the correct slow modes and associated time scales under a pre-defined bias. This can be conceived of as a first round of a GREST campaign in which the user applies an intuitive biasing potential to accelerate sampling of the system prior to any data-driven learning of the slow modes. The 4-well potential serves as the target potential under which we wish to estimate the slow modes $V_{target} = V_{4-well} = 2x^8 + 1.6\exp(-80x^2) + 0.4\exp(-80(x - 0.5)^2) + \exp(-40(x + 0.5)^2)$ but we perform simulations under a 2-well potential $V_{sim} = V_{2-well} = 2x^8 + 1.75\exp(-80x^2)$ (Fig. 8.1a). Formally, since the simulation and target potentials are related via the applied bias as $V_{sim} = V_{target} - U_{bias}$, this amounts to adopting

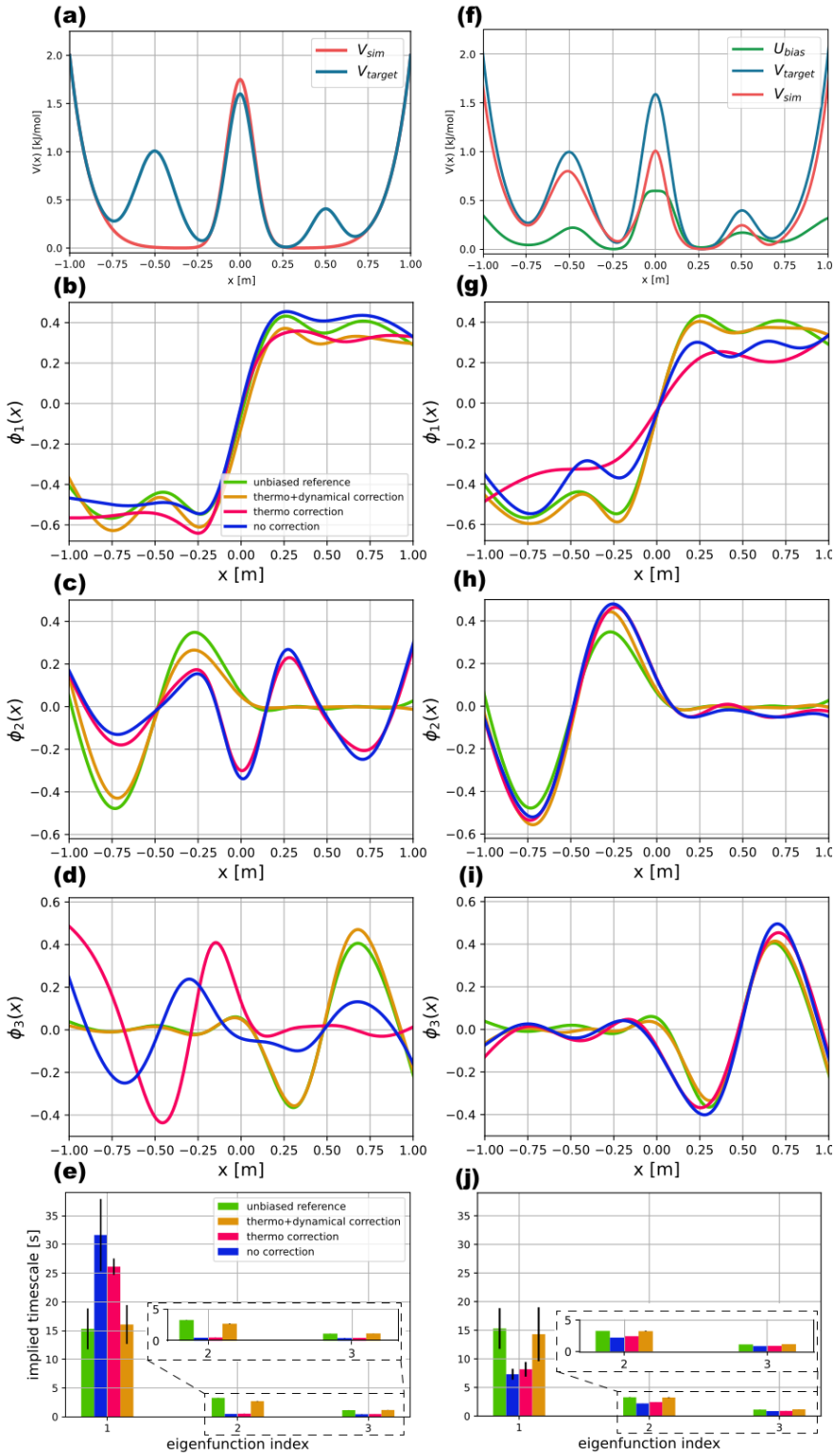

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
V_{sim}
V_{target}
U_{bias}
V(x) [kJ/mol]
x [m]
$\phi_1(x)$
$\phi_2(x)$
$\phi_3(x)$
unbiased reference
thermo+dynamical correction
thermo correction
no correction
implied timescale [s]
eigenfunction index

Figure 7.2

Figure 7.2: Application of GREST to the 1D 4-well potential as an illustrative toy example. **(Left column.)** Application of GREST under pre-defined bias. **(a)** The target 4-well potential V_{target} (blue) is shown alongside the 2-well potential under which the biased simulation is performed V_{sim}. The challenge is for path reweighted SRVs to correctly recover the slowly relaxing modes and associated time scales of the 4-well system in simulations conducted under the 2-well potential. **(b-d)** Comparison of the leading three leading slow eigenfunctions $\{\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \phi_3(\mathbf{x})\}$ recovered by path reweighted SRVs from long unbiased reference data V_{target} (green), alongside eigenfunctions learned using biased simulation data collected under V_{sim} to which we apply path weights comprising dynamical and thermodynamic corrections (orange), only thermodynamic corrections (red), and no reweighting corrections (blue). Eigenvectors are defined only up to scaling and sign so are presented under unit normalization and, if necessary, inversion of sign to align them with the unbiased reference data. **(e)** Associated implied time scales $t_i = -\tau/\ln(\lambda_i)$ for the SRV eigenfunctions in panels **(b-e)** with error bars showing the standard deviation over 10 independently trained models with different random seeds. **(Right column.)** Application of GREST under a learned bias. **(f)** A biasing potential U_{bias} is constructed by projecting the unbiased reference data into the leading slow mode $\phi_1(\mathbf{x})$ learned by a path reweighted SRV fit to the unbiased 4-well simulation trajectories. Application of the bias to the target potential results in a simulation potential $V_{sim} = V_{target} - U_{bias}$ that attenuates the highest central potential energy barrier by $\sim$0.5 kJ/mol and the other potential energy barriers by $\sim$0.25 kJ/mol. **(g-i)** Comparison between the reference eigenfunctions and those estimated from biased simulation data under V_{sim} in panel **(f)** to which we apply various path weights. **(j)** Associated implied time scales $t_i = -\tau/\ln(\lambda_i)$ for the SRV eigenfunctions in panels **(g-i)** with error bars showing the standard deviation over 10 independently trained models with different random seeds.

$U_{bias} = V_{4-well} - V_{2-well}$ as our biasing potential. The effect of this bias is to flatten the two smaller barriers to leave only the single high barrier at the origin intact. We therefore challenge our path reweighted SRVs to recover the correct slow modes and time scales of the 4-well system under simulations conducted under the 2-well potential.

We present in Fig. 8.1b-d (green) approximations to the three leading slow eigenfunctions $\{\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \phi_3(\mathbf{x})\}$ learned using SRVs from long unbiased reference trajectories collected under $V_{target} = V_{4-well}$. The leading eigenfunction $\phi_1(\mathbf{x})$ changes sign at $\mathbf{x} = 0$ m and corresponds to the slowest occurring transitions over the largest potential barrier. The next eigenfunction $\phi_2(\mathbf{x})$ is approximately zero in the right half space and changes sign at $\mathbf{x} =$ (-0.5) m to characterize transitions over the next highest potential barrier. Finally, $\phi_3(\mathbf{x})$

is approximately zero in the left half space and changes sign at $\mathbf{x} = +0.5$ m to characterize transitions over the lowest potential barrier. The associated implied time scales for these relaxations are presented in Fig. 8.1e.

We also present in Fig. 8.1b-e the eigenfunctions and implied time scales learned by path reweighted SRVs from biased simulation data collected under $V_{sim} = V_{2-well}$ under different choices of path weights: dynamical and thermodynamic corrections (orange), only thermodynamic corrections (red), and no reweighting corrections (blue). The 2-well landscape differs from the 4-well landscape largely by flattening of the two lower potential barriers and leaving the largest central potential barrier largely intact. Recognizing that the slowest relaxing mode is associated with the same dynamical process of transitions over this central barrier, we anticipate good agreement between the leading slow mode estimated under the 4-well and 2-well simulation trajectories. Inspection of Fig. 8.1b shows that this is indeed the case: the estimate of $\phi_1(\mathbf{x})$ from unbiased reference data collected under V_{4-well} (green) is in good agreement in the uncorrected estimate of $\phi_1(\mathbf{x})$ computed from simulation trajectories collected under V_{2-well} (blue). Application of thermodynamic only ($W(\boldsymbol{\omega}_{t\rightarrow t+\tau}) = g(\mathbf{x}_t)$, red) or thermodynamic and dynamical ($W(\boldsymbol{\omega}_{t\rightarrow t+\tau}) = g(\mathbf{x}_t) \times M(\boldsymbol{\omega}_{t\rightarrow t+\tau})$, orange) reweighting corrections within the SRV estimation (cf. Eqn. 7.29) therefore does little to improve the agreement with the $\phi_1(\mathbf{x})$ unbiased reference. Reweighting is, however, critical in accurate estimation of the correct implied time scale associated with this mode (Fig. 8.1e, left). The largest error in predicting the reference time scale of the leading mode (green) occurs when no reweighting corrections are applied (blue). Incorporating thermodynamic corrections alone improves agreement (red), but only by incorporating both the thermodynamic and dynamical corrections do we accurately recover the true time scale (orange).

The 2-well potential lacks potential barriers at $\mathbf{x} = (-0.5)$ m and $\mathbf{x} = +0.5$ m meaning that biased simulations conducted under $V_{sim} = V_{2-well}$ do not contain any signatures of relaxations associated with these barrier crossings and these modes are completely obfuscated

due to obliteration of these barriers by the applied bias (Fig. 8.1b-e, blue). Nevertheless, $\phi_2(\mathbf{x})$ and $\phi_3(\mathbf{x})$ and the associated implied time scales are accurately recovered when incorporating both thermodynamic and dynamical reweighting corrections into SRV learning (orange). Interestingly, the thermodynamic correction alone does very little to improve estimation of the true slow eigenfunctions and implied time scales, cautioning against omission of the dynamical path weights in estimating slow CVs [426]. This example illustrates that path reweighting can accurately learn the slow dynamical modes and time scales of a dynamical system from biased simulation trajectories.

Learned bias. We now test the capacity of the learned slow CVs to support an accelerating biasing potential and promote enhanced sampling of configurational phase space. This can be approximately conceived as the second round of a GREST campaign in which sampling is enhanced by accelerating biases in learned slow modes. We construct the bias potential by assuming that the projection of the unbiased reference simulation into the leading slow mode $P(\phi_1(\mathbf{x}))$ is Boltzmann distributed and may therefore be inverted to construct an effective potential $F(\phi_1(\mathbf{x})) = -k_B T \log(P(\phi_1(\mathbf{x})))$. This represents a good candidate biasing potential as it flattens out the portion of the potential associated with dynamical transitions in $\phi_1(\mathbf{x})$ and can therefore help accelerate sampling of these events. This leads us to define the biasing potential as the functional composition $U_{bias} = \mathcal{F}(\phi_1(\mathbf{x}))$, where $\mathcal{F}(\phi_1)$ is an analytical fit to $F(\phi_1)$ that we find to be well described by a super-Gaussian $F(\phi_1) \approx \mathcal{F}(\phi_1) = A \exp(-(\frac{(\phi_1 - x_0)^2}{2\sigma^2})^P)$ with $A = 3.1$ kJ/mol, $x_0 = 0.049$ m, $\sigma^2 = 0.058$ m^2, and $P = 2$ (see Supporting Information). This functional fit also admits continuous first derivatives required to compute the biasing forces on the particle through the chain rule, $\mathbf{f}_{bias} = -\frac{dU_{bias}}{d\mathbf{x}} = -\frac{d\mathcal{F}}{d\phi_1}\frac{d\phi_1}{d\mathbf{x}}$, where $\frac{d\mathcal{F}}{d\phi_1}$ is available analytically and $\frac{d\phi_1}{d\mathbf{x}}$ from automatic differentiation through the trained SRV neural network. In Fig. 8.1g, we illustrate $V_{target} = V_{4-well}$, U_{bias}, and $V_{sim} = V_{target} - \gamma U_{bias} = V_{4-well} - 0.2 U_{bias}$, where we have chosen to set the attenuation factor to $\gamma = 0.2$ (cf. Section 7.3.5). The biasing potential

depresses the height of the central highest potential barrier by ∼0.5 kJ/mol, resulting in an observed ∼150% increase in the number of $(\mathbf{x} < 0) \leftrightarrow (\mathbf{x} > 0)$ transitions. It also suppresses the two auxiliary barriers at $\mathbf{x} = -0.5$ m and $\mathbf{x} = 0.5$ m by ∼0.25 kJ/mol, resulting in modest increases in the number of $(\mathbf{x} < -0.5) \leftrightarrow (\mathbf{x} > -0.5)$ and $(\mathbf{x} < 0.5) \leftrightarrow (\mathbf{x} > 0.5)$ transitions by ∼20% and ∼14%, respectively.

We present in Fig. 8.1g-j the reference and learned eigenfunction approximations under different reweighting schemes. Due to the relatively small perturbation to the target potential, we see qualitative agreement with the unbiased reference eigenfunctions (green) even in the absence of any corrections (blue) or thermodynamic-only corrections (red), but the full thermodynamic and dynamical corrections (orange) are required to achieve quantitative agreement. The second and third implied time scales are in good agreement for all reweighting schemes since the barriers associated with these dynamical transitions are only moderately affected by the application of the bias, but we see that both the thermodynamic and dynamical corrections are imperative for accurate estimation of the leading time scale associated with transitions over the highest central potential barrier. This example demonstrates that an SRV-learned slow mode can be used to effectively accelerate slow dynamical transitions in a biased simulation from which path reweighted SRVs can accurately recover the true slow modes.

7.4.2 Alanine dipeptide

We now demonstrate our approach in an application to the widely studied biomolecular system alanine dipeptide (ADP). This system represents the "hydrogen atom of molecular simulation" [483] and we challenge GREST to simultaneously estimate the slow CVs governing its biomolecular dynamics and recover the underlying molecular free energy landscape without any prior knowledge of this system. Full details of the application of GREST to this system including all path reweighted SRV training settings and metadynamics hyperparam-

eters are provided in the Supporting Information.

Unbiased reference benchmark. We first perform a long unbiased simulation to serve as a baseline reference for the thermodynamics and kinetics of our ADP system against which to benchmark the performance of GREST. We present in Fig. 7.3a-i the ADP free energy surface as a ϕ-ψ projection into the backbone dihedral angles that are known to be good CVs with which to differentiate the metastable conformational states of the system [413, 414]. To ensure comprehensive sampling of configurational space, the reference simulation data was gathered from 8×0.5 μs simulations initialized from configurations nearest k-means centroids identified by k-means clustering the ϕ-ψ projection of an initial 250 ns unbiased simulation. The slowest relaxing implied time scale for this system is $\sim$1.4 ns, so these reference trajectories provide excellent sampling of all relevant transitions in this ADP system. We train an SRV over these unbiased simulation data in which we represent the instantaneous state of the ADP system through all $\binom{22}{2} = 231$ pairwise atomic distances to define a translationally and rotationally invariant featurization $\mathbf{x} \in \mathbb{R}^{231}$ of the system.

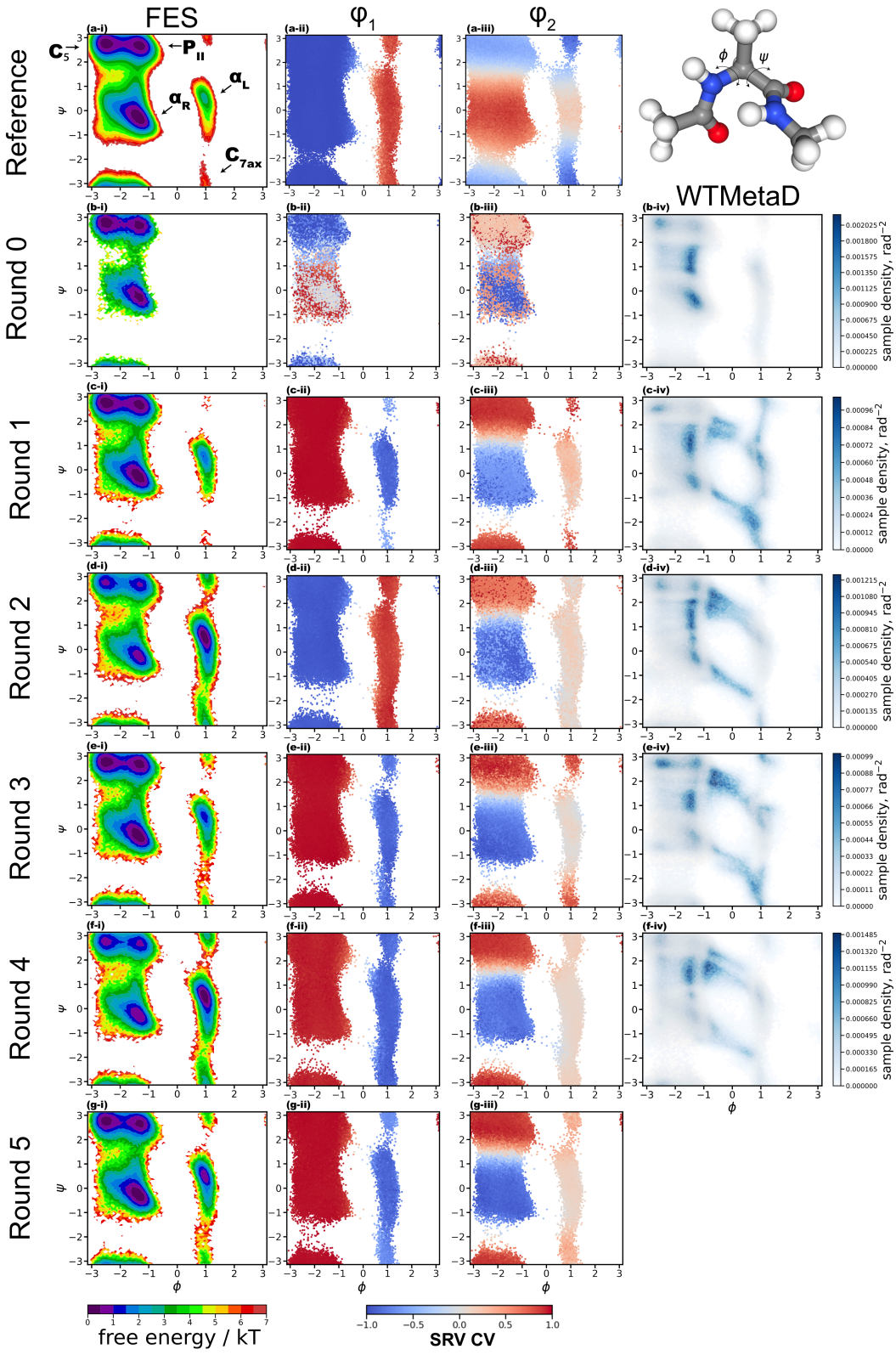

FES
φ1
φ2
WTMetaD
Reference
Round 0
Round 1
Round 2
Round 3
Round 4
Round 5
C5
PII
αL
αR
C7ax
ϕ
ψ
free energy / kT
SRV CV
sample density, rad−2

Figure 7.3

Figure 7.3: Application of GREST to ADP. **(a-i)** Reference free energy profile constructed from 4 μs of unbiased simulation data projected into the ϕ-ψ backbone angles. The local free energy minima are annotated with their associated macrostate labels: C_5, P_{II}, α_R, α_L, and C_{7ax}. **(a-ii,iii)** Leading two slow CVs $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ identified by application of SRVs to the unbiased reference trajectories. $\phi_1(\mathbf{x})$ characterizes transitions in the ϕ dihedral between $\{C_5, P_{II}, \alpha_R\}$ and $\{\alpha_L, C_{7ax}\}$. $\phi_2(\mathbf{x})$ primarily characterizes transitions in the ψ dihedral between $\{C_5, P_{II}\}$ and α_R. Rows **(b)-(g)** index successive rounds of GREST, commencing with the initial unbiased Round 0 in **(b)** and ending with the terminal adaptive sampling Round 5 in **(g)**. Columns **(i)-(iv)** index various facets of each round of GREST: **(i)** illustrates the unbiased free energy landscape estimated by thermodynamic reweighting of the enhanced sampling simulations performed under the fixed terminal metadynamics bias projected into the ϕ-ψ backbone angles, **(ii)** and **(iii)** illustrate heatmaps of the path reweighted SRV slow modes $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ learned from the enhanced sampling trajectories in the terminal metadynamics bias potential projected into ϕ-ψ space, and **(iv)** illustrates the phase space sampling density projected in ϕ-ψ space from well-tempered metadynamics simulations performed in the two learned slow modes $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ illustrated in the preceding columns. The GREST iterations are terminated when the CVs learned in successive rounds stabilize, indicating no additional enhancement in configurational and dynamical exploration of phase space. Convergence of GREST at Round 4 means that additional enhanced sampling calculations are not conducted in Round 5.

The slowest process captured by the first SRV mode $\phi_1(\mathbf{x})$ characterizes transitions between the triplet of metastable states $\{C_5, P_{II}, \alpha_R\}$ where $\phi < 0$ and the doublet of metastable states $\{\alpha_L, C_{7ax}\}$ where $\phi > 0$ (Fig. 7.3a-ii). The second slow mode $\phi_2(\mathbf{x})$ largely subpartitions the $\{C_5, P_{II}, \alpha_R\}$ triplet by characterizing transitions in ψ between $\{C_5, P_{II}\}$ and α_R (Fig. 7.3a-iii). Throughout our adaptive sampling workflow we constrain the path reweighted SRV to return two slow CVs since it is known that ADP possesses a 2D intrinsic manifold lying on the surface of a flat torus that is well parameterized by the ϕ and ψ backbone dihedrals [404, 484, 485, 486, 487]. In general, one would determine the appropriate number of CVs to retain by searching for a gap in the SRV eigenvalue spectrum at each round of GREST. Our goal is to learn to recapitulate these kinetics by biasing along dynamically estimated slow modes to visit relevant macrostates and efficiently sample state-to-state transitions.

Application of GREST. We commence Round 0 of our adaptive sampling workflow

by running an initial unbiased 5 ns simulation with the resultant free energy profile shown in (Fig. 7.3b-i). This initial simulation only samples the triplet of metastable states C_5, P_{II}, and α_R residing in the low-ϕ/high-ψ superbasin and fails to experience any transitions to the $\phi > 0$ region. In the next step, we use this initial data to estimate the leading two SRV slow modes (Fig. 7.3b-ii,iii). The paucity of training data in this initial round contributes to learning a pair of seemingly degenerate slow modes that only capture transitions in ϕ. Equipped with these learned slow modes and initial simulation data, we perform k-means clustering in the feature space of the 231 pairwise atomic distances to identify configurations nearest 25 k-means centroids that are used to initialize parallel walker metadynamics along our learned SRV CVs in the subsequent step of GREST. Each walker is simulated for 1 ns yielding an aggregate 25 ns of metadynamics simulation data. In Fig. 7.3b-iv we show the sampling density from this metadynamics simulation as a visualization of the phase space explored due to accelerating along the learned SRV slow modes $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$. The resultant phase space exploration along these learned CVs reveals that biasing has largely driven transitions in ψ across the free energy barrier bridging metastable states P_{II} and α_R. To complete Round 0, we perform enhanced sampling simulations using a time invariant and attenuated version of the termminal metadynamics bias potential, which we reduce in magnitude by an attenuation factor of $\gamma = 0.05$ to control the numerical instability associated with very large and very small magnitude reweighting factors during the subsequent SRV training [110, 112]. These enhanced sampling simulations are similarly executed in parallel with 25×1 ns simulations initialized from configurations identified with the same k-means clustering protocol as in the previous step but applied to these 25 ns of metadynamics simulation trajectories.

In Fig. 7.3c, we illustrate Round 1 of our adaptive sampling protocol which proceeds analogously to the initial round except that the training data now comprises enhanced sampling trajectories biased in the learned slow CVs from the previous round under the $\gamma = 0.05$

attenuation factor. Although the applied biasing forces are relatively modest, the reweighted free energy surfaces in Fig. 7.3c-i demonstrate that the biased trajectories experience transitions in ϕ and now visit metastable states α_L and C_{7ax} that were absent in the initial unbiased data. Training path reweighted SRVs over these trajectories with thermodynamic and dynamical corrections exposes two leading slow modes illustrated in Fig. 7.3c-ii,iii that are now in much better agreement with those extracted from the long unbiased reference data. (We recall that the sign of the learned eigenfunctions is immaterial.) We also see the subsequent WTMetaD simulations performed along these SRV CVs by initializing 25×1 ns walkers by k-means clustering in the feature space of 231 pairwise atomic distances drives more comprehensive sampling of configurational phase space (Fig. 7.3c-iv). Visual analysis reveals that this time the enhanced sampling has driven transits over the free energy barriers connecting P_{II} and α_R, P_{II} and α_L, α_R and C_{7ax}, and α_L and C_{7ax}. To complete Round 1, we perform a final enhanced sampling simulation applying the terminal WTMetaD potential under a $\gamma = 0.05$ attenuation factor using 25×1 ns walkers initialized by k-means clustering in the feature space of 231 pairwise atomic distances, and pass this forwards as the biased trajectory training data for Round 2.

We iterate this process to complete Rounds 2-5 of GREST that are illustrated in Fig. 7.3d-g. Convergence is assessed by tracking the round-to-round change in the learned slow SRV modes. Stabilization of the learned CVs between successive GREST rounds indicates that no important new dynamical modes or regions of configurational phase space are being explored under additional rounds of biased sampling. We quantify convergence by examining the Pearson correlation coefficient of SRV slow modes learned in consecutive rounds. Importantly, we maximize the correlation coefficient under rigid rotations of the subspace spanned by the slow modes in successive rounds to account for possible rotational symmetries in which the same subspace could be spanned by an equivalent rotation of the learned slow eigenfunctions. Mathematically, this corresponds to identifying the affine transformation that maximizes the

correlation coefficient. Convergence is defined by stabilization in the number n of recovered slow modes and round-to-round Pearson correlation coefficients $\{\rho_i > 0.8\}_{i=1}^{n}$ for all modes. In the present case the former criterion is satisfied by default since we fix the dimensionality of the SRV bottleneck layer to 2D. Consistent with expectations from inspection of the free energy surfaces and slow CVs in Fig. 7.3, we illustrate in Fig. 7.4 that the latter criterion is quickly satisfied after only Round 2, but we continue for three additional rounds for the purposes of illustration.

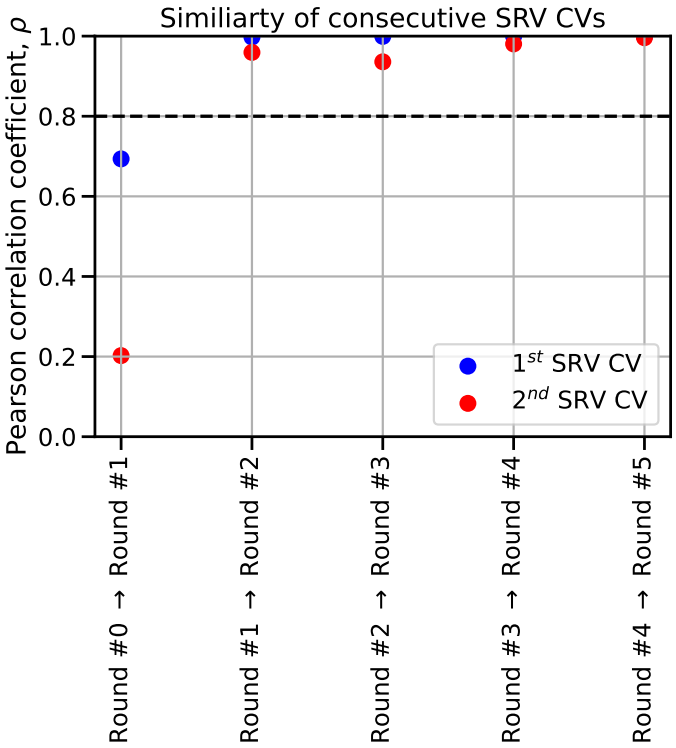


Figure 7.4: Convergence assessment of GREST applied to ADP by measuring the Pearson correlation coefficient between SRV slow CVs discovered in successive rounds. Stabilization of the CVs to $\rho > 0.8$ indicates that no important new dynamical modes or regions of configurational phase space are being explored under additional rounds of biased sampling.

The excellent agreement between the free energy surface and slow modes of the terminal round (Fig. 7.3g) and the unbiased reference data (Fig. 7.3a) indicate that the GREST iterations successfully recovered the true slow dynamical modes and used these to efficiently enhance sampling of configurational phase space. While the ADP system does represent a

demonstrative and instructive molecular example for which the ground truth is well understood, its small size and efficiency to simulate make it a relatively simple test of the GREST methodology. In particular, expending the same computational effort expended within the GREST pipeline (250 ns of cumulative simulation time) in unbiased simulations would accurately recover the free energy surface by brute force and enable SRV discovery of the slow modes without path reweighting. Moreover, the relatively low free energy barriers mean that it is difficult to assess the gains realized by the relatively mild biases offered by the $\gamma = 0.05$ attenuated enhanced sampling simulations since these barriers can be surmounted by sufficiently long unbiased trajectories. In our next application, we apply GREST to a larger, more complex molecular system where we demonstrate that the protocol does realize substantial efficiency gains.

7.4.3 WLALL pentapeptide

As a more complex biomolecular system, we apply GREST to the Trp-Leu-Ala-Leu-Leu (WLALL) pentapeptide [418]. This system exhibits a number of metastable states separated by free energy barriers that are sufficiently high to demonstrate the capabilities of GREST to identify and surmount these barriers while still being sufficiently low to enable determination of a ground truth benchmark by long unbiased simulations.

Unbiased reference benchmark. We generate unbiased reference simulation data for the WLALL system in a similar manner to ADP. We first conduct an initial unbiased 250 ns simulation then apply k-means clustering in a 2D TICA embedding using pairwise backbone distances as features to seed initial configurations for 8×0.75 μs independent simulations and produce 6 μs of aggregated reference data. As we will see, the slowest relaxing time scale in this system is on the order of 19.6 ns, so these data provide comprehensive sampling of configurational space. For the WLALL system, and many other systems of practical interest, intuitive low dimensional coordinates suitable for visualization, such as the two backbone

dihedrals for ADP, are not known *a priori*. For convenient visualization of the WLALL free energy surface, we use TICA [421, 413, 423] as a simple and convenient dimensionality reduction technique to recover a two-dimensional representation of the WLALL phase space independently of our SRVs. We represent the WLALL system to both the TICA and SRV approaches as the $\binom{20}{2} = 190$ pairwise atomic distances between the 20 backbone atoms to define a translationally and rotationally invariant featurization $\mathbf{x} \in \mathbb{R}^{190}$. The free energy surface resulting from projection of the unbiased reference trajectories into the leading two TICA coordinates TIC_1 and TIC_2 is presented in Fig. 7.5a and reveals three predominant macrostates: the unfolded ensemble (labeled *1*), the folded state (*2*), and the misfolded state (*3*). With the global free energy minima located in the unfolded ensemble under the prevailing thermodynamic conditions, the compact folded state is differentiated from the misfolded state with a comparatively lower free energy minima.

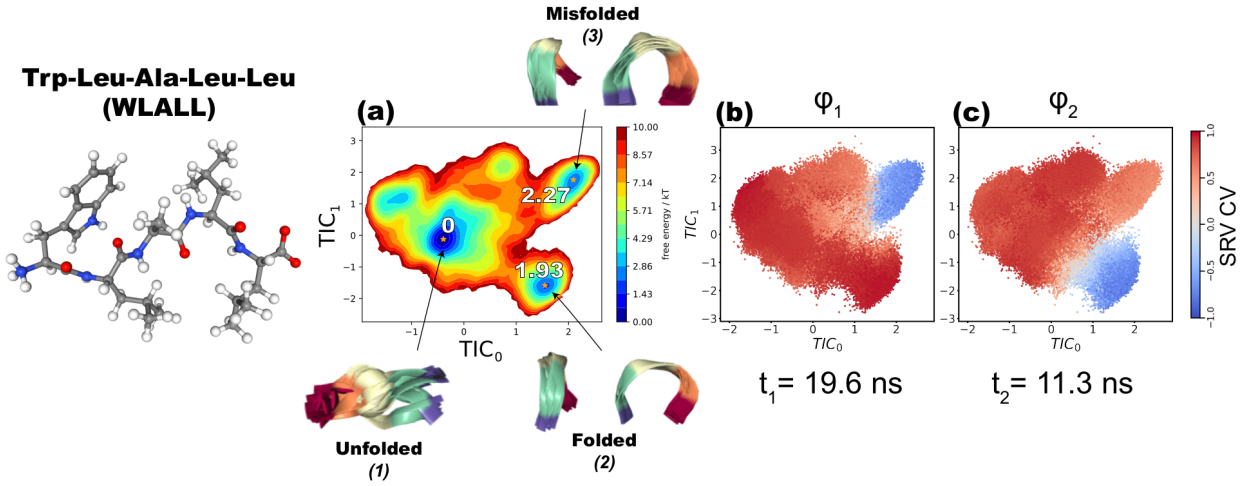


Figure 7.5: Unbiased reference simulations of WLALL pentapeptide. **(a)** The free energy surface resulting from projection of 6 μs of unbiased simulation data into the leading two TICA coordinates identifies three predominant metastable minima marked by golden stars and annotated with their corresponding relative free energy. Insets show a collection of representative structures from each minima identifying the *(1)* unfolded state, *(2)* folded state, and *(3)* misfolded state. **(b)** Application of SRVs to the unbiased reference data resolves a leading slow mode $\phi_1(\mathbf{x})$ with an implied time scale of $t_1 = 19.6$ ns characterizing transitions in and out of the misfolded state. **(c)** The next slowest mode $\phi_2(\mathbf{x})$ possesses an implied time scale of $t_2 = 11.3$ ns and corresponds to transitions in and out of the folded state.

We next train an SRV over the unbiased reference data. An implied time scale analysis reveals a gap in the eigenvalue spectrum after 2-3 leading slow modes (Fig. S3 in the Supporting Information). For simplicity and visualization convenience, choose to retain the leading two slow modes throughout our analyses of the WLALL system. As we shall see, this results in empirically good performance in recapitulating and enhancing sampling of the pentapeptide dynamics. The first slow mode $\phi_1(\mathbf{x})$ possesses an implied time scale of t_1 = 19.6 ns and characterizes transitions in and out of the misfolded state (Fig. 7.5b). The second slow mode $\phi_2(\mathbf{x})$ possesses an implied time scale of t_2 = 11.3 ns and corresponds to transitions in and out of the folded state (Fig. 7.5c). Having established this baseline understanding of the structure, thermodynamics, and kinetics from a 6 μs unbiased reference simulation, we now challenge GREST to automatically discover these slow modes and exploit them for efficient enhanced sampling.

Application of GREST. We follow the same adaptive sampling routine for the WLALL system as we demonstrated with ADP, beginning with a short unbiased 20 ns simulation to initialize Round 0. We illustrate the resulting free energy surface in Fig. 7.6a-i by projecting these data into the two leading TICA coordinates TIC_1 and TIC_2 identified from our analysis of the 6 μs unbiased reference simulations. (We emphasize that the TICA coordinates are used exclusively to provide a consistent and interpretable visual presentation of our data as we progress through the GREST iterations, and the unbiased simulation data and TICA analysis thereof play no role in the GREST analysis pipeline.) This initial Round 0 remains trapped in the unfolded ensemble and fails to explore any of the folded or misfolded states, which results in learned SRV slow modes that only characterize transitions within the unfolded ensemble (Fig. 7.6a-ii,iii). The subsequent WTMetaD run (Fig. 7.6a-iv) and subsequent simulation employing a time invariant $\gamma = 0.05$ attenuated terminal metadynamics bias are executed in the same manner as ADP, but in this instance we perform k-means clustering in the space of our pairwise backbone distance featurization $\mathbf{x} \in \mathbb{R}^{190}$ to

initialize 25 $\times$ 1 ns parallel walkers. In Round 1, we observe that the system escapes the unfolded ensemble and successfully explores the folded state (Fig. 7.6b-i), which is now captured in the leading SRV slow mode (Fig. 7.6b-ii) and results in accelerated metadynamics sampling along the transition path connecting the unfolded and folded states (Fig. 7.6b-iv). This process continues with the discovery and SRV characterization of the misfolded state in Round 2 (Fig. 7.6c), and convergence of the SRV slow modes after approximately five rounds (Fig. 7.6a-f, Fig. 7.7).

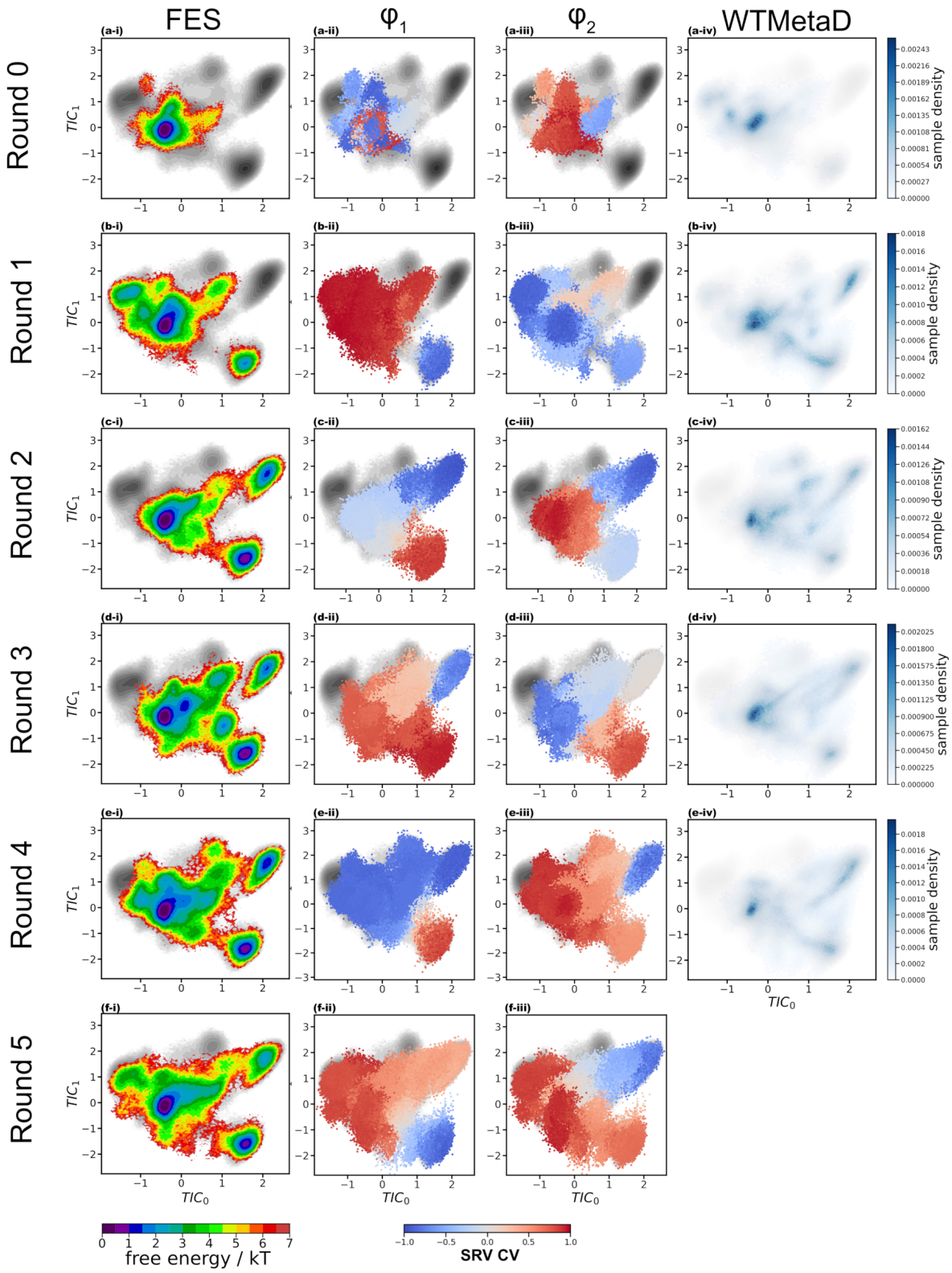

FES
φ1
φ2
WTMetaD
Round 0
Round 1
Round 2
Round 3
Round 4
Round 5
(a-i)
(a-ii)
(a-iii)
(a-iv)
(b-i)
(b-ii)
(b-iii)
(b-iv)
(c-i)
(c-ii)
(c-iii)
(c-iv)
(d-i)
(d-ii)
(d-iii)
(d-iv)
(e-i)
(e-ii)
(e-iii)
(e-iv)
(f-i)
(f-ii)
(f-iii)
TIC0
TIC1
sample density
free energy / kT
SRV CV

Figure 7.6

Figure 7.6: Application of GREST to WLALL pentapeptide. Rows **(a)-(f)** index successive rounds of GREST, commencing with the initial unbiased Round 0 in **(a)** and ending with the terminal adaptive sampling Round 5 in **(f)**. Columns **(i)-(iv)** index various facets of each round of GREST: **(i)** illustrates the unbiased free energy landscape estimated by thermodynamic reweighting of the enhanced sampling simulations performed under the fixed terminal metadynamics bias data projected into the two leading TICA coordinates TIC_1 and TIC_2, **(ii)** and **(iii)** illustrate heatmaps of the path reweighted SRV slow modes $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ learned from the enhanced sampling trajectories in the terminal metadynamics bias potential projected into TIC_1-TIC_2 space, and **(iv)** illustrates the phase space sampling density projected into TIC_1-TIC_2 space from well-tempered metadynamics simulations performed in the two learned slow modes $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ illustrated in the preceding columns. The GREST iterations are terminated when the CVs learned in successive rounds stabilize, indicating no additional enhancement in configurational and dynamical exploration of phase space. Convergence of GREST at Round 4 means that additional enhanced sampling calculations are not conducted in Round 5.

Visual comparison of the free energy surface in the terminal round of GREST (Fig. 7.6f-i) with that computed from the long unbiased reference data (Fig. 7.5a) demonstrates that GREST learned to identify the relevant transitions between states and to recover all of the important metastable states. Relative to the ∼20 ns relaxation time of the slowest mode in the system, GREST also accurately learned the leading slow relaxations from quite parsimonious simulation data. A path reweighted SRV fitted to the accumulated 125 ns of biased GREST simulation data identifies two slow modes that exhibit Pearson correlations of $\rho_1 = 0.91$ and $\rho_2 = 0.99$, respectively, with those estimated from the 6 μs of unbiased reference data. An implied time scale analysis of the biased data using a path reweighted SRV also reveals converged timescales of the leading two slow modes in good agreement with the unbiased reference data (Fig. S4 in the Supporting Information).

Finally, we verify that the enhanced sampling procedure effected within GREST does substantially improve sampling even under the $\gamma = 0.05$ attenuation factor in the metadynamics bias potential by surgically directing sampling along the relevant slow modes albeit with relatively modest accelerations to preserve numerical stability. Employing the same computational budget of 250 ns of simulation time, we demonstrate that GREST outper-

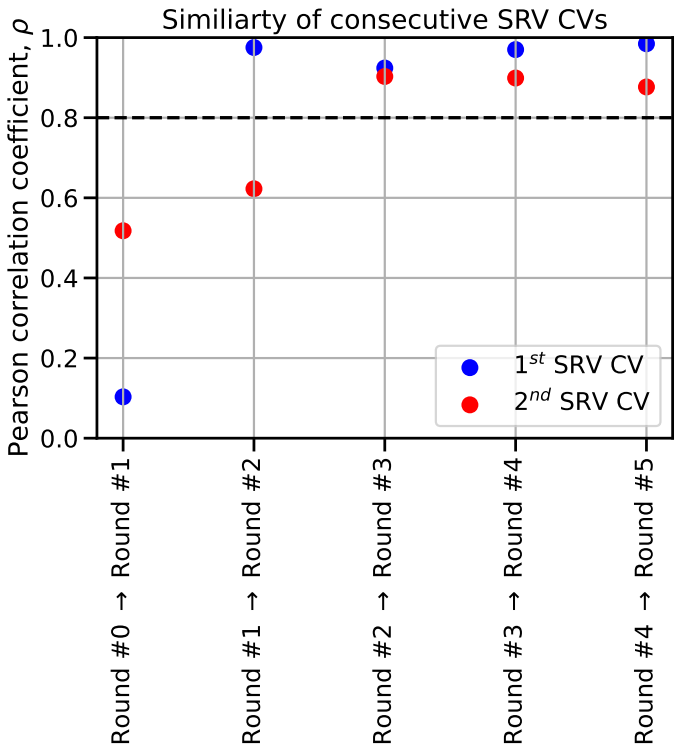


Figure 7.7: Convergence assessment of GREST applied to WLALL pentapeptide by measuring the Pearson correlation coefficient between SRV slow CVs discovered in successive rounds. Stabilization of the CVs to $\rho > 0.8$ indicates that no important new dynamical modes or regions of configurational phase space are being explored under additional rounds of biased sampling.

forms (i) a single long 250 ns unbiased simulation, (ii) 125 ns of unbiased simulation + 10 $\times$ 12.5 ns of unbiased simulations seeded by k-means clustering in pairwise backbone atom distances, and (iii) 250 ns of metadynamics performed in the squared molecular radius of gyration R_g^2 as an intuitive CV (Fig. S5 in the Supporting Information). These alternative approaches fail to identify all relevant metastable states and result in free energy surfaces with inferior coverage of the thermally accessible phase space compared to GREST.

7.5 Discussion & Conclusions

In this work we introduce Girsanov Reweighting Enhanced Sampling Technique (GREST) as an adaptive sampling scheme that interleaves rounds of data-driven slow collective variable

discovery and enhanced sampling along these coordinates. We employ state-free reversible VAMPNets (SRVs) to perform data-driven discovery of slow collective variables. The estimation of dynamical observables under an unbiased Hamiltonian from biased trajectories collected under a perturbed Hamiltonian is enabled by the Girsanov formalism [108, 109, 110, 111, 112] that prescribes how to apply reweighting of trajectories in the biased path ensemble according to both thermodynamic and integrator-specific dynamical path weights. This unlocks the ability to bias along slow coordinates to accelerate exploration of configurational phase space and then recover improved estimates for these slow coordinates under the enhanced sampling by applying Girsanov reweighting to the biased trajectories. This capability is the key feature underpinning GREST that interleaves successive rounds of path reweighted SRV slow collective variable discovery and metadynamics enhanced sampling in the learned coordinates until we achieve convergence.

In an application to a toy one-dimensional four-well system, we developed intuition for the behavior of GREST and showed how biasing along SRV learned CVs serve to reduce potential energy barriers mediating slow dynamical processes and accelerate sampling of phase space. In applications to two biomolecular systems – alanine dipeptide and the WLALL pentapeptide – we demonstrated how GREST can accelerate sampling of configurational phase space and furnish accurate estimates of the slow collective variables without any prior knowledge of the system. We achieve significant speedups in sampling the configurational phase space of the WLALL pentapeptide that enable GREST to discover all relevant metastable states whereas unbiased simulations and enhanced sampling in intuitive coordinates that are not well correlated with the slow dynamical relaxations fail to explore all metastable states in an equivalent simulation time.

In future work, we anticipate that GREST may be a valuable approach in realizing significant efficiency gains in exploring the configurational phase space of large biomolecular and macromolecular systems. We also observe that enhanced sampling need not be limited

to just those learned slow coordinates, and we propose that it may be profitable to augment GREST with sampling in both intuited and learned coordinates as an inexpensive means of incorporating prior knowledge into the approach. GREST is generically compatible with any off-the-shelf collective variable biasing technique, but may also benefit from integration with parallel tempering / replica exchange / Hamiltonian exchange formalisms to perform generic acceleration of "all" system degrees of freedom as well as targeted enhanced sampling along the learned slow modes [488, 489, 490, 491, 492, 493]. Path weights under Hamiltonian exchange can be treated under the same formalism presented in this work; analytical expressions for temperature reweighting and efficient numerical solution schemes have been previously reported by Chodera and co-workers [447, 82, 466]. Another future extension of GREST would involve application to explicit solvent systems. Exposing both solvent and solute degrees of freedom to the SRV network via permutationally invariant featurizations, such as the Permutation Invariant Vector (PIV) formalism proposed by Fabio Pietrucci and co-workers [494, 495], would enable learning of slow CVs over the solute and (permutationally invariant) solvent degrees of freedom. However, Basconi and Shirts have shown that Langevin dynamics, along with other velocity randomizing thermostats, may not be ideal for simulating solvated systems because of their tendency to artificially dampen dynamical processes due to disturbances of the natural time correlations of particle velocities and that these artifacts can only be avoided by using weak coupling constants that can lead to imprecise temperature control [496]. As such, care must be taken that the Langevin integration schemes for which analytical expressions for Girsanov weights are available must be deployed with care. A further avenue of exploration would be to conduct a comparison of the Girsanov reweighting formalism with the square root approximation (SqRA) to dynamical reweighting [497, 498, 499]. The SqRA requires quite strong assumptions regarding the character of the potential energy landscape and constancy of the diffusivity under an applied bias and so can only offer an approximate reweighting scheme, but can convey advantages in terms of

numerical robustness when these conditions are satisfied and it offers an accurate approxiamtion to the exact Girsanov expressions. The SqRA has been explored as an alternative to Girsanov reweighting by Donati, Weber, and Keller in the context of Markov state models [499], and it would be interesting to explore the quality of this approximation and its range of applicability in the context of GREST. Finally, we observe that GREST may readily be applied to generic dynamical systems beyond molecular dynamics to accelerate phase space exploration in numerical simulations in, for example, ecological modeling, climate science, process systems modeling, or financial markets. The absence of detailed balance guarantees in generic dynamical systems mean that the SRVs must be generalized to SNRVs (state-free non-reversible VAMPNets) using the non-reversible VAC formalism [101, 500], but otherwise the approach may be deployed in a largely analogous manner to that for molecular systems herein.

Data availability

A Python package implementing SRVs with capabilities for training over biased simulation data is publicly available via GitHub at `https://github.com/andrewlferguson/snrv`. Example Jupyter notebooks illustrating the application of GREST to the 4-well potential, alanine dipeptide, and WLALL pentapeptide systems along with all associated simulation data are hosted on Zenodo at DOI:10.5281/zenodo.7552310 [334].

Supporting Information

The Supporting Information is available free of charge at `https://pubs.acs.org/doi/10.1021/acs.jpca.3c00505`

CHAPTER 8

DIAMONDBACK: DIFFUSION-DENOISING AUTOREGRESSIVE MODEL FOR NON-DETERMINISTIC BACKMAPPING OF Cα PROTEIN TRACES

8.1 Abstract

Coarse-grained molecular models of proteins permit access to length and time scales unattainable by all-atom models and the simulation of processes that occur on long-time scales such as aggregation and folding. The reduced resolution realizes computational accelerations but full atomistic resolution can be vital for a complete understanding of mechanistic details. Backmapping is the process of restoring all-atom resolution to coarse-grained molecular representations. In this work, we report DiAMoNDBack (Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping) as an autoregressive denoising diffusion probability model to restore all-atom details to coarse-grained protein representations retaining only Cα coordinates. The autoregressive generation process proceeds from the protein N-terminus to C-terminus in a residue-by-residue fashion conditioned on the Cα trace and previously backmapped backbone and side chain atoms within the local neighborhood. The local and autoregressive nature of our model makes it transferable between proteins. The stochastic nature of the denoising diffusion process means that the model generates a realistic ensemble of backbone and side chain all-atom configurations consistent with the coarse-grained Cα trace. We train DiAMoNDBack over 65k+ structures from Protein Data Bank (PDB) and validate it in applications to a hold-out PDB test set, intrinsically-disordered protein structures from the Protein Ensemble Database (PED), molecular dynamics simulations of fast-folding mini-proteins from DE Shaw Research, and coarse-grained simulation data. We achieve state-of-the-art reconstruction performance in terms of correct bond formation, avoidance of side chain clashes, and diversity of the generated side chain configurational

states. We make DiAMoNDBack model publicly available as a free and open source Python package.

8.2 Introduction

Coarse-grained molecular models of proteins can substantially reduce the cost of molecular dynamics simulations and permit access to time and length scales and direct simulation of long-time processes such as folding, aggregation, and self-assembly that are inaccessible to all-atom molecular simulations [501, 120, 502, 503, 132, 131, 504, 505, 506]. Coarse-graining achieves these accelerations by selectively clumping groups of atoms into super-atoms, or coarse-grained 'beads' and deriving an associated coarse-grained force field with which to propagate the system dynamics. Eliminating degrees of freedom in the coarse-grained representation leads to computational accelerations associated with the reduced cost of tracking fewer particles, the possibility of larger numerical integration time steps, and smoothening of the underlying free energy landscape that accelerates phase space exploration of the coarse-grained system. Applications of coarse-graining to biomolecular dynamics have a rich history commencing with the pioneering work of Levitt and Warshel [507] and have led to a plethora of modern-day coarse-grained force fields such as MARTINI [124, 194, 508], SPICA [509], Rosetta [510], PACE [511], CABS [512], AWSEM [513], and Upside [514]. In recent years, there has been an explosion of interest in machine-learned coarse-grained potentials [515, 409, 144, 152, 516, 162, 517, 518, 519, 520] that can be constructed from all-atom simulation data in a bottom-up fashion by rigorous variational techniques such as force matching [521, 522] or relative entropy minimization [523].

The primary concession of coarse-graining is, of course, a loss of atomistic detail that can be important in many applications such as determining atomistic contacts in protein-protein or protein-ligand interactions [524] or in downstream *ab initio* calculations that require atomistic detail to compute properties such as dipole moments or NMR spectra [387].

Backmapping is the process of reintroducing the lost degrees of freedom into a coarse-grained representation. This procedure can be conceived as a super-resolution task going from a coarse-grained to an atomistic geometry. The intrinsic loss of resolution in constructing a coarse-grained model means that the backmapping operation is one-to-many, and the primary challenge backmapping algorithms must engage is the generation of one or more physically realistic atomistic configurations associated with a particular coarse-grained structure. Contemporary backmapping methods can typically be categorized into either rules-based or data-driven approaches. Rules-based approaches use heuristics to produce an initial guess for the atomistic structure that is then refined using geometry optimization and/or energy minimization. The initial structures can be generated by querying fragment libraries [525, 136, 115], using random arrangements [137], or geometrically-guided initialization [526, 138, 527, 528, 529, 530]. Subsequent structural refinement and/or energy minimization is often necessary as the initial structures generated with rules-based approaches introduce unphysical artifacts such as atomistic clashes and/or anomalous bonds and dihedrals [531]. The requirement to manually adjust each backmapped structure introduces significant computational cost while also inherently biasing the final atomistic structure towards the particular choice of minimization scheme [524]. Rules-based approaches also tend to be deterministic in the sense that a particular coarse-grained structure will yield a single all-atom backmapped configuration. This can be an undesirable property since they fail to capture the thermodynamic ensemble of atomistic structures faithful to the coarse-grained representation.

Data-driven techniques seek to remedy shortcomings of rules-based approaches by training neural networks to produce atomic structures conditioned on the coarse-grained geometry [141, 140, 143, 145, 142, 153, 532]. These data-driven methods benefit from higher throughput compared to rules-based approaches as the models are trained to produce better well-equilibrated geometries that do not require a second stage of refinement or energy

minimization. The most successful data-driven approaches tend to employ generative models, such as Variational AutoEncoders (VAEs) [214] and Generative Adversarial Networks (GANs) [533], that produce atomistic structures conditioned on the coarse-grained structure as a model input and can learn to produce a distribution of backmapped atomistic structures faithful to a particular coarse-graining [141, 140, 143, 142, 153, 532]. While many of these data-driven techniques have demonstrated good performance when applied to relatively small biomolecules such as alanine dipeptide and chignolin [143, 142, 153], they typically require training of bespoke models using atomistic training data and are not transferable to other molecules outside of the training data. A lack of transferability strongly limits the broader applicability of a backmapping model since training costs for one molecule cannot be amortized over other systems, and models cannot be developed for systems for which only coarse-grained trajectories are accessible and atomistic training data is either unavailable or insufficient to converge a model. Work by Stieffenhofer et al. demonstrated potential for a transferable model by training on small molecule data and applying to polymer systems with their monomer units corresponding to the small molecules [141, 140]. More recently, Yang and Gómez-Bombarelli present the first instance of a chemically transferable backmapping model designed to backmap Cα traces into full-resolution atomistic protein structures using a VAE architecture operating in the internal coordinate representation (dihedrals, angles, bond-lengths) of the protein [532]. The authors train on data from the Protein Ensemble Database (PED) [534], which largely represents intrinsically disordered (IDP) proteins, and held out four PED proteins for testing and evaluation.

In this work, we present a transferable backmapping model for proteins termed DiAMoNDBack (Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping) (Fig. 8.1). The model is based on the recently popularized class of generative Denoising Diffusion Probabilistic Model (DDPM) [154, 155]. DDPMs have demonstrated impressive performance within a number of molecular domains such as protein-ligand docking [535],

generation of molecular conformers [536, 537], learning of coarse-grained potentials [519], and protein structure generation [538, 539, 540, 541, 542]. Our model is tasked to backmap atomistic proteins from Cα traces by autoregressively generating atomistic structures in a residue-by-residue fashion from the N-terminus to C-terminus of the chain. Backmapping is performed by aligning each residue into a canonical alignment that permits us to directly generate the Cartesian coordinates of each backmapped atom in a rotationally and translationally invariant reference frame that avoids the need for costly data augmentations otherwise required to implicitly learn insensitivity to rigid atomic rotations and translations [153, 156]. Empirically, we found this approach to be simpler and more stable than operating in an internal coordinate basis or employing equivariant architectures. Structure generation for each residue is conditioned on the Cα trace and any previously backmapped residues within the local neighborhood. The full protein structure is assembled by stitching together the backmapped residues along the coarse-grained Cα backbone. Local conditioning accounts for the structure of backmapped residues that are spatially proximate (i.e., have been brought into proximity by a secondary structural element, the tertiary fold, or quaternary complex) but which may be distantly separated in the primary structure (i.e., amino acid sequence). Importantly, the local and autoregressive nature of our model makes it transferable between proteins, and the stochastic nature of the denoising diffusion process means that the model generates an ensemble of backbone and side chain configurations consistent with the coarse-grained Cα trace. This means that we can both amortize the training cost of the model by applying it to arbitrary proteins outside of the training data, apply it to coarse-grained simulation trajectories for which no accompanying all-atom training data exists, and generate multiple physically-consistent realizations of the backbone and side chain configurations.

We train DiAMoNDBack over 65k+ structures from the curated from the ProteinNet [157, 158] database containing structures from the Protein Data Bank (PDB) [159, 160]

to expose the model to a rich variety of local residue environments and construct a general-purpose generative model for backmapping protein structures from Cα traces. We validate our model in applications to a hold-out ProteinNet test set, and demonstrate the transferability of the model in applications to intrinsically-disordered protein structures from the Protein Ensemble Database (PED) [534], molecular dynamics simulations of fast-folding mini-proteins from DE Shaw Research [161], and coarse-grained simulation data generated from bespoke coarse-grained potentials by Majewski et al. [162]. We benchmark against the PULCHRA rules-based approach developed by Rotkiewicz and Skolnick [530] and the data-driven VAE-based GenZProt model developed by Yang and Gómez-Bombarelli [532]. We achieve state-of-the-art reconstruction performance in terms of (i) correct formation of bonds, (ii) avoidance of side chain steric clashes, and (iii) diversity of the generated side chain configurational states. Contrary to the rules-based approach we do not suffer from a deterministic backmapping to a single structure [530], and we better reproduce the natural distribution of side chain configurational states by avoiding the mode-collapse associated with the VAE-based approaches [532]. One drawback of the model is that the DDPM generation process is relatively slow, making it approximately 10$\times$ slower than GenZProt and 100$\times$ slower than PULCHRA. We make DiAMoNDBack model publicly available to the community as a free and open source Python package.

8.3 Methods

Given a coarse-grained Cα trace of a protein $\mathbf{x} = [x_1|x_2|\dots|x_{n-1}|x_n] \in \mathbb{R}^{n\times 3}$ with $x_i \in \mathbb{R}^{1\times 3}$ representing the spatial localization of each of the $i = 1\dots n$ Cα beads and $|$ designating concatenation, the backmapping process seeks to learn the distribution of atomistic structures $p(\mathbf{X} \in \mathbb{R}^{N\times 3}|\mathbf{x})$ containing N atoms consistent with and conditioned on the coarse-grained structure $\mathbf{x}$. We choose frame this reconstruction task in an autoregressive formulation where we generate the atomistic structure $\mathbf{X} = [X_1, X_2, \dots, X_{n-1}, X_n]$ residue-

by-residue, where $X_i \in \mathbb{R}^{p_i \times 3}$ represents the p_i atoms associated with residue i such that $\sum_{i=1}^{n} p_i = N$. By decomposing the full distribution into a product of conditional distributions $p(\mathbf{X}|\mathbf{x}) = p(X_1|\mathbf{x})p(X_2|\mathbf{x}, \{X_1\})p(X_3|\mathbf{x}, \{X_1, X_2\}) \dots p(X_n|\mathbf{x}, \{X_1, X_2, \dots, X_{n-2}, X_{n-1}\})$ we simplify the learning problem to backmapping individual residues $p(X_i|\mathbf{x}, \{X_1, X_2, \dots, X_{i-2}, X_{i-1}\})$ in an autoregressive fashion rather than attempting to generate the full protein structure in one-shot. We solve this learning problem by training an autoregressive denoising diffusion probabilistic model to generatively restore atomistic detail in a residue-wise fashion from the N to C-terminus of the protein chain. We pass conditional information on the global Cα trace and any previously backmapped residues using a conditional U-net architecture.

In the following sections, we provide details of our mathematical formalism and implementation, including the DDPM process and loss function, conditional U-net architecture, residue featurization and alignment, treatment of N- and C-terminal residues, training and testing data curation, and our evaluation metrics.

8.3.1 Denoising diffusion probabilistic model (DDPM) formalism

Denoising Diffusion Probabilistic Models (DDPMs) are latent variable models employing a neural network model to gradually denoise samples drawn from a simple, easy-to-sample-from distribution, frequently an isotropic Gaussian, into denoised samples indistinguishable from the training distribution [154, 155, 543]. We train a DDPM to estimate the conditional distribution for autoregressive residue backmapping $q_\theta(X_i|\mathbf{x}, \{X_1, X_2, \dots, X_{i-2}, X_{i-1}\}) \approx p(X_i|\mathbf{x}, \{X_1, X_2, \dots, X_{i-2}, X_{i-1}\})$ by defining an M-step discrete noising process that takes the data samples $p(X_i, t_0|\mathbf{x}, \{X_1, X_2, \dots, X_{i-2}, X_{i-1}\}) = p(X_i|\mathbf{x}, \{X_1, X_2, \dots, X_{i-2}, X_{i-1}\})$ defined at $t = t_0$ and transforms them by adding noise until $p(X_i, t_M|\mathbf{x}, \{X_1, X_2, \dots, X_{i-2}, X_{i-1}\}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at $t = t_M$ resembles isotropic Gaussian noise (Fig. 8.1b). While the noising process is straightforwardly defined, the challenge lies in approximating the reverse denoising process $q_\theta(X_i, t_j|\mathbf{x}, \{X_1, X_2, \dots, X_{i-2}, X_{i-1}\}, t_{j+1})$ such that starting at $t = t_M$ and progres-

sively applying q_θ through M denoising steps reverts isotropic Gaussian noise to realistic-looking samples from the training distribution. In practice, a U-net architecture [544] is frequently used to learn $q_\theta(X_i, t_j|\mathbf{x}, \{X_1, X_2, \ldots, X_{i-2}, X_{i-1}\}, t_{j+1})$ (Fig. 8.1c). The network can be conceived as a noise predictor with each training sample querying a random time step $t_j \in \{t_0, t_1, \ldots, t_{M-2}, t_{M-1}\}$, corrupting the sample with the associated level of noise, and then learning to predict the noise conditioned on the corrupted sample and the time step t_j.

The M-step noising and denoising diffusion processes are formally Markov chains, and it is useful to frame these forward and backward operations in terms of intermediate latent representations $\{\mathbf{z}_{t=0}, \mathbf{z}_{t=1}, \mathbf{z}_{t=2}, \ldots, \mathbf{z}_{t=M}\}$. The fixed forward noising diffusion process $p(\mathbf{z}_{t=i}|\mathbf{z}_{t=i-1})$ gradually adds Gaussian noise to samples drawn from the training data $\mathbf{z} = \mathbf{z}_{t=0} \sim \mathbf{X}$ in M discrete steps concluding with $\mathbf{z}_{t=M} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ resembling isotropic Gaussian noise. The learned backward denoising diffusion process $q_\theta(\mathbf{z}_{t=i-1}|\mathbf{z}_{t=i})$ transforms isotropic Gaussian noise into samples indistinguishable from the training data distribution. The forward diffusion processes is defined to have a Gaussian form with $p(\mathbf{z}_{t=i}|\mathbf{z}_{t=i-1}) \sim \mathcal{N}(\sqrt{1-\beta_{t=i}}\mathbf{z}_{t=i-1}, \beta_{t=i}\mathbf{I})$ where $\{\beta_{t=1}, \beta_{t=2}, \ldots, \beta_{t=M}\}$ define an ascending set of variables bounded between 0 and 1 known as the variance schedule. A sample $\mathbf{z}_{t=i}$ at an arbitrary diffusion step $t = i$ can therefore be drawn from a conditional Gaussian distribution with mean $\mu_{t=i} = \sqrt{1-\beta_{t=i}}$ and variance $\sigma^2_{t=i} = \beta_{t=i}$. Assuming the reverse processes has the same Gaussian form we seek to learn $q_\theta(\mathbf{z}_{t=i-1}|\mathbf{z}_{t=i}) \sim \mathcal{N}(\mu_\theta(\mathbf{z}_{t=i}, t=i), \sigma^2_{t=i}\mathbf{I})$. Here the reverse time diffusion processes is conditioned on the time step with a learnable mean function μ_θ and an untrained time dependent variance $\sigma^2_{t=i}\mathbf{I} = \beta_{t=i}\mathbf{I}$ [154]. Our learned model, which will be encapsulated within a U-net neural network, is essentially tasked with learning the mean function μ_θ of this reverse time denoising diffusion process.

An important observation in formulating the objective function for learning the reverse processes relies on noticing that the combination of the forward process p and the reverse pro-

cess q_θ resembles a hierarchical variational autoencoder (VAE) [543]. This permits the loss for the full reverse process $\mathcal{L}$ to be decomposed as a sum of per-time-step terms $\mathcal{L} = \mathcal{L}_1 + \ldots \mathcal{L}_M$, where each term in the sum can be written as a Kullback-Leibler (KL) divergence between two Gaussian distributions and expressed in closed form. During training we can optimize the loss function by leveraging the Gaussian construction of the forward processes to define $p(\mathbf{z}_{t=i}|\mathbf{z}_{t=0}) = \mathcal{N}(\sqrt{\bar{\alpha}_{t=i}}\mathbf{z}_{t=0}, (1-\bar{\alpha}_{t=i})\mathbf{I})$, with $\alpha_{t=i} = 1-\beta_{t=i}$ and $\bar{\alpha}_{t=i} = \prod_{j=1}^{j=i} \alpha_{t=j}$ that permits us to sample $\mathbf{z}_{t=i}$ at an arbitrary time point $t = i$ conditioned on only the original sample $\mathbf{z}_{t=0} = \mathbf{z} \sim \mathbf{X}$. As shown by Ho et al. [154], the mean function can be reparameterized $\mu_\theta(\mathbf{z}_{t=i}, t = i) = \frac{1}{\sqrt{\alpha_{t=i}}}\left(\mathbf{z}_{t=i} - \frac{\beta_{t=i}}{\sqrt{1-\bar{\alpha}_{t=i}}}\epsilon_\theta(\mathbf{z}_{t=i}, t = i)\right)$ introducing $\epsilon_\theta(\mathbf{z}_{t=i}, t = i)$ as the noise predicting network and recasting the learning problem as predicting the noise added to corrupt the original sample $\mathbf{z}_{t=0}$. Each component of the loss may then be conveniently written as $\mathcal{L}_{t=i}$ used to train our DDPM may then be written as a simple regression loss,

$$\mathcal{L}_{t=i}(\mathbf{z}_{t=0}, t = i) = ||\epsilon - \epsilon_\theta(\mathbf{z}_{t=i}, t = i)||^p = ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_{t=i}}\mathbf{z}_{t=0} + \sqrt{1-\bar{\alpha}_{t=i}}\epsilon, t = i)||^p \quad (8.1)$$

where we employ $p = 1$ corresponding to an L1-loss, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{z}_{t=0} \sim \mathbf{X}$ is a sample from the training data. The training procedure therefore involves gathering a sample $\mathbf{z}_{t=0}$, selecting a random time step t between 1 and M and sampling Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ used to corrupt the original sample given the selected time step t, and then training the noise predicting neural network $\epsilon_\theta(\mathbf{z}_{t=i}, t = i)$ to predict the added noise ϵ based on the corrupted sample $\mathbf{z}_{t=i}$. At inference time, the trained noise prediction network ϵ_θ specifies the mean function μ_θ used to define the reverse denoising diffusion processes that progressively transforms randomly sampled isotropic Gaussian noise into realistic samples from the training distribution.

8.3.2 U-net neural network architecture

A U-net architecture [544] is selected to serve as the noise prediction neural network $\epsilon_\theta(\mathbf{z}_{t=i}, t = i)$ within our DDPM process. The U-net accepts corrupted samples $\mathbf{z}_{t=i}$ as inputs and is trained to output the predicted noise ϵ_θ conditioned on the particular time step $t = i$ (Fig. 8.1c). We employ a U-net architecture composed of 1D convolutional layers adapted from the implementation of Wang et al. [545] who employed a similar DDPM model to generate backbone angles of biomolecules conditioned on temperature. As we are interested in autoregressive generation of residues X_i, we condition on structural information from the Cα trace $\mathbf{x}$ and previously backmapped residues $\{X_1, X_2, \ldots, X_{i-2}, X_{i-1}\}$. We incorporate this conditioning information by allowing only a subset of the input representation to vary throughout the diffusion process and treating the remaining components of the input as fixed [155]. In this manner, the regression task training the noise prediction neural network ϵ_θ as defined in Eqn. 8.1 only receives contributions from the components that are being noised, with the other fixed components serving as conditioning information provided to the network. Details on the construction and generation of the input to U-net architecture detailing which components of the representation are noised and which are treated as fixed conditioning variables is further explained Sec. 8.3.3.

We employ a four-layer U-net architecture with an initial channel dimension of 32. The channel dimension is multiplied by a factor of two for each subsequent layer while the feature dimension is reduced by a factor of two. A weight-standardized 1D convolution [546] and residual connections are used as implemented in the original DDPM model implementation `https://github.com/lucidrains/denoising-diffusion-pytorch`. A cosine noise scheduler [547] is used for adding noise throughout the diffusion time-steps. We use 50 diffusion steps and 1,000,000 training steps for our production models trained on PED and PDB data, and apply an additional 500,000 training steps when fine-tuning. An exponential learning rate scheduler is used during the initial 1,000,000 step training run with an initial

learning rate of 3×10^{-4} is annealed each step by a factor γ=0.9999965988084024 such that after 1,000,000 steps the learning rate is decayed to $1\times10^{-5} \approx 3\times10^{-4} \times \gamma^{1000000}$. A batch size of 128 is used for training and we also maintain a separate exponential moving average of the U-net parameters that start tracking after 2,000 steps and is updated every 10 training steps employing a 0.995 decay rate. The input representation is scaled using a min/max scaler where each component of the 19×3 residue prediction block and one-hot residue type encoding is scaled according to the maximum and minimum values present in the training dataset for each associated feature dimension. The local environment block containing structural information on the 14 spatially most proximate residues to the current residue being decoded is then normalized by scaling values at -20/+20 Å to -1/+1, which serves to provide the same numerical normalization for features across all components of the local environment block.

8.3.3 Autoregressive backmapping of Cα traces

To backmap a Cα protein trace we adopt an autoregressive approach that performs successive residue-by-residue backmapping proceeding from the protein N-terminus to C-terminus. Although we could have adopted alternative autoregressive orderings, such as random or data-driven ordering, empirically we find N-to-C backmapping performs well and displays no correlation in generating clashing residues as a function of position along the chain. We also hypothesize that maintaining an N-to-C ordering during both training and inference means that the partially backmapped environment experienced by the model during inference is more representative of the training data. For a contiguous protein containing n residues, this process begins at the residue adjacent to the N-terminus (X_2) and ends at the residue adjacent to the C-terminus (X_{n-1}). Terminal residues are handled separately after first backmapping all of the internal residues as described in Sec. 8.3.4. We define this process as a transformation from a coarse-grained configuration $\mathbf{x}$ to an entirely all-atom

configuration $\mathbf{X}$, with intermediate partially decoded configurations $\mathbf{X}_i$ representing a structure with all residues up to i decoded. The backmapping process for each residue i proceeds in four steps:

1. **Alignment:** Superpose the target and adjacent backbone Cα atoms (Cα_{i-1}, Cα_i, Cα_{i+1}) to a reference configuration and transform all previously decoded residues in the coordinate set to this canonical orientation (see Sec. 8.3.3): $\tilde{\mathbf{X}}_i = \mathbf{X}_i \cdot R_i + T_i$, where R_i and T_i represent rotation and translation matrices that place residue i into this frame.

2. **Featurization:** Build the input to the DDPM by extracting (i) the identity and geometry of the target residue, (ii) the local atomic environment of previously backmapped residues, and (iii) the Cα trace (see Sec 8.3.3).

3. **Inference:** Perform DDPM inference to generate the backbone and side chain atoms X_i.

4. **Re-alignment:** Incorporate the predicted residue X_i into the coordinate set and reverse the alignment transformation to reset the structure into the original orientation now including the newly backmapped residue $\mathbf{X}_i \leftarrow R_i^{-1} \cdot (\tilde{\mathbf{X}}_i - T_i)$.

This process is iterated until all internal residues are backmapped. During the initial N-to-C reconstruction of internal residues only the termini Cα positions are exposed to the network. Multi-chain proteins are decoded sequentially in the order in which they appear in the structure file in the same N-to-C fashion with information from backbapped residues in neighboring chains treated analogously to inter-chain residues when building the local environment information. As in the case of single-chain proteins, N-termini and C-termini are backmapped once all internal residues have been placed in all chains.

Alignment

A key component in our backmapping routine is generating predictions for each residue within a canonical reference frame. By adopting a fixed alignment we avoid the need for computationally costly data augmentation required to learn insensitivity to rigid translations and rotations of the atomic positions. Each residue at a given position i along the sequence is aligned by superposing the neighboring triplet of Cα atoms (Cα_{i-1}, Cα_i, Cα_{i+1}) to a reference structure. The reference structure is derived from the average angle formed between the three Cα atoms (106°) and the average Cα-Cα distance (3.81 Å) across all PDBs in our training set. The reference triplet is oriented to lie in the xy plane with a center-of-mass at the origin of the reference coordinate frame, such that the associated side chain atoms are aligned to approximately point along the positive y-axis. The Kabsch algorithm [548] is used to compute the rotation R_i and translation T_i transformations that superpose the Cα-triplet at residue i to the reference structure using the singular value decomposition implemented in Biopython [549, 550].

Featurization and conditioning

We expose the spatial arrangement of the residue atoms to our model in the form of three-dimensional Cartesian coordinates within the reference frame. When input to the U-net architecture this Cartesian-style representation is ultimately flattened into a 1D input that is partitioned into three blocks: (i) the residue prediction block, (ii) the local environment block, and (iii) the residue type one-hot-encoding (OHE) (Fig. 8.1c).

For the residue prediction block, the target residue at position i comprises all heavy (i.e., non-hydrogen) atoms between the backbone Cα_{i-1} and Cα_{i+1} atoms. This includes the nine backbone atoms – the three C, N, O atoms on either side of Cα_i plus Cα_{i-1}, Cα_i, and Cα_{i+1} – and between 0 and 10 side chain atoms. (Considering only the 20 naturally occurring amino acids, glycine is the smallest and contains zero heavy side chain atoms while

tryptophan is the largest and contains ten.) The side chain atom positions are populated for each residue according to the SidechainNet canonical ordering scheme, and we maintain a fixed-sized input for each residue type by padding any unutilized positions within the allocated 10-atom block with zeros.

The local environment block then contains information on neighboring residues comprising the local atomic environment of each target residue being predicted. This block contains the coordinates of the N residues nearest to the target Cα_i sorted by the distance of the closest atomic contact. Clearly more proximate residues play a more important role in informing accurate placement of the backmapped atoms to avoid spurious clashes and we can take advantage of this locality principle to accelerate model training and deployment. Empirically, we found model performance to have effectively converged at N=14. We maintain the N-to-C backmapping ordering during both training and inference. During training, this means that when predicting on residue i, residues $j < i$ are represented in all-atom resolution whereas residues $j > i$ are represented in coarse-grained Cα resolution. A consequence of this are that the ordinal ordering of residues based on the proximity of the closest atom is judged based on any atom within the residue for $j < i$ and only on the Cα for $j > i$. Each of the N=14 slots within the local environment block contains sufficient space to represent 11 atoms – the Cα and up to 10 side chain atoms. Residues selected to form part of the local environment block with indices $j < i$ are represented atomistically with unused side chain positions padded with zeros, whereas those with indices $j > i$ only populate the Cα position and all other positions are padded with zeros.

The residue prediction block and local environment block are concatenated along the atom index and flattened to form a $(19{+}N{\times}11){\times}3$ 1D array. Finally, a 20-dimensional residue type one-hot encoding is appended to this flattened array of structural information to encode the target residue identity as one of the 20 naturally occurring amino acids.

Mutable (predicted) and immutable (conditioning) components of the featurization

Our featurization approach is used to generate training samples for the DDPM from all-atom protein structures, but certain components of the featurization are fixed and serve as conditioning variables whereas others are variables we train the U-net to predict. During *training*, the non-Cα atom positions pertaining to the target residue i are corrupted with noise. The Cα triplet positions (Cα_{i-1}, Cα_i, Cα_{i+1}), the positions of all atoms in the local environment block, and the one-hot encoding block remain unaltered and are fixed and immutable in all diffusion time steps. During *inference*, the Cα trace together with any already decoded residues is used to build the fixed conditioning information, and the remaining components corresponding to the unknown backbone and side chain atoms are initialized with random Gaussian noise. This partially noised representation initiates the DDPM denoising process with the fixed conditioning components held unchanged throughout each diffusion time step and the U-net operates only on the atom positions that we wish to backmap. All immutable conditioning components and zero-padded components are masked in evaluation of the loss function (Eqn. 8.1). After completing the M steps of the denoising processes, the predicted atomistic structure of the target residue i generated by the DDPM is extracted and incorporated into the coordinate set of the partially backmapped protein structure. In our autoregressive inference scheme, this updated coordinate set is then used when building the input for predicting the next residue $(i + 1)$ along the chain. As each adjacent prediction block contains three overlapping C, N, and O backbone atoms, we choose to update these atom positions with the prediction from the most recently generated residue. Therefore the backmapping of residue $(i+1)$ will overwrite the C, N, and O backbone atoms predicted in the inference of residue i.

8.3.4 Terminal residues

Terminal residue alignment

Our backmapping procedure for internal residues relies on placement of each target residue i within a reference frame based on the Cα triplet (Cα_{i-1}, Cα_i, Cα_{i+1}). Since terminal residues lack one or other of these Cα coordinates we backmap the termini after first backmapping all of the internal residues and adopt a different set of atoms using the nearest backbone carbon with which to perform the reference alignment: (Cα_i, C$_i$, Cα_{i+1}) for N-termini and (Cα_{i-1}, C$_{i-1}$, Cα_i) for C-termini. The only other modifications to our internal residue backmapping process are that (i) during inference we perform an additional translation to guarantee coincidence of the Cα positions during our re-alignment step, (ii) we pad the missing atom positions within the residue prediction block – (Cα_{i-1}, C$_{i-1}$, O$_{i-1}$, N$_i$) for N-termini and (C$_{i+1}$, O$_{i+1}$, Cα_{i+1}) for C-termini – with zeros, but do not mask the backbone atom positions in calculation of the loss function to teach the network to differentiate between an internal and terminal residue prediction task, (iii) all residues in the local environment block are represented in all-atom resolution since the terminal residues are decoded last, and (iv) terminal residues are prohibited from appearing in one another's local environment block.

8.3.5 Data Curation

We collated four datasets for model training and testing: (i) PED – structural ensembles of primarily intrinsically disordered proteins [534], (ii) PDB – structures drawn from the RCSB Protein Data Bank [159, 160], (iii) DES – D.E. Shaw Research molecular dynamics simulations of fast folding mini proteins [161], and (iv) CG – coarse-grained simulations conducted by Majewski et al. [162].

PED. For the purpose of comparison to the GenZProt model of Yang and Gómez-

Bombarelli [532] we train over the Protein Ensemble Database (PED) [534], which contains structural ensembles of proteins including many intrinsically disordered proteins (IDPs). We discard three sequences – PED00125e000, PED00126e000, and PED00161e002 – that contain non-canonical amino acids. Following Yang and Gómez-Bombarelli [532], we adopted four PED proteins – PED00151ecut0, PED00090e000, PED00055e000, and PED00218e000 – containing 20-140 frames and including one two-chain protein (PED00218e000) as our test set, and employed the remaining data as out training set. Since GenZProt does not support backmapping of terminal residues, in the interest of making head-to-head comparisons with this model we report all quantitative analyses and comparisons restricted to internal residues only.

PDB. Our primary production-level model was trained over protein structures collated from the Protein Data Bank (PDB) [159, 160] held in the SidechainNet database [158] extension of the ProteinNet database [157] that itself builds on the data for the biennial Critical Assessment of protein Structure Prediction (CASP) challenges. For this PDB training dataset we retain a majority of configurations but filter according to a number of criteria. We discarded any configuration that had four or more disconnected chains or contained a chain less than five residues long, leaving 98,665/103,716 sequences. Next, we remove any structures that include incomplete side-chain coordinates for any non-terminal residues resulting in the elimination of an additional 32,403 structures. Finally, we discard 2,562 problematic structures containing one or more malformed neighboring Cα-Cα distances lying outside the range of 2.7-4.1 Å, where our cutoffs were informed by collating histograms of neighboring Cα-Cα distance distributions to identify outliers. We retained a total 63,700 structures containing over 13 million individual residue training samples. For the PDB test set, we employ the same test set as that provided by the ProteinNet database for the CASP12 blind structure prediction challenge. We filter the test set consistent with the criterion we used to filter the training data set removing structures that were missing some portion of the side

chain atoms. After data cleaning, we extracted 24 test set proteins ranging in size from 60 residues to 599 residues that include eight multi-chain proteins.

DES. The PED and PDB training data comprise static protein structures derived, respectively, from simulation and experiment. These training examples are therefore expected to correspond to structures lying in local or global minima of the configurational free energy landscape. We were interested to test if the performance of our model would improve with additional fine tuning on all-atom molecular dynamics (MD) trajectories containing a greater diversity of configurations including metastable states and transition states. We refined the model trained over the PDB training data by subjecting it to additional training over MD trajectories of 11 fast-folding mini proteins conducted by D.E.Shaw Research (DES) [161]. We aggregated one complete trajectory of each protein, except for Villin (2F4K) that contains a non-canonical amino acid residue, and strided each trajectory into 10,000 equally spaced frames. Using the procedure described in Sidky et al. [445] we separated these frames into 100 contiguous chunks and randomly shuffled these chunks to form an 80/20 train/test split for each protein. In this way, the model is exposed to configurations across the full trajectory, but the test set retains regions that are distinct and not temporally neighboring frames in the training set. To compile the fine-tuning dataset we combined the training splits from each of the 11 proteins totaling 88,000 frames with an aggregate of 3,860,000 distinct residue training samples. When constructing our fine-tuning data set we find that performance for terminal residue prediction can substantially improve by over-representing terminal residue training examples by repeating their occurrences in the training data set. Empirically we find there exists a trade-off where performance on internal residues begins to suffer if termini are too over-represented, which we find to occur at $\sim$20$\times$ more terminal residue representation. For the fine-tuned models presented here, we train with termini examples occurring 5$\times$ more than if no data augmentation were performed–which we find to be a good balance in both resolving terminal residues and internal residues with high fidelity.

CG. Finally, we collected coarse-grained trajectories from the work of Majewski et al. [162] for three proteins of varying size: 1FME (28 residues), PRB (47 residues), and A3D (73 residues). In contrast to previous data sets, these are simulations carried out using a bespoke coarse-grained force field and contain no all-atom reference structures so serve purely as testing data and a means to evaluate the out-of-domain generalization and transferability of our model. We performed even striding across all available 32 trajectories for each protein to obtain 2,000 frames for each system.

A visual comparison of the training and test data is presented in Fig. 8.2. In Fig. 8.2a we illustrate the distribution of sequence lengths. The PED training data comprises proteins from 13-260 residues in length with a total of 96 sequences and 9788 configurations. The DES data contains 11 sequences ranging from the small chignolin protein containing just 10 residues to the large λ-repressor containing 80 residues for a total of 88,000 structures. The PDB training data set contains proteins ranging from 5-2082 residues and totals 65,360 structures. Alongside the sequence diversity, we can also conceptualize the structural diversity of our training datasets by visualizing the distribution of our structures in the space of alpha-helical and beta-sheet content (Fig. 8.2b). The PED data tends toward regions with relatively low alpha-helix and beta-sheet content, reflecting the intrinsically disordered character of the dataset. Conversely, the PDB data spans a wide range of alpha-helix and beta-sheet content indicative of the more globular and ordered structures that originate from crystallography data. The DES data, while containing the least sequence diversity, covers a wide range of structural diversity due to the sampling of both folded and unfolded configurations in the MD simulations. In Fig. 8.2c, we illustrate the hold out test set proteins corresponding to each of the three training datasets in the space of their alpha-helix and beta-sheet content, along with visualizations of selected structures.

8.3.6 Evaluation metrics

We define three metrics to evaluate the quality of our backmapping predictions that measure (1) the quality of the reconstructed atomistic bond graph, (2) the degree of non-bonded steric clashes between residues, and (3) the diversity of atomistic structures produced.

Bonds. The quality of the bond graph for our backmapped structures is determined by calculating the percentage of bonded atom distances that lie within 10% of the bond distance in the reference atomistic structure. For the coarse-grained trajectories in the CG test data that do not possess corresponding atomistic reference trajectories, we use the average values for bond lengths from the test set DES data for the associated protein as the reference. The bond score varies between 0-100% with higher values indicating better performance.

Clash. A major failure mode for backmapping is the construction of side-chain placements that result in unphysical steric clashes. We quantify the degree of steric clashes by computing the fraction of residues in which one or more atoms lie within 1.2 Åof the atoms of another residue. For neighboring residues in the protein backbone, only clashes between the side chain atoms are considered. The clash score varies between 0-100% with lower values indicating better performance.

Diversity. A single coarse-grained Cα trace is consistent with multiple all-atom configurations. A desirable trait of a backmapping model is its capacity to generate an ensemble of all-atom predictions each of which is compatible with the Cα trace and itself contains well-formed bonds and avoids steric clashes. Previous work has used the root-mean squared distance (RMSD) between samples as a metric for generative diversity, where a high RMSD between samples indicates high diversity [142]. However, a low RMSD between the samples and the atomistic reference has also been used to show good adherence to conditioning and faithful reconstruction [532]. We combine these two concepts to posit that the single all-atom reference structure should be indistinguishable from the distribution of generated configurations, and therefore the average pairwise RMSDs between (i) all generated samples

and the reference and (ii) all generated samples with themselves should be approximately equal. Based on this framework we define a generative diversity score DIV as,

$$\begin{aligned} RMSD_{ref} &= \frac{1}{G}\sum_{i}^{G} RMSD(\mathbf{X}_i^{gen}, \mathbf{X}^{ref}) \\ RMSD_{gen} &= \frac{2}{G(G-1)}\sum_{i}^{G}\sum_{j}^{(i-1)} RMSD(\mathbf{X}_i^{gen}, \mathbf{X}_j^{gen}) \\ DIV &= 1 - \frac{RMSD_{gen}}{RMSD_{ref}} \end{aligned} \tag{8.2}$$

where G is the number of generated samples $\{\mathbf{X}_1^{gen}, \ldots, \mathbf{X}_G^{gen}\}$ conditioned on the Cα trace possessing a single reference configuration $\mathbf{X}^{ref}$. We find that

$$RMSD_{gen} \approx \frac{2}{(G-1)\sqrt{G}}\sum_{i}^{G} RMSD(\mathbf{X}_i^{gen}, \overline{\mathbf{X}^{gen}})$$

, where $\overline{\mathbf{X}^{gen}} = \frac{1}{G}\sum_i^G \mathbf{X}_i^{gen}$ is the mean generated configuration. This approximation, approximating $\frac{2}{(G-1)\sqrt{G}} \approx \frac{1}{G}$ and subject to the tightness of the inequalities used in its derivation, allows us to roughly interpret $RMSD_{gen}$ as the average RMSD of the generated configurations around their own mean. This construction is instructive as it allows us to then interpret the diversity score DIV as a comparison of the average RMSD of the generated configurations around the reference configuration $\mathbf{X}^{ref}$ relative to that around their own mean $\overline{\mathbf{X}^{gen}}$. In general, one would anticipate a tighter distribution around the mean of a distribution than any other imposed point, such that we would expect $RMSD_{gen}$<$RMSD_{ref}$. This, however, is not a strict inequality and one may expect it to be violated, particularly for small values of G that produces noisy estimates of $\overline{\mathbf{X}^{gen}}$. Nevertheless, our empirical calculations show that in the present applications that it generally holds with only occasional violations, and that our calculated diversity scores approximately lie on the interval [0,1]. Diversity scores of unity are obtained for deterministic backmapping since $RMSD_{gen}$=0,

while diversity scores close to zero are achieved for $RMSD_{ref} \approx RMSD_{gen}$ and are indicative of better model performance. We note that the diversity metric cannot be calculated for the coarse-grained data that lack reference atomistic structures, but we can still compute $RMSD_{gen}$ as a proxy metric for generative diversity. Finally, we support our quantitative diversity score analysis with inspection of a number of representative generative samples to visually assess the quality and diversity of the backmapped ensemble.

8.4 Results and Discussion

We train and evaluate DiAMoNDBack on four train/test data sets comprising a variety of protein structures drawn from conformational ensembles, crystal structures, and molecular dynamics trajectories. We benchmark against the recently published data-driven approach by Yang and Gómez-Bombarelli that trains a graph neural network variational autoencoder (VAE) to generatively decode an internal coordinate representation (angles, bonds, dihedrals) of the protein structure in one-shot from a Cα trace that is then used to assemble a fully atomistic structure using a custom geometry builder [532]. We also compare against the PULCHRA rules-based approach developed by Rotkiewicz and Skolnick that generates an approximate structure using heuristics and atomic fragments followed by rotating side chain dihedrals to resolve steric clashes [530]. We choose not perform any energy minimization or molecular mechanics relaxations for any of the three models to compare and evaluate the backmapping approaches independent of any molecular force-field and to avoid the high computational cost associated with these operations for large proteins. We also note that while we provide routines for handling prediction of terminal residues, for the purposes of comparison to GenZProt that does not support termini prediction all analyses herein are performed on protein structures stripped of terminal residues.

8.4.1 PED training and evaluation

Following Yang and Gómez-Bombarelli [532], we first train and evaluate DiAMoNDBack on the PED dataset comprising conformational ensembles of intrinsically disordered proteins [534]. A numerical comparison of DiAMoNDBack and GenZProt on the test set of four held-out structures is presented in Table 8.1.

While performance is split between our approach and GenZProt in terms of bond quality, we believe bond quality may be a deceiving metric for the PED dataset due to the presence of unphysical bond lengths in the training dataset. While we found that none of the frames from the four test proteins were problematic, remaining faithful to the original data splits and including all training data frames could potentially be detrimental as our model is effectively learning to recapitulate these unphysical examples. Nevertheless, the bond quality is overall still excellent showing an average of ∼97.6% of bond lengths being within 10% of the reference values for our model.

In terms of clash score our approach outperforms GenZProt with an overall improvement of ∼50% fewer clashing residues across the four structures with superior clash scores to GenZProt on all test proteins except PED00151ecut0 where our mean clash score is slightly better than GenZProt but not outside of error. The area of greatest improvement we notice is generative diversity, where our model yields diversity scores less than 0.23 across all four test proteins while all diversity scores for GenZProt are greater than 0.85. With this PED dataset serving to provide a common benchmark between our method and GenZProt, the largely disordered character of the proteins may not be entirely representative of MD data or more globular structures represented in the PDB. We, therefore, propose training on data originating from PDB structures can provide a more broadly applicable model for transferable backmapping of generic Cα protein traces.

	Test protein			
	PED00055e000	PED00090e000	PED00151ecut0	PED00218e000
	Bonds [%]↑	Bonds [%]↑	Bonds [%]↑	Bonds [%]↑
GenZProt	97.55±0.02	95.30±0.02	**97.05±0.01**	98.06±0.01
DiAMoNDBack (PED)	97.70±0.08	**97.94±0.07**	96.76±0.05	98.07±0.10
	Clash [%]↓	Clash [%]↓	Clash [%]↓	Clash [%]↓
GenZProt	4.67±0.12	9.27±0.13	0.34±0.07	5.42±0.28
DiAMoNDBack (PED)	**2.82±0.18**	**5.04±0.68**	0.30±0.03	**1.76±0.38**
	Diversity↓	Diversity↓	Diversity↓	Diversity↓
GenZProt	0.9048±0.0004	0.888±0.002	0.8533±0.0007	0.862±0.001
DiAMoNDBack (PED)	**0.221±0.002**	**0.208±0.003**	**0.148±0.001**	**0.187±0.006**

Table 8.1: Comparison of our model trained on the PED dataset, labeled 'DiAMoNDBack (PED)', compared to the work of Yang and Gómez-Bombarelli [532], labeled as 'GenZProt'. Errors for bond and clash scores are reported as standard deviations over five generated samples averaged over frames, while diversity score errors are calculated using Jackknife resampling over generated samples according to Eqn. 8.2. The bolded values indicate superior performance if the results are outside of error.

8.4.2 PDB training with MD evaluation

We train our model on a corpus of 65k+ PDB structures evaluating performance on 24 held-out PDB proteins and MD simulation data from 11 fast-folding proteins generated by D.E.Shaw and co-workers [161]. In comparison to the PED dataset, this PDB training set contains 680× more protein sequences and 6.7× more configurations, amounting to 15× more individual residue training examples. After training on this voluminous dataset of PDB structures, we also fine-tune this generic model on MD trajectories of 11 fast-folding proteins. Better calibrating our model to a specific MD force field and class of protein sequences improves performance and provides a way to purpose small amounts of available atomistic data for more bespoke backmapping models.

In Table 8.2 a numerical comparison is presented between our PDB-trained model, the PDB-trained model fine-tuned on DES data, GenZProt [532], and the rules-based PULCHRA approach [530]. Comparisons are shown for aggregate evaluations overall 24 test set

PDB structures and 11 MD trajectories of different protein sequences alongside specific comparisons to the best and worst-performing structures and trajectories determined by clash score. We note the GenZProt model here is trained on a much smaller volume of protein structures compared to the PDB training set our model is trained on, but nevertheless serves as a comparison to another transferable backmapping approach such as the rules-based baseline. While we notice that our models produce superior bond quality to GenZProt and the rules-based baseline, a more substantial impact is seen in clash score and diversity in comparison to GenZProt for both the PDB and DES datasets. For both the aggregate PDB and DES datasets our backmapping approaches yield generated structures with an average of fewer than 0.6% clashing residues, with 4/24 test set PDB structures recording zero clashes across all five generated samples for the PDB-trained model. The rules-based approach outperforms our PDB-trained model on both PDB and DES data in terms of clash score, but after fine-tuning on DES data our model then outperforms the rules-based approach on the test-set MD trajectories. The jump in performance as a result of fine-tuning highlights the ability to produce custom models by better calibrating our generic backmapping model with available atomistic training data, while also not degrading performance on the PDB dataset.

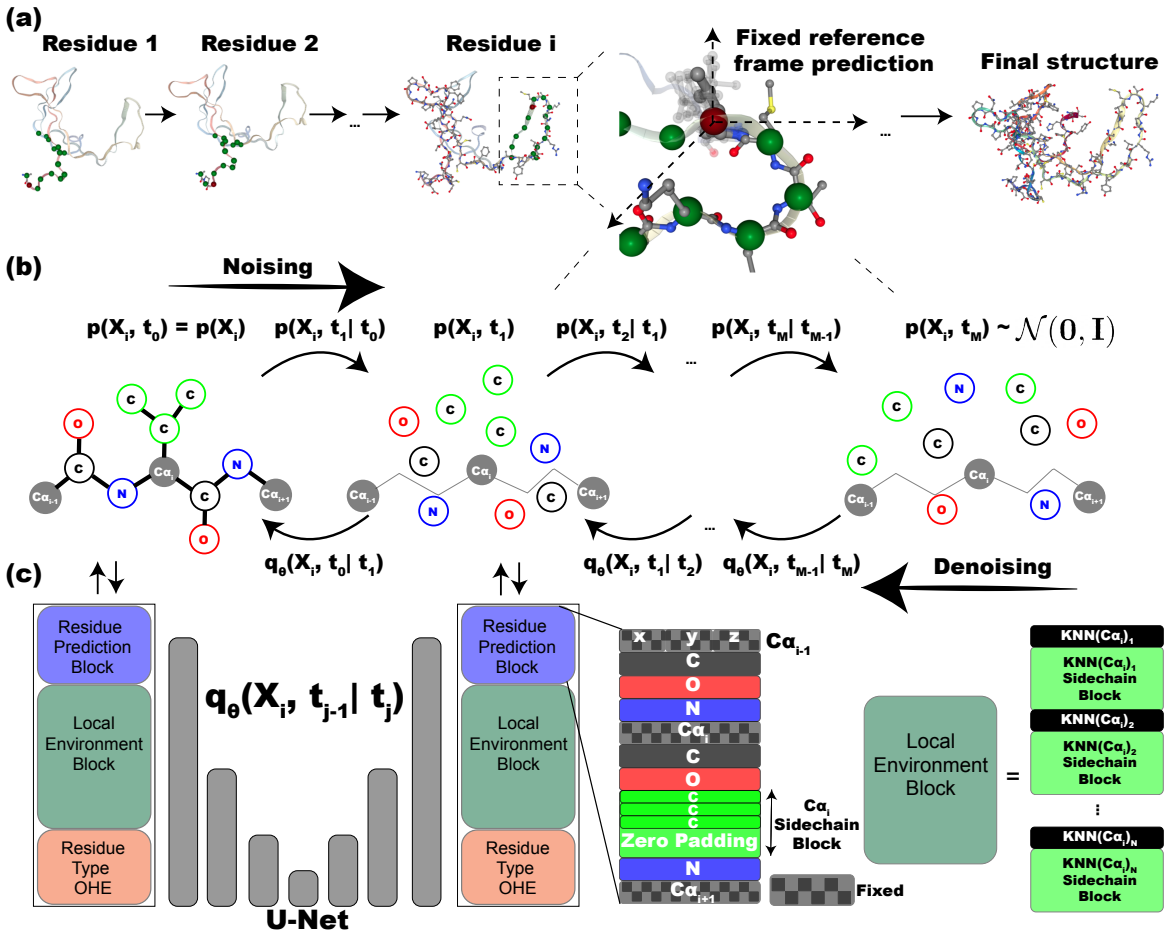


Figure 8.1: Schematic illustration of DiAMoNDBack (Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping). **(a)** Atomistic detail is restored from Cα traces by backmapping the structure residue-by-residue from the N-terminus to the C-terminus. Each residue prediction task is performed within a canonical reference frame and depends on the local environment around the target Cα (red sphere) comprising the N most spatially proximate residues to the target Cα (green spheres). **(b)** The residue prediction task is performed using a Denoising Diffusion Probabilistic Model (DDPM) that learns to reverse an M-step noising process applied to real samples. The trained DDPM model can then operate on random noise to recover realistic-looking samples. **(c)** Learning the denoising process involves training a U-net model that is designed to predict the noise added to a corrupted sample. The input to the network are Cartesian coordinate representations of the target residue, the local environment, and a one-hot encoding of the residue identity. Conditioning is achieved by only noising and regressing on components of the representation that are allowed to change throughout the diffusion steps in the residue prediction block such as the backbone C, N and O atoms and the side chain atoms. Atoms comprising the partially-decoded local environment and the Cα backbone, along with the one-hot residue type encoding only serve as conditioning information passed to the network, are fixed throughout the diffusion steps, and do not contribute to the loss.

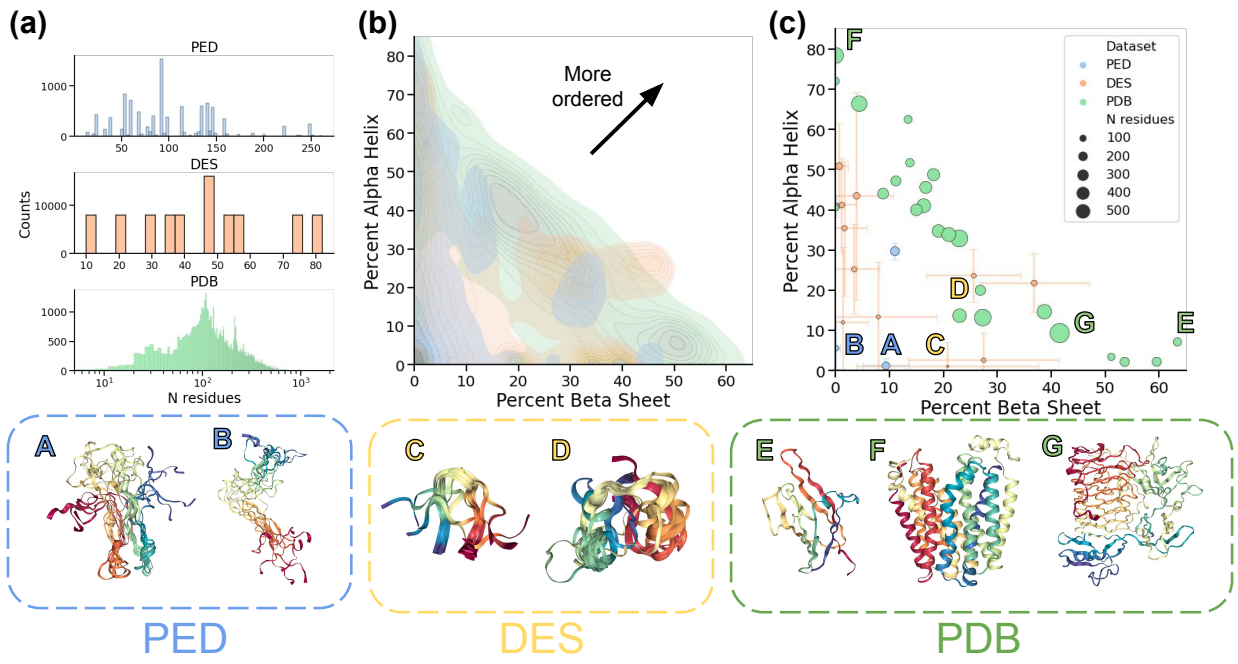


Figure 8.2: Visualization of protein structures used for training and testing in the PED, PDB, and DES training sets. **(a)** Distribution of sequence lengths. **(b)** Distribution of structural diversity in the space of average percent alpha-helix and beta-sheet content. Disordered/unstructured structures reside close to the origin. Colors correspond to the histograms in panel (a): PED = blue, PDB = green, and DES = orange. **(c)** Test set proteins for each training data set shown in the same space of alpha-helix and beta-sheet content. The size of each point corresponds to the number of residues in the protein. For the PED and DES data, multiple frames are present for each protein and error bars indicate the standard deviation across frames. Selected representative training structures are labeled A-F and the corresponding structures rendered along the bottom of the figure.

	PDB overall	PDB best (TBM#T0922)	PDB worst (TBM-hard#T0912)	DES overall	DES best (NTL9)	DES worst (UVF)
	Bonds [%]↑	Bonds [%]↑	Bonds [%]↑	Bonds [%]↑	Bonds [%]↑	Bonds [%]↑
PULCHRA	98.91 *	99.82	98.41	98.45	98.7	99.2
GenZProt	96.253±0.005	97.71±0.22	94.20±0.024	94.855±0.002	96.41±0.003	95.905±0.002
DiAMoNDBack (PDB)	99.18±0.04	99.78±0.18	98.97±0.09	97.981±0.002	98.24±0.01	97.22±0.01
DiAMoNDBack (PDB;DES-FT)	98.99±0.06	99.56±0.09	98.61±0.25	98.725±0.004	98.88±0.01	98.26±0.01
	Clash [%]↓	Clash [%]↓	Clash [%]↓	Clash [%]↓	Clash [%]↓	Clash [%]↓
PULCHRA	0.15 *	0	1.01	0.20	0.26	0.24
GenZProt	8.43±0.22	10.56±1.11	14.81±0.43	6.01±0.04	1.43±0.01	6.96±0.04
DiAMoNDBack (PDB)	0.57±0.09	0.00±0.00	1.07±0.44	0.33±0.01	0.12±0.02	0.58±0.04
DiAMoNDBack (PDB;DES-FT)	0.52±0.16	0.00±0.00	0.91±0.47	0.175±0.007	0.04±0.01	0.38±0.02
	Diversity↓	Diversity↓	Diversity↓	Diversity↓	Diversity↓	Diversity↓
PULCHRA	1.0	1.0	1.0	1.0	1.0	1.0
GenZProt	0.865±0.002	0.84±0.01	0.884±0.001	0.8316±0.0001	0.8758±0.0001	0.8962±0.0001
DiAMoNDBack (PDB)	0.037±0.004	0.16±0.03	-0.004±0.015	0.0329±0.0002	0.0706±0.0004	0.0230±0.0002
DiAMoNDBack (PDB;DES-FT)	0.064±0.004	0.09±0.03	0.076±0.008	0.0223±0.0002	0.0235±0.0002	0.0206±0.0005

Table 8.2: Comparison of backmapping performance on PDB data and MD trajectories between the rules-based PULCHRA approach [530], GenZProt [532], and DiAMoNDBack trained on the PDB dataset, and the same PDB-trained model fine-tuned on MD trajectory data. The 'PDB overall' and 'DES overall' columns report aggregate metrics averaged over all samples and frames for the two data sets with additional columns reporting metrics on the worst- and best-performing systems for each dataset as determined by clash score. Errors for bond and clash scores are reported as standard deviations over five model-generated samples for our model and GenZProt, with no associated error for the deterministic backmapping of the rules-based approach. Errors in the diversity score are calculated using Jackknife resampling over generated samples according to Eqn. 8.2.

Another area of significant improvement compared to GenZProt lies in generative diversity, where both our PDB-trained and DES fine-tuned models achieve good generative diversity metrics near zero for both data sets. In contrast to the deterministic rules-based approach that only produces a single atomistic backmapping, we can achieve competitive accuracy in bond quality and clash score while also recapitulating a great diversity of atomistic structures faithful to a Cα coarse-graining. Examples of the best and worst-performing PDB and DES test set sequences are visualized in Fig. 8.3 where we can see substantial diversity in the side-chain atom placements by our model-generated samples.

. *Aggregate metrics for the rules-based approach on the PDB test set are only evaluated on 16/24 single-chain proteins, as we were unable to get multi-chain proteins properly recognized by the backmapping software.

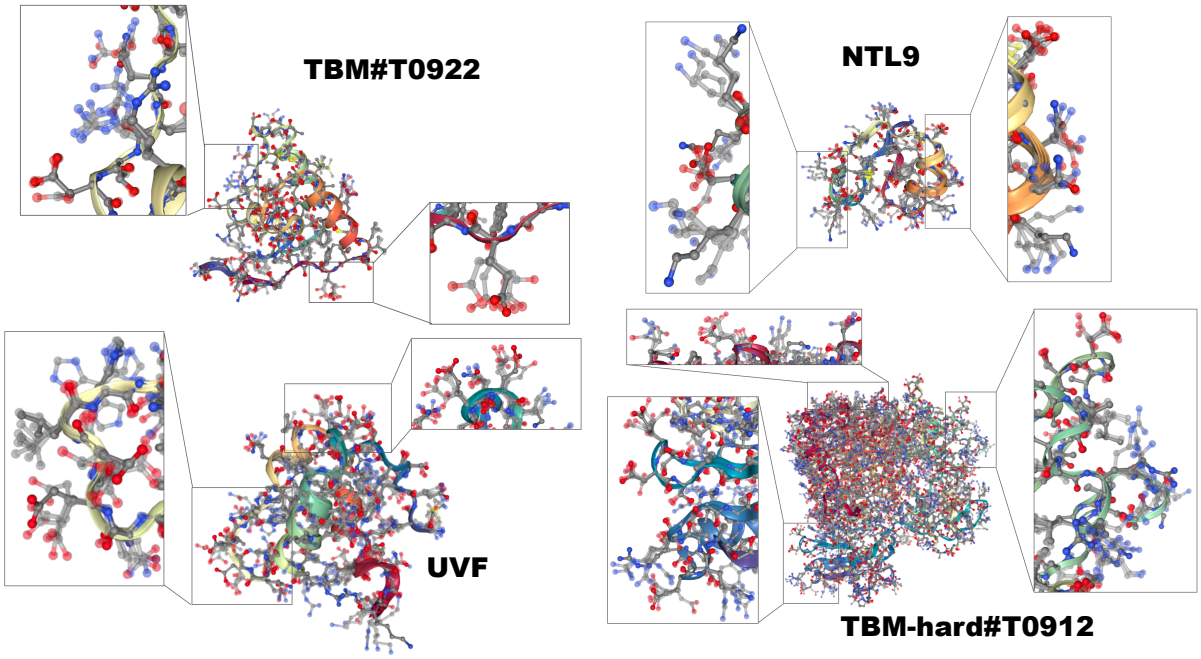


Figure 8.3: Visualization of four protein structures backmapped using our PDB-trained model from the PDB test set (TBM#T0922 and TBM-hard#T0912) and the DES MD test set (NTL9 and UVF). Five samples generated using the Cα backbone as conditioning are shown as translucent structures, with the associated atomistic reference structure shown as a single opaque structure. Insets for each protein zoom into regions highlighting the generative diversity in the side chain placements.

8.4.3 Analysis of side chain dihedral angles in generated atomistic structures

To further evaluate the fidelity of our backmapped structures we examine the distribution of side chain dihedral angles in comparison to the reference atomistic structure. We calculate the C-Cα-Cβ-Cγ dihedral angle for 17/20 amino acids with these available atoms throughout the test set MD simulation data, not applicable for GLY and ALA and there are no CYS residues present in the 11 MD sequences (Fig. 8.4). We measure Jensen-Shannon divergence scores (JS) to quantify the correspondence between our model backmapped dihedral angle distributions and the reference structure distributions for our test-set MD data. Comparing also the dihedral angle distributions produced by the rules-based backmapping we notice overall improved agreement for our PDB-trained model, and further improved JS scores thereafter for the DES fine-tuned model. Exemplar distributions highlighted in

Fig. 8.4 demonstrate less precise dihedral angle generation for the rules-based backmapping in comparison to our model predictions. The rules-based approach tends to produce longer tails in the dihedral distributions consistent with the rules-based backmapping methodology to rotate side chain dihedral angles to avoid steric clashes. This manual adjustment of the structure ultimately results in populating regions of side chain dihedral space that are otherwise unrepresented in the reference structure. Our model appropriately matches the reference dihedral angle distributions that become even more precisely replicated in our backmapped structures after fine-tuning. On average across all residues, we find 26% lower JS scores compared to the rules-based approach for our PDB-trained model which then improves to 49% lower JS scores for the DES fine-tuned model.

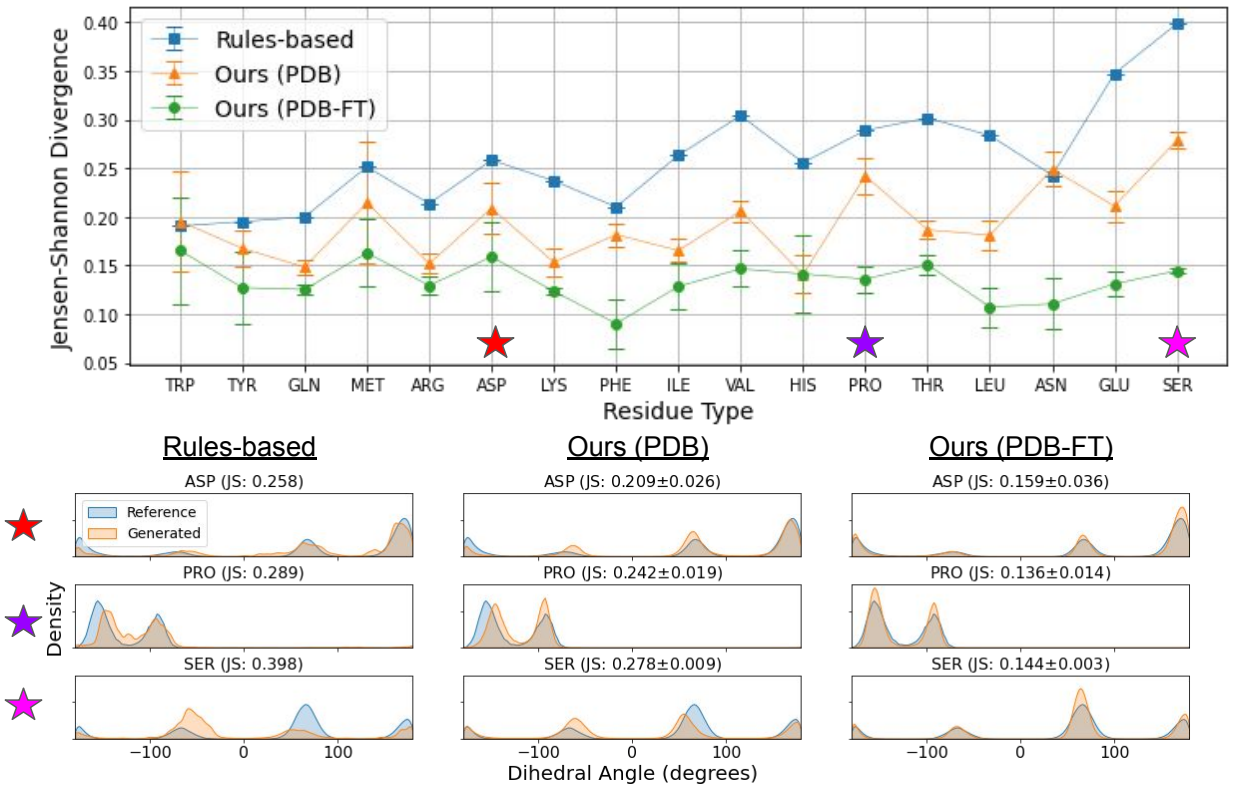


Figure 8.4: Comparison of side chain dihedral angle distributions in the DES test-set data partitioned by residue identity for generated structures produced by the rules-based approach, our PDB-trained model, and the PDB-trained model fine-tuned on DES data. Jensen-Shannon divergence scores for dihedral distributions across all 17 enumerated residues, with colored stars indicating the select distributions visualized below. Dihedral angles are shown for the C-Cα-Cβ-Cγ atoms enumerated for 17/20 available amino acids containing these atoms, not applicable for GLY and ALA and there are no CYS residues present in the 11 MD sequences. Distribution of dihedral angles for three exemplar residues ASP (top), PRO (middle), and SER (bottom) labeled with Jensen-Shannon divergence scores (JS) between the reference and generated distributions. Errors in the JS score are calculated as standard deviations over generated samples for our model predictions, with no associated error prescribed to the deterministic rules-based approach.

8.4.4 Fine-tuning impact on MD trajectories

Fine-tuning our generic PDB-trained model on atomistic data from MD trajectories serves to better calibrate our backmapping to particular sequences and force fields. To visualize the impact of this fine-tuning we compare distributions of bond quality and clash score across all 11 test-set MD trajectory sequences for the PDB-trained and DES fine-tuned models

(Fig. 8.5a). An improvement in the bond quality is evident for all sequences after fine-tuning, indicative of the model becoming better aligned the MD simulation data. While the improvement in clash score is quantitatively significant (Table 8.2), the shift in the per-sequence distributions is less apparent due to the large population of frames with no clashing structures. An average of ∼74% frames across all sequences are generated with no clashes in all five samples for the PDB-trained model, which improves to ∼84% frames with no clashes for the fine-tuned model. Visualizing the performance of the PDB-trained model on the 24 PDB test set structures reveals many PDB sequences with worse clash scores, but improved bond scores, compared to the DES data (Fig. 8.5b). The generally worse clash score for the PDB data is likely a result of the PDB structures being much larger and more globular, while the MD simulation data represents smaller sequences with many frames each that undergo many folding transitions and spend time in more extended configurations that are less susceptible to clashes.

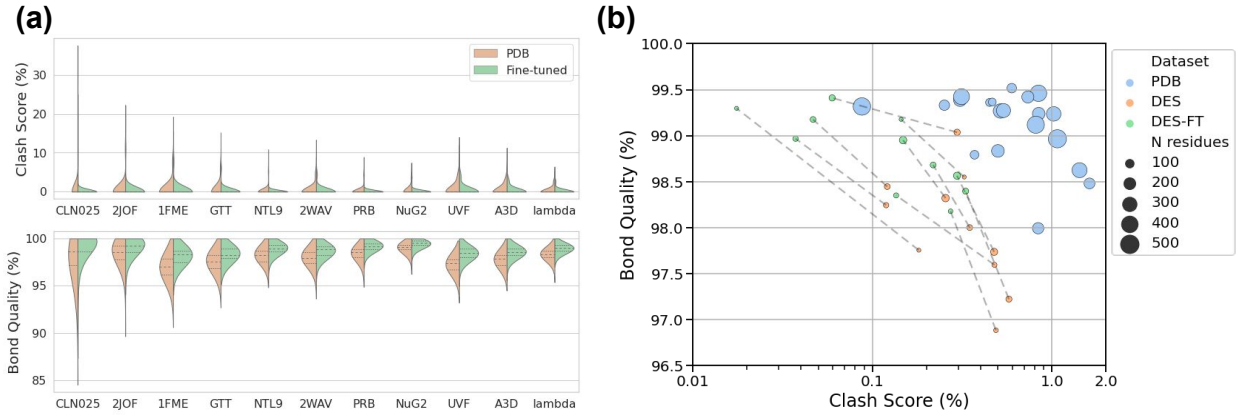


Figure 8.5: Comparison of clash score and bond quality evaluation metrics over the 11 sequences from the MD trajectory data for the PDB-trained and DES fine-tuned models. **(a)** A split violin plot shows the distribution of clash scores (top) and bond qualities (bottom) between generated samples produced by the PDB-trained and fine-tuned models. **(b)** Scatter plot showing the mean bond quality and clash score for the 24 test-set PDB sequences evaluated with the PDB-trained model (blue points) and the 11 MD trajectory sequences for the PDB-trained model (orange) and DES fine-tuned model (green). Grey dashed lines connect the points corresponding to the mean values associated with each sequence between the PDB-trained and DES fine-tuned models.

8.4.5 Analysis of residue-wise performance

Our backmapping approach relies on building the protein structure residue-by-residue, and therefore we analyze trends in performance across residues to probe possible failure modes of our model. Similar to how the clash score is calculated as the percent of clashing residues throughout a structure, we can evaluate per-residue clash scores by determining if the sidechain on the residue of interest lies within a 1.2 Angstrom cutoff of any other residues. This yields a binary value for each residue indicating a clash, which can then be averaged across all residues of a given type to yield a per-residue clash percentage. The per-residue performance of the PDB-trained model on the PDB test set reveals all but three residues display 98%+ bond quality, with only TRP, ARG, and LYS performing below or near this mark (Fig. 8.6a). ARG and LYS, in particular, are both long and charged amino acids that are exceptionally dynamic and rapidly exchange their protons with water, making them challenging to resolve experimentally [551, 552]. This observation is further supported in our data cleaning procedure, where we found that ARG and LYS had significantly more incomplete side chains across the full PDB compared to other residues, indicating that confidence in identifying all side chain positions is comparatively lower than other residues. TRP performs better in terms of bond quality but worse compared to ARG and LYS in clash score likely due to TRP being the largest amino acid and therefore most susceptible to being involved in clashes. TRP is also the least represented residue type in the PDB training data (Fig. 8.6d), which may hinder interpolation to new environments and lead to higher average clash score. Evaluating per-residue performance on the MD simulation data for our PDB-trained model (Fig. 8.6b) and the fine-tuned model (Fig. 8.6c) also reveals the impact that fine-tuning has to improve performance in bond quality and clash score across all residues.

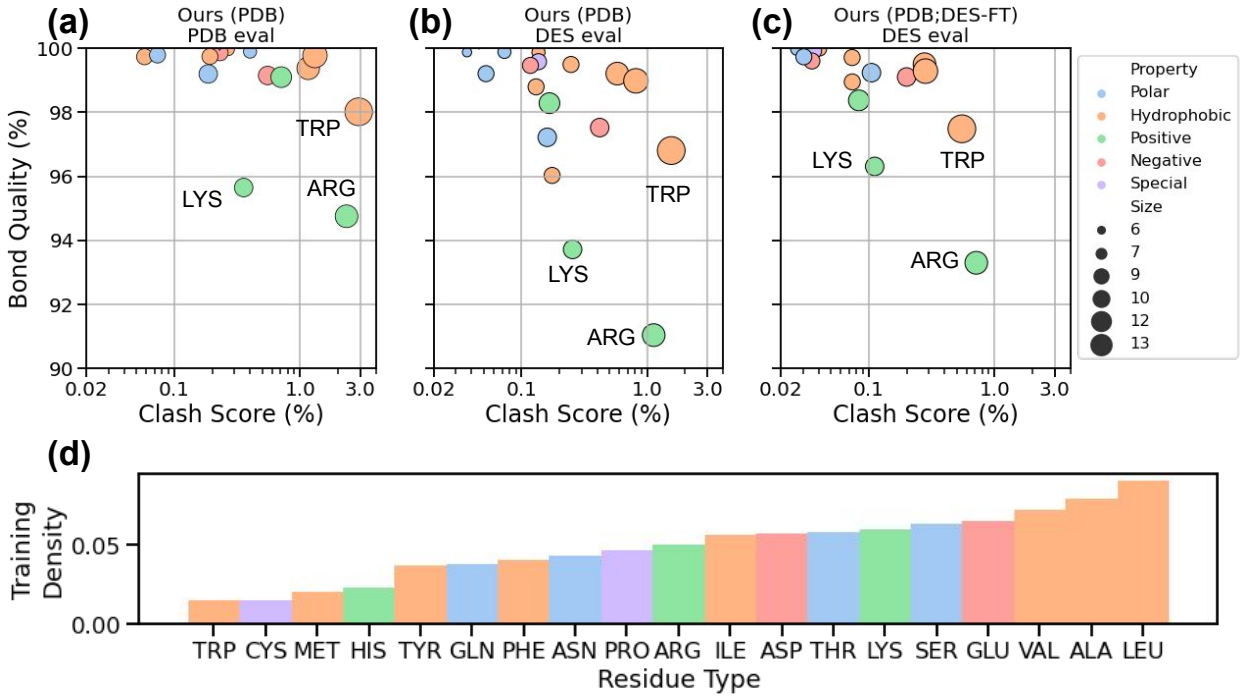


Figure 8.6: Analysis of per-residue bond qualities and clash scores. Scatter plots show the bond quality and clash score for each residue sized by the number of associated side chain atoms and colored by residue-specific properties such as charge, hydrophobicity, polarity, and special (PRO and CYS) with specific more poorly performing residues TRP, LYS, and ARG annotated. Evaluations with the PDB-trained model on the 24 test-set PDB sequences are shown in **(a)**, the same PDB-trained model evaluated on the 11 test-set MD trajectories are shown in **(b)** with the same MD test-set evaluated with the DES fine-tuned model is shown in **(c)**. **(d)** Representation of per-residue training samples in the PDB data training set.

8.4.6 Application to bespoke coarse-grained trajectories

As a final test of the transferability of our backmapping to generic Cα traces, we apply our model to MD trajectories originating from bespoke coarse-grained force fields produced by Majewski et al. [162]. We perform backmapping to evaluate performance in Table 8.3 on three coarse-grained trajectories spanning a range of sequence sizes 1FME (28 residues), PRB (47 residues), and A3D (73 residues). These coarse-grained trajectories are notably out-of-distribution compared to their atomistic MD counterparts. The rules-based approach, for instance, yielding 1.29% clashing residues on the coarse-grained MD 1FME data but only

0.25% clashing residues on the atomistic MD 1FME data. Nevertheless, our backmapping models still yield under 2% clashing residues for all three systems improving on GenZProt that produces an average of over ∼10% clashing residues across the three sequences. Fine-tuning especially serves to improve bond quality, but does not quantitatively degrade and for 1FME and PRB improves upon the clash score. Without any available reference atomistic structure for this data, we cannot use Eqn. 8.2 to calculate generative diversity but instead use the average pairwise RMSD between generated samples as a proxy. Here we notice our model is still capable of producing good generative diversity compared to GenZProt, reporting an average increase in RMSD of ∼630% across these three sequences for the PDB trained model. Overall, we find our model capable of backmapping these bespoke coarse-grained trajectories representing an out-of-domain evaluation with minimal error.

	CG 1FME	CG PRB	CG A3D
	Bonds [%]↑	Bonds [%]↑	Bonds [%]↑
PULCHRA	98.93	99.18	99.60
GenZProt	95.57±0.01	97.210±0.003	96.930±0.001
DiAMoNDBack (PDB)	96.72±0.01	98.23±0.01	98.25±0.01
DiAMoNDBack (PDB;DES-FT)	97.81±0.03	98.97±0.02	98.64±0.01
	Clash [%]↓	Clash [%]↓	Clash [%]↓
PULCHRA	1.29	0.673	0.298
GenZProt	14.01±0.03	8.20±0.05	11.88±0.01
DiAMoNDBack (PDB)	1.96±0.04	1.12±0.05	1.06±0.04
DiAMoNDBack (PDB;DES-FT)	1.75±0.06	0.97±0.03	1.08±0.04
	RMSD [nm]↑	RMSD [nm]↑	RMSD [nm]↑
PULCHRA	–	–	–
GenZProt	0.23±0.09	0.225±0.070	0.18±0.02
DiAMoNDBack (PDB)	1.96±0.29	1.43±0.16	1.67±0.17
DiAMoNDBack (PDB;DES-FT)	1.80±0.27	1.30±0.14	1.57 ±0.15

Table 8.3: Application of backmapping to bespoke coarse-grained trajectories generated by Majewski et al. [162]. Comparisons are presented for three sequences spanning a range of lengths: 1FME (28 residues), PRB (47 residues), and A3D (73 residues). Errors represent standard deviations over generated samples for the bond and clash scores, with generative diversity measured here as the average pairwise RMSD between generated samples.

8.5 Conclusions

In this work, we present a transferable approach for backmapping Cα protein traces into complete atomistic structures leveraging Denoising Diffusion Probabilistic Models (DDPMs) as a generative deep learning model. Our backmappring approach builds the protein structure in an autoregressive manner residue-by-residue, at each instance predicting the Cartesian coordinates of the target residue aligned to a canonical reference frame conditioned on all previously decoded residues and the existing Cα trace. By operating in a fixed reference frame for each residue prediction task we avoid costly data augmentations that would otherwise be required to learn sensitivity to rotational and transitional Cartesian coordinate symmetries. We train models on a corpus of 65k+ PDB structures exposing our model to a rich variety of local residue environments and providing a general-purpose model for backmapping generic Cα traces. Evaluating our backmapping approach on both held-out PDB structures and MD simulation data reveals competitive performance in bond quality ($\sim$98%+ bond lengths within 10% of reference structures) and low steric clashes (fewer than 0.6% clashing residues). Fine-tuning on atomistic MD simulation data improves these metrics for trajectories by better calibrating our model to a particular set of sequences and the MD force field. Our model also produces diverse atomistic reconstructions faithful to a fixed input Cα trace. While we find improved structural metrics compared to the previous transferable backmapping work of Yang and Gómez-Bombarelli [532] (GenZProt) and are competitive with rules-based approaches [530], this aspect of generative diversity is where our model excels producing substantially more diverse atomistic reconstructions than the deterministic rules-based approach or GenZProt. Analysis of side chain dihedral angles in reconstructed structures also reveal our model generates more physically plausible distributions of internal coordinates compared to the rules-based approach that involves manually adjusting side chain dihedral angles to resolve clashes. Finally, we also show our model is applicable to MD trajectories generated by bespoke coarse-grained force fields generating

high-quality bonds (∼98% bond lengths within 10% of reference data) and low quantities of clashing residues (∼1.25%).

Although our backmapping approach demonstrates good performance across a variety of protein structures and data sources, there may still be some possible improvements that can further enhance the quality of our atomistic reconstructions. Our model currently generates atomistic residues from Cα traces conditioned on the N-nearest neighboring residues. Our current data representation exposes the Cartesian coordinates of the local environment and imposes practical limits on the conditioning size (N) due to increasing dimentionality slowing training and inference efficiencies. Residues outside the purview of this local neighborhood can therefore be excluded from the conditioning effectively hidden from the model and potentially leading to clashes. A more comprehensive conditioning scheme that incorporates the full protein structure, for instance by using a graph neural network to assemble the conditioning information, could potentially improve the quality of our generated structures.

CHAPTER 9

CONCLUSIONS AND FUTURE WORK

Artificial intelligence and machine learning techniques have emerged as transformative tools with the potential to revolutionize molecular design and simulation. The integration of machine learning techniques into molecular modeling and design has resulted in new approaches for addressing previously insurmountable challenges. Discovering new molecules and materials with desirable properties is plagued with massive chemical search spaces that render traditional approaches hopelessly inefficient and prohibitively expensive. Machine learning techniques provide a framework to intelligently navigate large search spaces, leveraging data-driven approaches to guide the search and minimize the screening burden. Another area of innovation lies in improving the efficiency of molecular simulation through the development of new algorithms and computational methods. Despite impressive supercomputing hardware and efficient simulation codes, practical simulation times are still limited by the necessity for small integration time steps that restrict the length and time scales that can be reliably resolved. Machine learning techniques provide a toolkit to intervene in molecular simulation workflows, enabling the development of data-driven approaches that accelerate simulations to surpass the time scale barrier. This thesis develops new machine learning and molecular simulation methods and deploys them in service of the discovery, design, and engineering of functional molecules and materials. Frameworks for directed molecular design and high-throughput virtual screening are established and machine learning-enabled approaches are developed for improving the efficiency and broadening the applicability of molecular simulations. These techniques provide a means to overcome the challenges posed by massive search spaces and time scale limitations, enabling more efficient and targeted discovery, design, and simulation of molecules and materials.

9.1 Conclusions

This thesis focuses on four themes of work: (i) studying and optimizing the self-assembly behavior of π-conjugated peptides, (ii) data-driven molecular design using active learning and Bayesian optimization, (iii) machine learning-enabled enhanced sampling techniques, and (iv) atomistic structure generation using generative deep learning for backmapping coarse-grained molecules and trajectories.

In Chapter 2, coarse-grained molecular dynamics simulations are used within an active learning loop to discover π-conjugated peptides demonstrating good self-assembly behavior. After simulating only 2.3% of the 8000-member DXXX-OPV3-XXXD family of π-conjugated peptides a converged rank-ordering of molecules that exhibit promising assembly behaviors into pseudo-1D nanoaggregates with the potential for optoelectronic activity are identified. This ranked list presents an inexpensive filtration of the entire DXXX-OVP3-XXXD sequence space to direct expensive experimental synthesis and characterization efforts toward the most promising candidate molecules. Many design precepts extracted from the high-performing molecules are substantiated by physiochemical intuition, such as the presence of bulky aromatic residues like tryptophan and phenylalanine hindering assembly by interfering with the π-π stacking of the aromatic cores. On the other hand, the absence of any introduced human bias within the active learning protocol also proved to be valuable in identifying high-performing candidates incorporating methionine residues that were not previously considered.

In Chapter 3, we combine all-atom molecular dynamics simulations with UV-vis spectroscopy measurements to perform a hybrid computational-experimental active learning loop. This work engages a $\sim$90$\times$ larger design space than in Chapter 2, containing 694,982 unique members from the X_n-4T-X_n π-conjugated peptide family comprising peptide wings between n =1-5 amino acids in length. This platform enabled the discovery of high-performing new molecules experimentally validated to assemble into pseudo-1D linear nanoaggregates after

testing only 1181 molecules (0.17% of the design space) by computation and 28 (0.004%) by experiment. Interrogating the experimental screening data revealed molecular design rules that link sequence to emergent structure and function of the self-assembled nanoaggregates. The examination of the computational screening simulations unveiled two typical assembly mechanisms and pathways observed in the high-performing molecules: (i) hierarchical assembly of small oligomers with in-register π-core stacking that combine into a single linear aggregate, and (ii) rapid hydrophobic collapse followed by slow structural ripening and the emergence of in-register ordering of the π-cores. This work exposes new understanding of how vaiation in oligopeptide sequence impacts assembly behavior, both corroborating prior physicochemical understanding and revealing new design rules of molecular assembly mechanisms. This integrated computational-experimental active learning platform demonstrates the force of tightly coupled collaboration between theory and experiment and provides a paradigm transferable to other applications in molecular design and discovery.

The study of π-conjugated peptides in this thesis concludes with Chapter 5, which now explores the influence of the π-core chemistry and the length of an alkyl spacer linking the core to the peptide wings on intermolecular charge mobility by combining molecular dynamics simulation and first principles electronic-structure calculations. This work complements the previous studies on the peptide wing composition in Chapters 2 and 3, culminating this exploration of the molecular building blocks that constitute π-conjugated peptides. High-throughput virtual screening is conducted here over a 24-molecule library of π-conjugated peptides containing one of six π-cores each of which has four different designs with between zero and three alkyl spacers connecting the aromatic cores to their flanking peptide wings. The results of the screening find a benzo-dithiophene core with no alkyl spacer as possessing the highest predicted hole mobility of 0.224 $cm^2/(Vs)$, and pyridine-diimide with no alkyl spacer as possessing the highest predicted electron mobility of 0.143 $cm^2/(Vs)$. Charge mobility is found to degrade with increasing spacer length, suggesting the absence of alkyl

spacers as a more promising chemical motif for high charge mobility in π-conjugated peptides. While these calcaulted mobilities are lower than values that may tpically be expected for crystalline semiconductors, the mechanical flexibility, solution processability, weight, aqueous stability, and biocompatibility of π-conjugated peptide-based semiconductors convey advantages well-suited to charge transport and energy harvesting applications in biological and living systems. The computational approach develoepd here is generically applicable for screening various other self-assembling and electronically active organic molecules.

For the final application of active learning presented in this thesis, in Chapter 4 we seek to discover small organic compounds capable of selectively permeating cardiolipin-containing membranes. Molecules that display selectively to cardiolipin over other common membrane phospholipids, such as phosphatidylglycerol, hold promise as biomarkers for staining and quantification of cardiolipin content useful as fluorescent dyes and diagnostic tools. Using transferable coarse-grained models a design space of $\sim 10^{60}$ small molecules under 500 Da can be tractably be represented with only 124,327 unique coarse-grained topologies that define the design space. Alchemical free energy calculations measuring $\Delta\Delta G$ between cardiolipin and phosphatidylglycerol membranes serve as the simulation engine, where after seven rounds of active learning and only testing 0.42% of the molecular design space 184 candidate molecules are discovered with better selectively than the established cardiolipin fluorescent dye 10-N-nonyl acridine orange. Using interpretable LASSO regression to analyze the topological motifs contributing to the cardiolipin selectivity reveals cardiolipin selectivity is associated with positively charged groups, hydrophobicity, and the capacity to form hydrogen bonds. Conversely, the presence of negatively charged groups and overall polarity causes the molecule to partition into the bulk water environment, while a predominant lack of polarity results in the candidate molecules tending to integrate into the central regions of the membrane. Experimental measurements of cardiolipin selectively for two candidate molecules corroborated the trends in the free energy calculations from the computational

models providing validation for the simulation approach. The active learning process and alchemical free energy calculations presented in this study can be readily applied to various molecular design applications. This can be achieved by substituting the thermodynamic selectivity prediction with a computational or experimental assay measuring the target property for optimization.

The work in Chapter 5 develops a new machine learning-enabled enhanced molecular simulation sampling workflow called the Girsanov Reweighting Enhanced Sampling Technique (GREST). Small integration time steps required for numerical stability in molecular simulation limit the practically achievable time scale that can be attained and have sparked the development of numerous computational techniques to accelerate rare event sampling and phase space exploration. Collective variable enhanced sampling techniques are one such method that relies on applying biases along predefined low-dimensional coordinates that are functions of the system degrees of freedom to accelerate barrier crossings and state transitions. While the manual definition of good collective variables can be cumbersome and require domain-specific knowledge, the GREST methodology seeks to provide a data-driven approach for the automatic discovery of collective variables interleaved within an adaptive sampling workflow. Good collective variables that correlate with the slowest evolving dynamical motions in a system are inherently dynamical quantities and therefore require the use of Girsanov formalism that prescribes how to properly reweight biased trajectories taking into account both thermodynamic and integrator-specific dynamical path weights. This capability is the central feature underpinning GREST which interleaves successive rounds of machine learning for slow collective variable discovery and metadynamics enhanced sampling within the slow collective coordinates to accelerate configurational phase space sampling. GREST is capable of furnishing accurate estimates of the slow collective variables without any prior knowledge of the system and achieves significant speedups compared to performing unbiased simulations and enhanced sampling in intuited coordinates.

In Chapter 6, a new temporally-aware approach is developed for backmapping coarse-grained trajectories into all-atom resolution. Coarse-graining is a complementary approach to collective variable enhanced sampling techniques designed to accelerate molecular simulations by simplifying the system retaining fewer degrees of freedom required to preserve relevant physicochemical properties. Reintroducing atomistic detail into coarse-grained representations is a necessary step for analysis that requires all-atom detail, but is a particularly challenging task due to the one-to-many mapping of coarse-grained to atomistic structures. The scheme developed in this work leverages the generative capabilities of conditional variation autoencoders to non-deterministically produce novel atomistic realizations of the same coarse-grained configurations. By providing previously decoded atomistic representations as conditioning, the model can is capable of preserving kinetic properties in the generated trajectories alongside matching thermodynamic and structural characteristics of the reference data. Remarkably, after training on reference atomistic data the model is capable of successfully generalizing the coarse-grained data generated using custom coarse-grained force fields. Temporal coherence sets this approach apart from previous deep learning-enabled backmapping strategies that rely on frame-by-frame reconstruction explicitly neglecting the trajectory information. This backmapping method provides a tool for achieving smooth atomistic reconstructions of coarse-grained trajectories that maintain kinetic, thermodynamic, and structural properties.

The final section of this thesis in Chapter 8 continues the backmapping work, now developing a chemically transferable backmapping model capable of restoring atomistic detail into generic Cα protein traces. This work addresses the major downfall of the previously developed backmapping method in Chapter 6, in that this backmapping strategy no longer requires training bespoke models for each new system. The approach is extensible to arbitrarily large proteins by the inherently local nature of the model that decodes the structure residue-by-residue in an autoregressive manner. Training on a broad and diverse dataset

of 65k+ PDB structures produces a generalizable model applicable to both held-out PDB structures and molecular dynamics trajectories of small fast-folding proteins, as well as trajectories generated using bespoke coarse-grained force fields. This method provides a practical tool for backmapping generic Cα protein traces without the need to re-train models for each new system, with the additional functionality to fine-tune the generic PDB-trained model on available atomistic data for task-specific calibration and improved reconstruction performance.

9.2 Future directions

In light of the findings and conclusions presented throughout this thesis, several promising avenues for future research and exploration emerge that pave the way for exciting new directions in the fields of data-driven molecular design and simulation.

One avenue for future research in the space of discovering new high-performing π-conjugated peptides is continuing to expand the molecular design space exposed to the active learning loop. In Chapters 2 and 3 active learning is executed to search the permutations of amino acids comprising the oligopeptide wings, while in Chapter 5 virtual high-throughput screening is performed over a library of different π-cores. The challenge with performing active learning over a combination of oligopeptide wings and π-core chemistries lies in generating a unified embedding space encompassing these two molecular building blocks. The embeddings generated in this thesis use autoencoder architectures to transform the discrete molecular representations into continuous vector representations that can then be directly exposed to the Bayesian optimization step within the active learning loop. This embedding task is relatively straightforward when the design space consists of only peptide sequences, simplifying the input to the network only needing to account for variations in the peptide sequence. Effectively accounting for both the peptide sequence and π-core chemistry is a more challenging problem due to their inherent differences in structure and properties. Pep-

tide sequences consist of a linear chain of amino acids, which have well-defined composition and a fixed set of amino acid building blocks. On the other hand, π-core chemistries involve aromatic systems and conjugated structures, which introduce a different set of chemical features and complexities. These differences make it difficult to create a unified embedding space that captures the chemical similarity between both the peptide sequence and π-core chemistry accurately.

Restricting the design space to only a fixed set of π-cores makes the problem more tractable, but still presents challenges in how to effectively featurize the complete π-conjugated peptide molecular structure to expose to the neural network. Graph neural networks (GNNs) representing each amino acid as a separate node are used to generate the embedding space in Chapter 3. This approach is less applicable to the heterogeneous structure of different π-cores due to an explosion in the number of nodes. To generate permutationally invariant embeddings requires graph matching to be performed during the reconstruction loss calculation of the autoencoder, this becomes computationally intractable when the graphs grow larger than around seven nodes in size which limits the number of nodes that could be used to represent these molecules. Recent research has proposed new neural network architectures for achieving permutationally invariant autoencoders [553, 554]. Another option is to entirely forgo attempting to produce continuous representations using deep learning-based approachs, and instead, simply leveraging the fact that building the Gaussian Process Regression (GPR) model for the active learning loop can be achieved through the definition of a similarity measure between candidates in the molecular design space. For example, featurizing each molecule using Morgan fingerprint representations of the complete π-conjugated peptide structure can be used with the Tanimoto similarity metric to quantify chemical similarity for the purposes of GPR construction [555]. Future research could explore new neural network architectures for permutationally invariant autoencoders or consider alternative approaches such as leveraging cheminformatic similarity measures for GPR construction. These

avenues hold promise for advancing the discovery of high-performing π-conjugated peptides within virtuous active learning loops.

Potential future work in the active learning project for discovering cardiolipin selective molecules discussed in Chapter 4 could seek to continue experimental testing of promising candidates and expand the virtual high-throughput screening to perform higher resolution all-atom simulations. Currently, a coarse-grained model is used in the screening procedure to reduce the chemical design space and speed up the simulations. However, this approach leads to a significant redundancy of all-atom structures that align with a single coarse-grained representation. To overcome this limitation, atomistic functional groups associated with specific coarse-grained motifs were analyzed in Chapter 4 to establish design rules from the active learning results. Yet, a more comprehensive understanding of the atomistic structures corresponding to the best-performing coarse-grained topologies for further computational screening can be achieved by employing a more rigorous backmapping strategy. One potential solution is to employ a data-driven approach that utilizes the existing coarse-graining function to map atomistic molecules to their corresponding coarse-grained topologies. This can help generate a dataset of exemplar all-atom-coarse-grained pairs, which can then be used to train a stochastic backmapping model, such as an autoencoder or generative adversarial network architecture. While pursuing all-atom calculations of the top-performing candidates identified from the coarse-grained calculations is a possibility, challenges related to the lack of validated all-atom force field parameters and the computational complexity of all-atom free-energy calculations in membrane systems must be addressed [338]. In conclusion, future work should focus on experimental testing, higher-resolution simulations, and the development of a robust backmapping strategy to enhance the effectiveness of discovering cardiolipin selective molecules.

Future work for the GREST enhanced sampling technique developed in Chapter 7 could seek to realize efficiency gains in applications to larger biomolecular systems and also poten-

tially introduce approximations in the workflow to help improve numerical stability. While in Chapter 7 GREST was only demonstrated on the relatively small biomolecular systems of alanine dipeptide and the WLALL pentapeptide, applications to larger proteins or explicit solvent systems would be interesting to pursue. Proper treatment of explicit solvent system would ideally consider both the solute and solvent degrees of freedom. Exposing the solvent to the State-free Reversible VAMPNet (SRV) network model could be achieved using the Permutation Invariant Vector (PIV) formalism proposed by Fabio Pietrucci and coworkers [494, 495] that would enable learning slow modes over the solute and permutationally invariant solvent degrees of freedom. As analytical expressions for the Girsanov pathweights have only been derived for specific integrators, deploying the established Langevin dynamics integrator may not be ideal for the treatment of solvated systems due to their tendency to artificially dampen dynamical processes due to disturbances of the natural time correlations of particle velocities [496]. Another potentially interesting avenue of exploration would be conducting a comparison Girsanov reweighting formalism with the square root approximation (SqRA) to dynamical reweighting [497, 498, 499]. The SqRA can potentially convey advantages compared to Girsanov reweighting in terms of improved numerical robustness and offers an accurate approximation to the exact Girsanov expressions. Donati, Weber, and Keller investigated the potential of utilizing the SqRA instead of Girsanov reweighting within the framework of Markov state models [499]. It would be intriguing to assess the effectiveness and applicability of this approximation technique in the context of GREST. In conclusion, it is evident that GREST can be effectively utilized in various dynamic systems apart from molecular dynamics, aiming to expedite the exploration of phase space in numerical simulations. This can be beneficial in diverse fields such as ecological modeling, climate science, process systems modeling, and financial markets.

Finally, there are numerous future research directions for the data-driven backmapping work developed in Chapters 6 and 8. The chemically transferable backmapping approach in

Chapter 8 addresses the major shortcoming of the approach in Chapter 6, no longer requiring training bespoke models for each new system intended for backmapping. The approach in Chapter 6 does benefit from the flexibility to define arbitrary mappings from coarse-grained to all-atom space, rather than being beholden to only backmapping Cα protein traces. At present, the backmapping model used in Chapter 6 relies on convolutional neural networks (CNNs) that operate on voxelized representations converted to and from Cartesian coordinates. By employing explicitly covariant network architectures [429, 430, 431], such as those utilized in the backmapping scheme presented by Wang et al. [142], there is a potential to achieve superior data efficiencies. This approach eliminates the need to train with random rotations and offers the possibility of significantly reducing the network size and memory requirements compared to voxelizations, thereby enhancing scalability. In this particular method, graph neural networks with E(3) equivariant operations are utilized for generative frame-by-frame backmapping. To incorporate temporal coherence, it may be possible to include preceding trajectory configurations as conditioning variables for the structure generation processes within this framework. Furthermore, while the approach developed in Chapter 8 is not explicitly designed for backmapping trajectories, future research could explore developing a model more geared toward trajectories by also incorporating temporal coherence through additional conditioning. One limitation of the chemically transferable backmapping model in Chapter 8 is the limited context size of the local environment exposed when backmapping each residue. Excluding residues beyond the scope of the nearby area effectively conceals them from the conditioning, which may result in clashes being produced by the model. Incorporating the full structure as conditioning, for example, by processing the structure with a graph neural network, could help provide a more global view and ensure all relevant components are being exposed to the model for each prediction task.

REFERENCES

[1] Sayak Subhra Panda, Howard E Katz, and John D Tovar. Solid-state electrical applications of protein and peptide based nanomaterials. *Chemical Society Reviews*, 47(10):3640–3658, 2018.

[2] Rohan J Kumar, James M MacDonald, Th Birendra Singh, Lynne J Waddington, and Andrew B Holmes. Hierarchical self-assembly of semiconductor functionalized peptide α-helices and optoelectronic properties. *Journal of the American Chemical Society*, 133(22):8564–8573, 2011.

[3] Herdeline Ann M Ardoña and John D Tovar. Peptide π-electron conjugates: organic electronics for biology? *Bioconjugate Chemistry*, 26(12):2290–2302, 2015.

[4] Mohammad Aref Khalily, Gokhan Bakan, Betul Kucukoz, Ahmet Emin Topal, Ahmet Karatay, H Gul Yaglioglu, Aykutlu Dana, and Mustafa O Guler. Fabrication of supramolecular n/p-nanowires via coassembly of oppositely charged peptide-chromophore systems in aqueous media. *ACS Nano*, 11(7):6881–6892, 2017.

[5] David A Stone, Alok S Tayi, Joshua E Goldberger, Liam C Palmer, and Samuel I Stupp. Self-assembly and conductivity of hydrogen-bonded oligothiophene nanofiber networks. *Chemical Communications*, 47(20):5702–5704, 2011.

[6] Hui Shao, James Seifert, Natalie C Romano, Min Gao, Jonathan J Helmus, Christopher P Jaroniec, David A Modarelli, and Jon R Parquette. Amphiphilic self-assembly of an n-type nanotube. *Angewandte Chemie*, 122(42):7854–7857, 2010.

[7] Wei-Wen Tsai, Ian D Tevis, Alok S Tayi, Honggang Cui, and Samuel I Stupp. Semiconducting nanowires from hairpin-shaped self-assembling sexithiophenes. *The Journal of Physical Chemistry B*, 114(45):14778–14786, 2010.

[8] Tetsu Kitamura, Suguru Nakaso, Norihiro Mizoshita, Yusuke Tochigi, Takeshi Shimomura, Masaya Moriyama, Kohzo Ito, and Takashi Kato. Electroactive supramolecular self-assembled fibers comprised of doped tetrathiafulvalene-based gelators. *Journal of the American Chemical Society*, 127(42):14769–14775, 2005.

[9] Reza Saberi Moghaddam, Emily R Draper, Claire Wilson, Hadi Heidari, and Dave J Adams. Effect of electric field on the electrical properties of a self-assembled perylene bisimide. *RSC Advances*, 8(59):34121–34125, 2018.

[10] Galen L Eakins, Rishi Pandey, Jonathan P Wojciechowski, Han Yue Zheng, James EA Webb, Céline Valéry, Pall Thordarson, Natalie OV Plank, Juliet A Gerrard, and Justin M Hodgkiss. Functional organic semiconductors assembled via natural aggregating peptides. *Advanced Functional Materials*, 25(35):5640–5649, 2015.

[11] M Pandeeswar, Harshavardhan Khare, Suryanarayanarao Ramakumar, and T Govindaraju. Crystallographic insight-guided nanoarchitectonics and conductivity modulation of an n-type organic semiconductor through peptide conjugation. *Chemical Communications*, 51(39):8315–8318, 2015.

[12] Javier López-Andarias, María José Rodriguez, Carmen Atienza, Juan Luis López, Tsubasa Mikie, Santiago Casado, Shu Seki, José L Carrascosa, and Nazario Martín. Highly ordered n/p-co-assembled materials with remarkable charge mobilities. *Journal of the American Chemical Society*, 137(2):893–897, 2015.

[13] John D Tovar. Supramolecular construction of optoelectronic biomaterials. *Accounts of Chemical Research*, 46(7):1527–1537, 2013.

[14] Moran Amit, Sivan Yuran, Ehud Gazit, Meital Reches, and Nurit Ashkenasy. Tailor-made functional peptide self-assembling nanostructures. *Advanced Materials*, 30(41):1707083, 2018.

[15] Danielle M Raymond and Bradley L Nilsson. Multicomponent peptide assemblies. *Chemical Society Reviews*, 47(10):3659–3720, 2018.

[16] Rein V Ulijn and Andrew M Smith. Designing peptide based nanomaterials. *Chemical Society Reviews*, 37(4):664–675, 2008.

[17] Amanda B Marciel, Melikhan Tanyeri, Brian D Wall, John D Tovar, Charles M Schroeder, and William L Wilson. Fluidic-directed assembly of aligned oligopeptides with π-conjugated cores. *Advanced Materials*, 25(44):6398–6404, 2013.

[18] Rachael A Mansbach and Andrew L Ferguson. Control of the hierarchical assembly of π-conjugated optoelectronic peptides by ph and flow. *Organic & Biomolecular Chemistry*, 15(26):5484–5502, 2017.

[19] Albertus PHJ Schenning and EW Meijer. Supramolecular electronics; nanowires from self-assembled π-conjugated systems. *Chemical Communications*, (26):3245–3258, 2005.

[20] Miriam Mba, Alessandro Moretto, Lidia Armelao, Marco Crisma, Claudio Toniolo, and Michele Maggini. Synthesis and self-assembly of oligo (p-phenylenevinylene) peptide conjugates in water. *Chemistry–A European Journal*, 17(7):2044–2047, 2011.

[21] Joseph K Gallaher, Emma J Aitken, Robert A Keyzers, and Justin M Hodgkiss. Controlled aggregation of peptide-substituted perylene-bisimides. *Chemical Communications*, 48(64):7961–7963, 2012.

[22] Edward R Jira, Kirill Shmilovich, Tejaswini S Kale, Andrew Ferguson, John D Tovar, and Charles M Schroeder. Effect of core oligomer length on the phase behavior and assembly of π-conjugated peptides. *ACS Applied Materials & Interfaces*, 12(18):20722–20732, 2020.

[23] Kalpana Besar, Herdeline Ann M Ardona, John D Tovar, and Howard E Katz. Demonstration of hole transport and voltage equilibration in self-assembled π-conjugated peptide nanostructures using field-effect transistor architectures. *ACS Nano*, 9(12):12401–12409, 2015.

[24] Brian D Wall, Ashley E Zacca, Allix M Sanders, William L Wilson, Andrew L Ferguson, and John D Tovar. Supramolecular polymorphism: tunable electronic interactions within π-conjugated peptide nanostructures dictated by primary amino acid sequence. *Langmuir*, 30(20):5946–5956, 2014.

[25] Brian D Wall, Stephen R Diegelmann, Shuming Zhang, Thomas J Dawidczyk, William L Wilson, Howard E Katz, Hai-Quan Mao, and John D Tovar. Aligned macroscopic domains of optoelectronic nanostructures prepared via shear-flow assembly of peptide hydrogels. *Advanced Materials*, 23(43):5009–5014, 2011.

[26] Sayak Subhra Panda, Kirill Shmilovich, Nicholas SM Herringer, Nicolas Marin, Andrew L Ferguson, and John D Tovar. Computationally guided tuning of peptide-conjugated perylene diimide self-assembly. *Langmuir*, 37(28):8594–8606, 2021. DOI: 10.1021/acs.langmuir.1c01213.

[27] Sayak Subhra Panda, Kirill Shmilovich, Andrew L Ferguson, and John D Tovar. Controlling supramolecular chirality in peptide- π-peptide networks by variation of the alkyl spacer length. *Langmuir*, 35(43):14060–14073, 2019.

[28] Taein Lee, Sayak Subhra Panda, John D Tovar, and Howard E Katz. Unusually conductive organic–inorganic hybrid nanostructures derived from bio-inspired mineralization of peptide/pi-electron assemblies. *ACS Nano*, 14(2):1846–1855, 2020.

[29] Antonio Facchetti. π-conjugated polymers for organic electronics and photovoltaic cell applications. *Chemistry of Materials*, 23(3):733–758, 2011.

[30] Linyi Bian, Enwei Zhu, Jian Tang, Weihua Tang, and Fujun Zhang. Recent progress in the design of narrow bandgap conjugated polymers for high-efficiency organic solar cells. *Progress in Polymer Science*, 37(9):1292–1331, 2012.

[31] Pierre M Beaujuge and John R Reynolds. Color control in π-conjugated organic polymers for use in electrochromic devices. *Chemical Reviews*, 110(1):268–320, 2010.

[32] Herdeline Ann M Ardoña and John D Tovar. Energy transfer within responsive pi-conjugated coassembled peptide-based nanostructures in aqueous environments. *Chemical Science*, 6(2):1474–1484, 2015.

[33] Bryce A Thurston, John D Tovar, and Andrew L Ferguson. Thermodynamics, morphology, and kinetics of early-stage self-assembly of π-conjugated oligopeptides. *Molecular Simulation*, 42(12):955–975, 2016.

[34] Lawrence R Valverde, Bryce A Thurston, Andrew L Ferguson, and William L Wilson. Evidence for prenucleated fibrilogenesis of acid-mediated self-assembling oligopeptides via molecular simulation and fluorescence correlation spectroscopy. *Langmuir*, 34(25):7346–7354, 2018.

[35] Lawrence R. Valverde, Bryce A. Thurston, Andrew L. Ferguson, and William L. Wilson. Evidence for prenucleated fibrilogenesis of acid-mediated self-assembling oligopeptides via molecular simulation and fluorescence correlation spectroscopy. *Langmuir*, 34(25):7346–7354, 6 2018.

[36] Rachael A. Mansbach and Andrew L. Ferguson. Control of the hierarchical assembly of π-conjugated optoelectronic peptides by pH and flow. *Organic and Biomolecular Chemistry*, 15:5484–5502, 2017.

[37] Rachael A. Mansbach and Andrew L. Ferguson. Coarse-grained molecular simulation of the hierarchical self-assembly of π-conjugated optoelectronic peptides. *Journal of Physical Chemistry B*, 121(7):1684–1706, 2017.

[38] Rachael A. Mansbach and Andrew L. Ferguson. Patchy particle model of the hierarchical self-assembly of π-conjugated optoelectronic peptides. *The Journal of Physical Chemistry B*, 122(44):10219–10236, 11 2018.

[39] Chiho Kim, Anand Chandrasekaran, Anurag Jha, and Rampi Ramprasad. Active-learning and materials design: The example of high glass transition temperature polymers. *MRS Communications*, 9(3):860–866, 2019.

[40] Julia Ling, Maxwell Hutchinson, Erin Antono, Sean Paradiso, and Bryce Meredig. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integrating Materials and Manufacturing Innovation*, 6(3):207–217, 2017.

[41] Rainier Barrett and Andrew D White. Iterative peptide modeling with active learning and meta-learning. *arXiv preprint arXiv:1911.09103*, 2019.

[42] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2 2018.

[43] G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2 2012.

[44] Gilles Horowitz. Organic field-effect transistors. *Advanced Materials*, 10(5):365–377, 1998.

[45] Amaresh Mishra, Chang-Qi Ma, and Peter Bauerle. Functional oligothiophenes: molecular design for multidimensional nanoarchitectures and their applications. *Chemical Reviews*, 109(3):1141–1276, 2009.

[46] Michael Kasha, HR Rawls, and M Ashraf El-Bayoumi. The exciton model in molecular spectroscopy. *Pure and Applied Chemistry*, 11(3-4):371–392, 1965.

[47] Herdeline Ann M Ardona, Kalpana Besar, Matteo Togninalli, Howard E Katz, and John D Tovar. Sequence-dependent mechanical, photophysical and electrical properties of pi-conjugated peptide hydrogelators. *Journal of Materials Chemistry C*, 3(25):6505–6514, 2015.

[48] Anatoliy N. Sokolov, Sule Atahan-Evrenk, Rajib Mondal, Hylke B. Akkerman, Roel S. Sánchez-Carrera, Sergio Granados-Focil, Joshua Schrier, Stefan C.B. Mannsfeld, Arjan P. Zoombelt, Zhenan Bao, and Alán Aspuru-Guzik. From computational discovery to experimental characterization of a high hole mobility organic crystal. *Nature Communications*, 2(1):437, September 2011.

[49] Allix M Sanders, Thomas J Dawidczyk, Howard E Katz, and John D Tovar. Peptide-based supramolecular semiconductor nanomaterials via pd-catalyzed solid-phase "dimerizations". *ACS Macro Letters*, 1(11):1326–1329, 2012.

[50] Anthony J Petty, Qianxiang Ai, Jeni C Sorli, Hamna F Haneef, Geoffrey E Purdum, Alex Boehm, Devin B Granger, Kaichen Gu, Carla Patricia Lacerda Rubinger, Sean R Parkin, Kenneth R Graham, Oana D Jurchescu, Yueh-Lin Loo, Chad Risko, and John E Anthony. Computationally aided design of a high-performance organic semiconductor: the development of a universal crystal engineering core. *Chemical Science*, 10(45):10543–10549, 2019.

[51] Zhaosheng Hu, Zhenhua Lin, Jie Su, Jincheng Zhang, Yue Hao, Jingjing Chang, and Jishan Wu. Controllable self-assembly of ptcdi-c8 for high mobility low-dimensional organic field-effect transistors. *ACS Applied Electronic Materials*, 1(10):2030–2036, 2019.

[52] Byung Jun Jung, Kyusang Lee, Jia Sun, Andreas G Andreou, and Howard E Katz. Air-operable, high-mobility organic transistors with semifluorinated side chains and unsubstituted naphthalenetetracarboxylic diimide cores: High mobility and environmental and bias stress stability from the perfluorooctylpropyl side chain. *Advanced Functional Materials*, 20(17):2930–2944, 2010.

[53] Masahiro Funahashi and Jun-Ichi Hanna. High carrier mobility up to 0.1 cm2/vs and a wide mesomorphic temperature range of alkynyl-substituted terthiophene and quaterthiophene derivatives. *Molecular Crystals and Liquid Crystals*, 436(1):225–1179, 2005.

[54] Antonio Facchetti, Myung-Han Yoon, Charlotte L Stern, Howard E Katz, and Tobin J Marks. Building blocks for n-type organic electronics: Regiochemically modulated

inversion of majority carrier sign in perfluoroarene-modified polythiophene semiconductors. *Angewandte Chemie*, 115(33):4030–4033, 2003.

[55] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[56] Regression. In *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[57] Burr Settles. Active learning literature survey. 2009.

[58] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[59] Dezhen Xue, Prasanna V Balachandran, John Hogden, James Theiler, Deqing Xue, and Turab Lookman. Accelerated search for materials with targeted properties by adaptive design. *Nature Communications*, 7(1):1–9, 2016.

[60] Ruihao Yuan, Zhen Liu, Prasanna V Balachandran, Deqing Xue, Yumei Zhou, Xiangdong Ding, Jun Sun, Dezhen Xue, and Turab Lookman. Accelerated discovery of large electrostrains in batio3-based piezoelectrics using active learning. *Advanced Materials*, 30(7):1702884, 2018.

[61] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA, 1967.

[62] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[63] Devinderjit Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, 2006.

[64] Mark Ebden. Gaussian processes: A quick introduction. *arXiv preprint arXiv:1505.02965*, 2015.

[65] Jonas Močkus. On bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pages 400–404. Springer, 1975.

[66] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[67] Daniel James Lizotte. *Practical Bayesian optimization*. PhD thesis, 2008.

[68] Romy Lorenz, Ricardo Pio Monti, Inês R Violante, Christoforos Anagnostopoulos, Aldo A Faisal, Giovanni Montana, and Robert Leech. The automatic neuroscientist: A framework for optimizing experimental design with closed-loop real-time fmri. *NeuroImage*, 129:320–334, 2016.

[69] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, volume 2, page 6. Citeseer, 2000.

[70] Andreas Vlachos. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312, 7 2008.

[71] Jingbo Zhu, Huizhen Wang, and Eduard Hovy. Multi-criteria-based strategy to stop active learning for data annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1129–1136. Association for Computational Linguistics, 2008.

[72] Michael Bloodgood and K Vijay-Shanker. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. *arXiv preprint arXiv:1409.5165*, 2014.

[73] Michael Bloodgood and John Grothendieck. Analysis of stopping active learning based on stabilizing predictions. *arXiv preprint arXiv:1504.06329*, 2015.

[74] Garrett Beatty, Ethan Kochis, and Michael Bloodgood. The use of unlabeled data versus labeled data for stopping active learning for text classification. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 287–294. IEEE, 2019.

[75] Jan Dudek. Role of cardiolipin in mitochondrial signaling pathways. *Frontiers in Cell and Developmental Biology*, 5:90, 2017.

[76] Giuseppe Paradies, Valeria Paradies, Francesca M Ruggiero, and Giuseppe Petrosillo. Role of cardiolipin in mitochondrial function and dynamics in health and disease: molecular and pharmacological aspects. *Cells*, 8(7):728, 2019.

[77] François Gonzalvez, Marilena D'Aurelio, Marie Boutant, Aoula Moustapha, Jean-Philippe Puech, Thomas Landes, Laeticia Arnauné-Pelloquin, Guillaume Vial, Nellie Taleux, Christian Slomianny, et al. Barth syndrome: cellular compensation of mitochondrial dysfunction and apoptosis inhibition due to changes in cardiolipin remodeling linked to tafazzin (taz) gene mutation. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1832(8):1194–1206, 2013.

[78] Hai-Feng Tian, Jin-Mei Feng, and Jian-Fan Wen. The evolution of cardiolipin biosynthesis and maturation pathways and its implications for the evolution of eukaryotes. *BMC Evolutionary Biology*, 12(1):1–15, 2012.

[79] Edgard M Mejia, Hieu Nguyen, and Grant M Hatch. Mammalian cardiolipin biosynthesis. *Chemistry and Physics of Lipids*, 179:11–16, 2014.

[80] Michelle Grace Acoba, Nanami Senoo, and Steven M Claypool. Phospholipid ebb and flow makes mitochondria go. *Journal of Cell Biology*, 219(8):e202003131, 2020.

[81] Christophe Chipot and Andrew Pohorille. *Free Energy Calculations*. Springer, 2007.

[82] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008.

[83] Beauchamp, Kyle A and Chodera, John D and Naden, Levi N and Shirts, Michael R. Python implementation of the multistate bennett acceptance ratio (mbar). Published under the MIT license.

[84] Christopher M Dobson et al. Chemical space and biology. *Nature*, 432(7019):824–828, 2004.

[85] Kiran H Kanekal and Tristan Bereau. Resolution limit of data-driven coarse-grained models spanning chemical space. *The Journal of Chemical Physics*, 151(16):164106, 2019.

[86] Myriam E Rodriguez, Kashif Azizuddin, Ping Zhang, Song-mao Chiu, Minh Lam, Malcolm E Kenney, Clemens Burda, and Nancy L Oleinick. Targeting of mitochondria by 10-n-alkyl acridine orange analogues: role of alkyl chain length in determining cellular uptake and localization. *Mitochondrion*, 8(3):237–246, 2008.

[87] Chris WT Leung, Yuning Hong, Jonas Hanske, Engui Zhao, Sijie Chen, Ekaterina V Pletneva, and Ben Zhong Tang. Superior fluorescent probe for detection of cardiolipin. *Analytical Chemistry*, 86(2):1263–1268, 2014.

[88] Jonathan D Hirst, David R Glowacki, and Marc Baaden. Molecular simulations and visualization: Introduction and overview. *Faraday Discuss.*, 169:9–22, 2014.

[89] David E Shaw, JP Grossman, Joseph A Bank, Brannon Batson, J Adam Butts, Jack C Chao, Martin M Deneroff, Ron O Dror, Amos Even, Christopher H Fenton, et al. Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 41–53. IEEE, 2014.

[90] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kale, and Klaus Schulten. Scalable molecular dynamics with namd. *J. Comput. Chem.*, 26(16):1781–1802, 2005.

[91] Edmond Chow, Charles A Rendleman, Kevin J Bowers, Ron O Dror, Douglas H Hughes, Justin Gullingsrud, Federico D Sacerdoti, and David E Shaw. Desmond performance on a cluster of multicore processors. *DE Shaw Research Technical Report DESRES/TR–2008-01*, 2008.

[92] Jens Glaser, Trung Dac Nguyen, Joshua A Anderson, Pak Lui, Filippo Spiga, Jaime A Millan, David C Morse, and Sharon C Glotzer. Strong scaling of general-purpose molecular dynamics simulations on gpus. *Comput. Phys. Commun.*, 192:97–107, 2015.

[93] Nikola Tchipev, Steffen Seckler, Matthias Heinen, Jadran Vrabec, Fabio Gratl, Martin Horsch, Martin Bernreuther, Colin W Glass, Christoph Niethammer, Nicolay Hammer, Bernd Krischok, Michael Resch, Dieter Kranzlmüller, Hans Hasse, Hans-Joachim Bungartz, and Philipp Neumann. Twetris: Twenty trillion-atom simulation. *Int J High Perform Comput Appl.*, 33(5):838–854, 2019.

[94] Ron Elber. Perspective: Computer simulations of long time dynamics. *J. Chem. Phys.*, 144(6):060901, 2016.

[95] Matt J Harvey, Giovanni Giupponi, and G De Fabritiis. Acemd: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.*, 5(6):1632–1639, 2009.

[96] David E Shaw, Peter J Adams, Asaph Azaria, Joseph A Bank, Brannon Batson, Alistair Bell, Michael Bergdorf, Jhanvi Bhatt, J Adam Butts, Timothy Correia, et al. Anton 3: twenty microseconds of molecular dynamics simulation before lunch. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11, 2021.

[97] Peter L Freddolino, Christopher B Harrison, Yanxin Liu, and Klaus Schulten. Challenges in protein-folding simulations. *Nat. Phys.*, 6(10):751–758, 2010.

[98] Nuria Plattner, Stefan Doerr, Gianni De Fabritiis, and Frank Noé. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling. *Nat. Chem*, 9(10):1005–1011, 2017.

[99] Julie Roy and Charles A Laughton. Long-timescale molecular-dynamics simulations of the major urinary protein provide atomistic interpretations of the unusual thermodynamics of ligand binding. *Biophys. J.*, 99(1):218–226, 2010.

[100] Pim WJM Frederix, Ilias Patmanidis, and Siewert J Marrink. Molecular simulations of self-assembling bio-inspired supramolecular systems and their connection to experiments. *Chem Soc Rev*, 47(10):3470–3489, 2018.

[101] Hythem Sidky, Wei Chen, and Andrew L Ferguson. Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Mol. Phys.*, 118(5):e1737742, 2020.

[102] Angel E García. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, 68(17):2696, 1992.

[103] Andrea Amadei, Antonius BM Linssen, and Herman JC Berendsen. Essential dynamics of proteins. *Proteins: Struct. Funct. Genet.*, 17(4):412–425, 1993.

[104] Rainer Hegger, Alexandros Altis, Phuong H Nguyen, and Gerhard Stock. How complex is the dynamics of peptide folding? *Phys. Rev. Lett.*, 98(2):028102, 2007.

[105] Pavel I Zhuravlev, Christopher Kroboth Materese, and Garegin A Papoian. Deconstructing the native state: energy landscapes, function, and dynamics of globular proteins. *J. Phys. Chem. B*, 113(26):8800–8812, 2009.

[106] Payel Das, Mark Moll, Hernan Stamati, Lydia E Kavraki, and Cecilia Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. U.S.A.*, 103(26):9885–9890, 2006.

[107] Mary A Rohrdanz, Wenwei Zheng, and Cecilia Clementi. Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annu. Rev. Phys. Chem.*, 64:295–316, 2013.

[108] Igor Vladimirovich Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory Probab. its Appl.*, 5(3):285–301, 1960.

[109] Lorenzo Donati, Carsten Hartmann, and Bettina G Keller. Girsanov reweighting for path ensembles and markov state models. *J. Chem. Phys.*, 146(24):244112, 2017.

[110] Luca Donati and Bettina G Keller. Girsanov reweighting for metadynamics simulations. *J. Chem. Phys.*, 149(7):072335, 2018.

[111] Stefanie Kieninger and Bettina G Keller. Path probability ratios for langevin dynamics—exact and approximate. *J. Chem. Phys.*, 154(9):094102, 2021.

[112] Jeffrey K Weber and Vijay S Pande. Potential-based dynamical reweighting for markov state models of protein dynamics. *J. Chem. Theory Comput.*, 11(6):2412–2420, 2015.

[113] Kurt Kremer and Florian Müller-Plathe. Multiscale simulation in polymer science. *Mol. Simul.*, 28(8-9):729–750, 2002.

[114] Mark F Horstemeyer. Multiscale modeling: a review. In *Practical Aspects of Computational Chemistry*, pages 87–135. Springer, 2009.

[115] Christine Peter and Kurt Kremer. Multiscale simulation of soft matter systems–from the atomistic to the coarse-grained level and back. *Soft Matter*, 5(22):4357–4366, 2009.

[116] Helgi I Ingólfsson, Cesar A Lopez, Jaakko J Uusitalo, Djurre H de Jong, Srinivasa M Gopal, Xavier Periole, and Siewert J Marrink. The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 4(3):225–248, 2014.

[117] Matej Praprotnik, Silvina Matysiak, Luigi Delle Site, Kurt Kremer, and Cecilia Clementi. Adaptive resolution simulation of liquid water. *J. Phys.: Condens. Matter*, 21(49):499801–499801, Dec 2008.

[118] Matej Praprotnik, Luigi Delle Site, and Kurt Kremer. Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly. *J. Chem. Phys.*, 123(22):224106, Dec 2005.

[119] Gregory A Voth. *Coarse-Graining of Condensed Phase and Biomolecular Systems*. CRC press, 2008.

[120] William George Noid. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.*, 139(9):09B201_1, 2013.

[121] Emiliano Brini, Elena A Algaer, Pritam Ganguly, Chunli Li, Francisco Rodríguez-Ropero, and Nico FA van der Vegt. Systematic coarse-graining methods for soft matter simulations–a review. *Soft Matter*, 9(7):2108–2119, 2013.

[122] Praveen K Depa and Janna K Maranas. Speed up of dynamic observables in coarse-grained molecular-dynamics simulations of unentangled polymers. *J. Chem. Phys.*, 123(9):094901, 2005.

[123] Praveen Depa, Chunxia Chen, and Janna K Maranas. Why are coarse-grained force fields too fast? a look at dynamics of four coarse-grained polymers. *J. Chem. Phys.*, 134(1):014903, 2011.

[124] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H De Vries. The martini force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.

[125] Siewert J Marrink and D Peter Tieleman. Perspective on the martini model. *Chem. Soc. Rev.*, 42(16):6801–6822, 2013.

[126] Steve O Nielsen, Carlos F Lopez, Goundla Srinivas, and Michael L Klein. Coarse grain models and the computer simulation of soft materials. *J. Phys.: Condens. Matter*, 16(15):R481, 2004.

[127] Marc Baaden and Siewert J Marrink. Coarse-grain modelling of protein–protein interactions. *Curr. Opin. Struct. Biol.*, 23(6):878–886, 2013.

[128] Chun Wu and Joan-Emma Shea. Coarse-grained models for protein aggregation. *Curr. Opin. Struct. Biol.*, 21(2):209–220, 2011.

[129] Guojie Zhang, Anthony Chazirakis, Vagelis A Harmandaris, Torsten Stuehn, Kostas Ch Daoulas, and Kurt Kremer. Hierarchical modelling of polystyrene melts: From soft blobs to atomistic resolution. *Soft Matter*, 15(2):289–302, 2019.

[130] Goundla Srinivas, Dennis E Discher, and Michael L Klein. Self-assembly and properties of diblock copolymers by coarse-grain molecular dynamics. *Nat. Mater.*, 3(9):638–644, 2004.

[131] Kirill Shmilovich, Rachael A Mansbach, Hythem Sidky, Olivia E Dunne, Sayak Subhra Panda, John D Tovar, and Andrew L Ferguson. Discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation. *The Journal of Physical Chemistry B*, 124(19):3873–3891, 2020.

[132] Bernadette Mohr, Kirill Shmilovich, Isabel S Kleinwächter, Dirk Schneider, Andrew L Ferguson, and Tristan Bereau. Data-driven discovery of cardiolipin-selective small molecules by computational active learning. *Chemical Science*, 13(16):4498–4511, 2022.

[133] Mohsen Sadeghi and Frank Noé. Large-scale simulation of biomembranes incorporating realistic kinetics into coarse-grained models. *Nat. Commun.*, 11(1):1–13, 2020.

[134] Ryan Bradley and Ravi Radhakrishnan. Coarse-grained models for protein-cell membrane interactions. *Polymers*, 5(3):890–936, 2013.

[135] Jiang Wang, Mohit Gayatri, and Andrew L Ferguson. Coarse-grained molecular simulation and nonlinear manifold learning of archipelago asphaltene aggregation and folding. *J. Phys. Chem. B*, 122(25):6627–6647, 2018.

[136] Berk Hess, Salvador León, Nico Van Der Vegt, and Kurt Kremer. Long time atomistic polymer trajectories from coarse grained simulations: bisphenol-a polycarbonate. *Soft Matter*, 2(5):409–414, 2006.

[137] Andrzej J Rzepiela, Lars V Schäfer, Nicolae Goga, H Jelger Risselada, Alex H De Vries, and Siewert J Marrink. Reconstruction of atomistic details from coarse-grained structures. *J. Comput. Chem.*, 31(6):1333–1343, 2010.

[138] Tsjerk A Wassenaar, Kristyna Pluhackova, Rainer A Böckmann, Siewert J Marrink, and D Peter Tieleman. Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *Journal of Chemical Theory and Computation*, 10(2):676–690, 2014.

[139] Allison P. Heath, Lydia E. Kavraki, and Cecilia Clementi. From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. *Proteins: Struct., Funct., Bioinf.*, 68(3):646–661, 2007.

[140] Marc Stieffenhofer, Michael Wand, and Tristan Bereau. Adversarial reverse mapping of equilibrated condensed-phase molecular structures. *Mach. Learn.: Sci. Technol.*, 1(4):045014, 2020.

[141] Marc Stieffenhofer, Tristan Bereau, and Michael Wand. Adversarial reverse mapping of condensed-phase molecular structures: Chemical transferability. *APL Mater.*, 9(3):031107, 2021.

[142] Wujie Wang, Minkai Xu, Chen Cai, Benjamin Kurt Miller, Tess Smidt, Yusu Wang, Jian Tang, and Rafael Gómez-Bombarelli. Generative coarse-graining of molecular conformations. *arXiv preprint arXiv:2201.12176*, 2022. Submission date: 28 Jan 2022, DOI:10.48550/arXiv.2201.12176.

[143] Wei Li, Craig Burkhart, Patrycja Polińska, Vagelis Harmandaris, and Manolis Doxastakis. Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach. *J. Chem. Phys.*, 153(4):041101, 2020.

[144] Wujie Wang and Rafael Gómez-Bombarelli. Coarse-graining auto-encoders for molecular dynamics. *npj Comput. Mater.*, 5(1):1–9, 2019.

[145] Yaxin An and Sanket A Deshmukh. Machine learning approach for accurate backmapping of coarse-grained models to all-atom models. *Chem. Comm.*, 56(65):9312–9315, 2020.

[146] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[147] Kunihiko Fukushima, Sei Miyake, and Takayuki Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Trans. Syst. Man Cybern. A Syst.*, (5):826–834, 1983.

[148] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, 1989.

[149] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[150] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Comput.*, 29(9):2352–2449, 2017.

[151] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.*, 2018, 2018.

[152] Brooke E Husic, Nicholas E Charron, Dominik Lemm, Jiang Wang, Adrià Pérez, Maciej Majewski, Andreas Krämer, Yaoyi Chen, Simon Olsson, Gianni de Fabritiis, et al. Coarse graining molecular dynamics with graph neural networks. *J. Chem. Phys.*, 153(19):194101, 2020.

[153] Kirill Shmilovich, Marc Stieffenhofer, Nicholas E Charron, and Moritz Hoffmann. Temporally coherent backmapping of molecular trajectories from coarse-grained to atomistic resolution. *The Journal of Physical Chemistry A*, 126(48):9124–9139, 2022.

[154] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[155] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[156] Kirill Shmilovich, Devin Willmott, Ivan Batalov, Mordechai Kornbluth, Jonathan Mailoa, and J Zico Kolter. Orbital mixer: Using atomic orbital features for basis-dependent prediction of molecular wavefunctions. *Journal of Chemical Theory and Computation*, 18(10):6021–6030, 2022.

[157] Mohammed AlQuraishi. Proteinnet: a standardized data set for machine learning of protein structure. *BMC bioinformatics*, 20(1):1–10, 2019.

[158] Jonathan Edward King and David Ryan Koes. Sidechainnet: An all-atom protein structure dataset for machine learning. *Proteins: Structure, Function, and Bioinformatics*, 89(11):1489–1496, 2021.

[159] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[160] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12):980–980, 2003.

[161] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

[162] Maciej Majewski, Adrià Pérez, Philipp Thölke, Stefan Doerr, Nicholas E Charron, Toni Giorgino, Brooke E Husic, Cecilia Clementi, Frank Noé, and Gianni De Fabritiis. Machine learning coarse-grained potentials of protein thermodynamics. *arXiv preprint arXiv:2212.07492*, 2022.

[163] Kirill Shmilovich, Sayak Subhra Panda, Anna Stouffer, John D Tovar, and Andrew L Ferguson. Hybrid computational–experimental data-driven design of self-assembling π-conjugated peptides. *Digital Discovery*, 1(4):448–462, 2022.

[164] Kirill Shmilovich, Yifan Yao, John D Tovar, Howard E Katz, André Schleife, and Andrew L Ferguson. Computational discovery of high charge mobility self-assembling π-conjugated peptides. *Molecular Systems Design & Engineering*, 7(5):447–459, 2022.

[165] Kirill Shmilovich and Andrew L Ferguson. Girsanov reweighting enhanced sampling technique (grest): On-the-fly data-driven discovery of and enhanced sampling in slow collective variables. *The Journal of Physical Chemistry A*, 127(15):3497–3517, 2023.

[166] Dorothea Pinotsi, Luca Grisanti, Pierre Mahou, Ralph Gebauer, Clemens F. Kaminski, Ali Hassanali, and Gabriele S. Kaminski Schierle. Proton transfer and structure-specific fluorescence in hydrogen bond-rich protein structures. *Journal of the American Chemical Society*, 138(9):3046–3057, 3 2016.

[167] Xin Guo, Martin Baumgarten, and Klaus Müllen. Designing π-conjugated polymers for organic electronics. *Progress in Polymer Science*, 38(12):1832–1908, 12 2013.

[168] Se Hye Kim and Jon R Parquette. A model for the controlled assembly of semiconductor peptides. *Nanoscale*, 4(22):6940–6947, 2012.

[169] Ullrich Mitschke and Peter Bäuerle. The electroluminescence of organic materials. *Journal of Materials Chemistry*, 10(7):1471–1507, 2000.

[170] Jean Roncali. Conjugated poly (thiophenes): synthesis, functionalization, and applications. *Chemical Reviews*, 92(4):711–738, 1992.

[171] Denis Fichou and Christiane Ziegler. *Structure and Properties of Oligothiophenes in the Solid State: Single crystals and thin films*. Wiley-VCH: Weinheim, 1999.

[172] Linyi Bian, Enwei Zhu, Jian Tang, Weihua Tang, and Fujun Zhang. Recent progress in the design of narrow bandgap conjugated polymers for high-efficiency organic solar cells. *Progress in Polymer Science*, 37(9):1292–1331, 2012.

[173] Xin Guo, Martin Baumgarten, and Klaus Müllen. Designing π-conjugated polymers for organic electronics. *Progress in Polymer Science*, 38(12):1832–1908, 2013.

[174] Christopher R Newman, C Daniel Frisbie, Demetrio A da Silva Filho, Jean-Luc Brédas, Paul C Ewbank, and Kent R Mann. Introduction to organic thin film transistors and design of n-channel organic semiconductors. *Chemistry of Materials*, 16(23):4436–4451, 2004.

[175] Seth R Marder and Kwang-Sup Lee. *Photoresponsive Polymers II*, volume 214. Springer, 2008.

[176] Roman Marty, Ruth Szilluweit, Antoni Sánchez-Ferrer, Sreenath Bolisetty, Jozef Adamcik, Raffaele Mezzenga, Eike-Christian Spitzner, Martin Feifer, Stephan N Steinmann, Clémence Corminboeuf, et al. Hierarchically structured microfibers of "single stack" perylene bisimide and quaterthiophene nanowires. *ACS Nano*, 7(10):8498–8508, 2013.

[177] Dennis W. P. M. Löwik, E. H. P. Leunissen, M. van den Heuvel, M. B. Hansen, and Jan C. M. van Hest. Stimulus responsive peptide based materials. *Chemical Society Reviews*, 39(9):3394, 8 2010.

[178] Rein V. Ulijn and Andrew M. Smith. Designing peptide based nanomaterials. *Chemical Society Reviews*, 37(4):664, 3 2008.

[179] Amanda B. Marciel, Melikhan Tanyeri, Brian D. Wall, John D. Tovar, Charles M. Schroeder, and William L. Wilson. Fluidic-directed assembly of aligned oligopeptides with π-conjugated cores. *Advanced Materials*, 25(44):6398–6404, 11 2013.

[180] Matthew J. Webber, Eric A. Appel, E. W. Meijer, and Robert Langer. Supramolecular biomaterials. *Nature Materials*, 15(1):13–26, 1 2016.

[181] Albertus P. H. J. Schenning and E. W. Meijer. Supramolecular electronics; nanowires from self-assembled π-conjugated systems. *Chemical Communications*, (26):3245–3258, 6 2005.

[182] Miriam Mba, Alessandro Moretto, Lidia Armelao, Marco Crisma, Claudio Toniolo, and Michele Maggini. Synthesis and self-assembly of oligo(p-phenylenevinylene) peptide conjugates in water. *Chemistry - A European Journal*, 17(7):2044–2047, 2 2011.

[183] Joseph K. Gallaher, Emma J. Aitken, Robert A. Keyzers, and Justin M. Hodgkiss. Controlled aggregation of peptide-substituted perylene-bisimides. *Chemical Communications*, 48(64):7961, 7 2012.

[184] Antonio Facchetti. π-conjugated polymers for organic electronics and photovoltaic cell applications. *Chemistry of Materials*, 23(3):733–758, 2 2011.

[185] Brian D. Wall, Ashley E. Zacca, Allix M. Sanders, William L. Wilson, Andrew L. Ferguson, and John D. Tovar. Supramolecular Polymorphism: Tunable electronic interactions within π-conjugated peptide nanostructures dictated by primary amino acid sequence. *Langmuir*, 30(20):5946–5956, 5 2014.

[186] Pierre M. Beaujuge and John R. Reynolds. Color control in π-conjugated organic polymers for use in electrochromic devices. *Chemical Reviews*, 110(1):268–320, 1 2010.

[187] Herdeline Ann M. Ardoña and John D. Tovar. Energy transfer within responsive pi-conjugated coassembled peptide-based nanostructures in aqueous environments. *Chemical Science*, 6(2):1474–1484, 1 2015.

[188] Bryce A. Thurston, John D. Tovar, and Andrew L. Ferguson. Thermodynamics, morphology, and kinetics of early-stage self-assembly of π-conjugated oligopeptides. *Molecular Simulation*, 42(12):955–975, 8 2016.

[189] Allix M. Sanders, Tejaswini S. Kale, Howard E. Katz, and John D. Tovar. Solid-phase synthesis of self-assembling multivalent π-conjugated peptides. *ACS Omega*, 2(2):409–419, 2 2017.

[190] Bryce Thurston, Ethan Shapera, John D Tovar, André Schleife, and Andrew L Ferguson. Revealing the sequence-structure-electronic property relation of self-assembling π-conjugated oligopeptides by molecular and quantum mechanical modeling. *Langmuir*, 35(47):15221–15231, 2019.

[191] Brian D. Wall and John D. Tovar. Synthesis and characterization of π-conjugated peptide-based supramolecular materials. *Pure and Applied Chemistry*, 84(4):1039–1045, 3 2012.

[192] Bryce A. Thurston and Andrew L. Ferguson. Machine learning and molecular design of self-assembling -conjugated oligopeptides. *Molecular Simulation*, 44(11):930–945, 7 2018.

[193] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H De Vries. The martini force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.

[194] Luca Monticelli, Senthil K Kandasamy, Xavier Periole, Ronald G Larson, D Peter Tieleman, and Siewert-Jan Marrink. The martini coarse-grained force field: extension to proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834, 2008.

[195] Djurre H. de Jong, Gurpreet Singh, W. F. Drew Bennett, Clement Arnarez, Tsjerk A. Wassenaar, Lars V. Schäfer, Xavier Periole, D. Peter Tieleman, and Siewert J. Marrink. Improved parameters for the Martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1):687–697, 1 2013.

[196] Alfonso Gautieri, Antonio Russo, Simone Vesentini, Alberto Redaelli, and Markus J Buehler. Coarse-grained model of collagen molecules using an extended martini force field. *Journal of Chemical Theory and Computation*, 6(4):1210–1218, 2010.

[197] Martina Pannuzzo, Djurre H De Jong, Antonio Raudino, and Siewert J Marrink. Simulation of polyethylene glycol and calcium-mediated membrane fusion. *The Journal of Chemical Physics*, 140(12):124905, 2014.

[198] Cesar A López, Alex H De Vries, and Siewert J Marrink. Computational microscopy of cyclodextrin mediated cholesterol extraction from lipid model membranes. *Scientific Reports*, 3:2071, 2013.

[199] Cong Guo, Yin Luo, Ruhong Zhou, and Guanghong Wei. Probing the self-assembly mechanism of diphenylalanine-based peptide nanovesicles and nanotubes. *ACS Nano*, 6(5):3907–3918, 2012.

[200] Mikyung Seo, Sarah Rauscher, Régis Pomès, and D Peter Tieleman. Improving internal peptide dynamics in the coarse-grained martini model: toward large-scale simulations of amyloid-and elastin-like peptides. *Journal of Chemical Theory and Computation*, 8(5):1774–1785, 2012.

[201] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[202] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 1 2007.

[203] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 12 1981.

[204] Roger W Hockney and James W Eastwood. *Computer Simulation Ssing Particles*. CRC Press, 1988.

[205] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 9 1997.

[206] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[207] Travis E. Oliphant. *Guide to NumPy*. CreateSpace Independent Publishing Platform, 2nd edition, 2015.

[208] François Chollet. keras. `https://github.com/fchollet/keras`, 2015.

[209] Tomaž Hočevar and Janez Demšar. A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4):559–565, 2 2014.

[210] Jiang Wang and Andrew L. Ferguson. Mesoscale simulation of asphaltene aggregation. *The Journal of Physical Chemistry B*, 120(32):8016–8035, 8 2016.

[211] W Humphrey, A Dalke, and K Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–8, 2 1996.

[212] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Technical Report, Los Alamos National Lab.(LANL), Los Alamos, NM, 2008.

[213] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[214] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[215] Roberto Calandra, Jan Peters, Carl Edward Rasmussen, and Marc Peter Deisenroth. Manifold gaussian processes for regression. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3338–3345. IEEE, 2016.

[216] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[217] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Submission date: 22 Dec 2014, DOI:10.48550/arXiv.1412.6980.

[218] Hossein Mohammadi, Rodolphe Le Riche, Nicolas Durrande, Eric Touboul, and Xavier Bay. An analytic comparison of regularization methods for gaussian processes. *arXiv preprint arXiv:1602.00853*, 2016.

[219] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, pages 401–406, 1946.

[220] Michael Bloodgood and K Vijay-Shanker. Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets. *arXiv preprint arXiv:1409.4835*, 2014.

[221] Andrew W Long and Andrew L Ferguson. Rational design of patchy colloids via landscape engineering. *Molecular Systems Design & Engineering*, 3(1):49–65, 2018.

[222] J. Wang and A. L. Ferguson. Nonlinear machine learning in simulations of soft and biological materials. *Molecular Simulation*, 44(13-14):1090–1107, 9 2018.

[223] Yutao Ma and Andrew L Ferguson. Inverse design of self-assembling colloidal crystals with omnidirectional photonic bandgaps. *Soft Matter*, 15(43):8808–8826, 2019.

[224] N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 1 2007.

[225] David Aparício, Pedro Ribeiro, and Fernando Silva. Temporal network comparison using graphlet-orbit transitions. *arXiv preprint arXiv:1707.04572*, 2017.

[226] Nino Shervashidze and Karsten Borgwardt. Fast subtree kernels on graphs. In *Advances in Neural Information Processing Systems*, pages 1660–1668, 2009.

[227] Hisashi Kashima, Koji Tsuda, and Akihiro Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 321–328, 2003.

[228] Stephen Bonner, John Brennan, G Theodoropoulos, I Kureshi, and AS McGough. Efficient comparison of massive graphs through the use of "graph fingerprints". In *Twelfth Workshop on Mining and Learning with Graphs (MLG) '16 SanFrancisco, California USA*. ACM, August 2016.

[229] Wesley F. Reinhart and Athanassios Z. Panagiotopoulos. Automated crystal characterization with a fast neighborhood graph analysis method. *Soft Matter*, 14(29):6083–6089, 7 2018.

[230] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 7 2006.

[231] Jiang Wang, Mohit A. Gayatri, and Andrew L. Ferguson. Mesoscale simulation and machine learning of asphaltene aggregation phase behavior and molecular assembly landscapes. *The Journal of Physical Chemistry B*, 121(18):4923–4944, 5 2017.

[232] Boaz Nadler, Stephane Lafon, Ioannis Kevrekidis, and Ronald R Coifman. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *Advances in Neural Information Processing Systems*, pages 955–962, 2006.

[233] Jiang Wang and Andrew L. Ferguson. A study of the morphology, dynamics, and folding pathways of ring polymers with supramolecular topological constraints using molecular simulation and nonlinear manifold learning. *Macromolecules*, 51(2):598–616, 1 2018.

[234] Geeta S Vadehra, Brian D Wall, Stephen R Diegelmann, and John D Tovar. On-resin dimerization incorporates a diverse array of π-conjugated functionality within aqueous self-assembling peptide backbones. *Chemical Communications*, 46(22):3947–3949, 2010.

[235] Kalpana Besar. *Organic semiconductor devices for chemical sensing and bio interfaces*. PhD thesis, Johns Hopkins University, 2016.

[236] Andrew L Ferguson, Athanassios Z Panagiotopoulos, Pablo G Debenedetti, and Ioannis G Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proceedings of the National Academy of Sciences*, 107(31):13597–13602, 2010.

[237] Michael Rubinstein and Ralph H Colby. *Polymer Physics*. Oxford University Press, 2003.

[238] Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 3 1963.

[239] Kirill Shmilovich, Rachael A. Mansbach, and Andrew L. Ferguson. Dataset for discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation. *Materials Data Facility*, pages DOI: 10.18126/xqiz–hzc2, 2019.

[240] Ben Blaiszik, Kyle Chard, Jim Pruyne, Rachana Ananthakrishnan, Steven Tuecke, and Ian Foster. The materials data facility: Data services to advance materials science research. *JOM*, 68(8):2045–2052, 2016.

[241] Ben Blaiszik, Logan Ward, Marcus Schwarting, Jonathon Gaff, Ryan Chard, Daniel Pike, Kyle Chard, and Ian Foster. A data ecosystem to support machine learning in materials science. *MRS Communications*, 9(4):1125–1133, 2019.

[242] Pim WJM Frederix, Gary G Scott, Yousef M Abul-Haija, Daniela Kalafatovic, Charalampos G Pappas, Nadeem Javid, Neil T Hunt, Rein V Ulijn, and Tell Tuttle. Exploring the sequence space for (tri-) peptide self-assembly to design and discover new hydrogels. *Nature chemistry*, 7(1):30, 2015.

[243] Rainier Barrett and Andrew D. White. Investigating active learning and meta-learning for iterative peptide design. *Journal of Chemical Information and Modeling*, 61(1):95–105, 2021.

[244] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre,

Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

[245] Alexander van Teijlingen and Tell Tuttle. Beyond tripeptides two-step active machine learning for very large data sets. *Journal of Chemical Theory and Computation*, 17(5):3221–3232, 2021.

[246] Fei Li, Jinsong Han, Tian Cao, William Lam, Baoer Fan, Wen Tang, Sijie Chen, Kin Lam Fok, and Linxian Li. Design of self-assembly dipeptide hydrogels and machine learning via their chemical features. *Proceedings of the National Academy of Sciences*, 116(23):11259–11264, 2019.

[247] Shinji Nagasawa, Eman Al-Naamani, and Akinori Saeki. Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. *The Journal of Physical Chemistry Letters*, 9(10):2639–2646, 2018.

[248] Bryce A Thurston and Andrew L Ferguson. Machine learning and molecular design of self-assembling-conjugated oligopeptides. *Molecular Simulation*, 44(11):930–945, 2018.

[249] Prasanna V Balachandran, Benjamin Kowalski, Alp Sehirlioglu, and Turab Lookman. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nature Communications*, 9(1):1–9, 2018.

[250] Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin A Blood-Forsythe, Hyun Sik Chae, Markus Einzinger, Dong-Gwang Ha, Tony Wu, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials*, 15(10):1120–1127, 2016.

[251] Kirill Shmilovich, Sayak Subhra Panda, Anna Stouffer, John D. Tovar, and Andrew L. Ferguson. Supporting data for: "Hybrid computational-experimental data-driven design of self-assembling π-conjugated peptides", June 2021.

[252] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[253] Vilfredo Pareto. *Cours d'économie politique*, volume 1. Librairie Droz, 1964.

[254] Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019. Submission date: 29 Mar 2019, DOI:10.48550/arXiv.1903.12436.

[255] Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. A flexible framework for multi-objective bayesian optimization using random scalarizations. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial*

Intelligence Conference, volume 115 of *Proceedings of Machine Learning Research*, pages 766–776. PMLR, 22–25 Jul 2020.

[256] Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017.

[257] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[258] Pekka Mark and Lennart Nilsson. Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k. *The Journal of Physical Chemistry A*, 105(43):9954–9960, 2001.

[259] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010.

[260] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 2007.

[261] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[262] Doros N Theodorou and Ulrich W Suter. Shape of unperturbed linear polymers: polypropylene. *Macromolecules*, 18(6):1206–1214, 1985.

[263] Manfred K Warmuth, Jun Liao, Gunnar Rätsch, Michael Mathieson, Santosh Putta, and Christian Lemmen. Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences*, 43(2):667–673, 2003.

[264] Dante A Pertusi, Matthew E Moura, James G Jeffryes, Siddhant Prabhu, Bradley Walters Biggs, and Keith EJ Tyo. Predicting novel substrates for enzymes with minimal experimental effort with active learning. *Metabolic Engineering*, 44:171–181, 2017.

[265] Rocco Varela, W Patrick Walters, Brian B Goldman, and Ajay N Jain. Iterative refinement of a binding pocket model: active computational steering of lead optimization. *Journal of Medicinal Chemistry*, 55(20):8926–8942, 2012.

[266] Varun Khanna and Shoba Ranganathan. In silico approach to screen compounds active against parasitic nematodes of major socio-economic importance. In *BMC bioinformatics*, volume 12, pages 1–12. BioMed Central, 2011.

[267] Daniel Reker, Petra Schneider, and Gisbert Schneider. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chemical Science*, 7(6):3919–3927, 2016.

[268] Armaghan W Naik, Joshua D Kangas, Christopher J Langmead, and Robert F Murphy. Efficient modeling and active learning discovery of biological responses. *PLoS One*, 8(12):e83996, 2013.

[269] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. Submission date: 20 Dec 2013, DOI:10.48550/arXiv.1312.6114.

[270] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017.

[271] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[272] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[273] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

[274] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.

[275] Shuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. Aaindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(suppl_1):D202–D205, 2007.

[276] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[277] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[278] Andrew L Ferguson and Keith A Brown. Data-driven design and autonomous experimentation in soft and biological materials engineering. *Annual Review of Chemical and Biomolecular Engineering*, 13(2):1–20, 2022.

[279] David Ginsbourger, Rodolphe Le Riche, and Laurent Carraro. A multi-points criterion for deterministic parallel global optimization based on gaussian processes. *HAL preprint hal-00260579*, 2008.

[280] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[281] Boaz Nadler, Stephane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. *arXiv preprint math/0506090*, 2005.

[282] Andrew L Ferguson, Athanassios Z Panagiotopoulos, Pablo G Debenedetti, and Ioannis G Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proceedings of the National Academy of Sciences*, 107(31):13597–13602, 2010.

[283] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[284] Lauri Himanen, Marc OJ Jäger, Eiaki V Morooka, Filippo Federici Canova, Yashasvi S Ranawat, David Z Gao, Patrick Rinke, and Adam S Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

[285] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[286] Marc OJ Jäger, Eiaki V Morooka, Filippo Federici Canova, Lauri Himanen, and Adam S Foster. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials*, 4(1):1–8, 2018.

[287] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

[288] Jeffrey D Hartgerink, Elia Beniash, and Samuel I Stupp. Self-assembly and mineralization of peptide-amphiphile nanofibers. *Science*, 294(5547):1684–1688, 2001.

[289] Sayak Subhra Panda and John D Tovar. Aqueous self-assembly of peptide-diketopyrrolopyrrole conjugates with variation of n-alkyl side chain and π-core lengths. *Organic Materials*, (AAM), 2021.

[290] Sayak Subhra Panda, Kirill Shmilovich, Andrew L Ferguson, and John D Tovar. Computationally guided tuning of amino acid configuration influences the chiroptical properties of supramolecular peptide-π-peptide nanostructures. *Langmuir*, 36(24):6782–6792, 2020.

[291] Giuseppe Paradies, Valeria Paradies, Valentina De Benedictis, Francesca M Ruggiero, and Giuseppe Petrosillo. Functional role of cardiolipin in mitochondrial bioenergetics. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1837(4):408–417, 2014.

[292] Federico Elías-Wolff, Martin Lindén, Alexander P Lyubartsev, and Erik G Brandt. Curvature sensing by cardiolipin in simulated buckled membranes. *Soft Matter*, 15(4):792–802, 2019.

[293] Edward Ross Pennington, Katsuhiko Funai, David A Brown, and Saame Raza Shaikh. The role of cardiolipin concentration and acyl chain composition on mitochondrial inner membrane molecular organization and function. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1864(7):1039–1052, 2019.

[294] Christopher A Lipinski. Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337–341, 2004.

[295] Andreas Uphoff, Martin Hermansson, Perttu Haimi, and Pentti Somerharju. Chapter 11 - analysis of complex lipidomes. In Károly Vékey, András Telekes, and Akos Vertes, editors, *Medical Applications of Mass Spectrometry*, pages 223–249. Elsevier, Amsterdam, 2008.

[296] HH Szeto and AV Birk. Serendipity and the discovery of novel compounds that restore mitochondrial plasticity. *Clinical Pharmacology & Therapeutics*, 96(6):672–683, 2014.

[297] AV Birk, WM Chao, C Bracken, JD Warren, and HH Szeto. Targeting mitochondrial cardiolipin and the cytochrome c/cardiolipin complex to promote electron transport and optimize mitochondrial atp synthesis. *British Journal of Pharmacology*, 171(8):2017–2028, 2014.

[298] Hazel H Szeto and Shaoyi Liu. Cardiolipin-targeted peptides rejuvenate mitochondrial function, remodel mitochondria, and promote tissue regeneration during aging. *Archives of Biochemistry and Biophysics*, 660:137–148, 2018.

[299] Frits A de Wolf. Binding of doxorubicin to cardiolipin as compared to other anionic phospholipids—an evaluation of electrostatic effects. *Bioscience Reports*, 11(5):275–284, 1991.

[300] Kyrylo Pyrshev, Semen O Yesylevskyy, and Mikhail Bogdanov. Ttape-me dye is not selective to cardiolipin and responds to main anionic phospholipids unspecifically. *bioRxiv*, 2020.

[301] Jake Jacobson, Michael R Duchen, and Simon JR Heales. Intracellular distribution of the fluorescent dye nonyl acridine orange responds to the mitochondrial membrane potential: implications for assays of cardiolipin and mitochondrial mass. *Journal of Neurochemistry*, 82(2):224–233, 2002.

[302] Vishal M Gohil, Jelena Gvozdenovic-Jeremic, Michael Schlame, and Miriam L Greenberg. Binding of 10-n-nonyl acridine orange to cardiolipin-deficient yeast cells: implications for assay of cardiolipin. *Analytical Biochemistry*, 343(2):350–352, 2005.

[303] Piercen M Oliver, John A Crooks, Mathias Leidl, Earl J Yoon, Alan Saghatelian, and Douglas B Weibel. Localization of anionic phospholipids in escherichia coli cells. *Journal of Bacteriology*, 196(19):3386–3398, 2014.

[304] Alex-Rose Pogmore, Kenneth H Seistrup, and Henrik Strahl. The gram-positive model organism bacillus subtilis does not form microscopically detectable cardiolipin-specific lipid domains. *Microbiology*, 164(4):475–482, 2018.

[305] Simona Lobasso, Matilde Sublimi Saponetti, Francesco Polidoro, Patrizia Lopalco, Jasna Urbanija, Veronika Kralj-Iglic, and Angela Corcelli. Archaebacterial lipid membranes as models to study the interaction of 10-n-nonyl acridine orange with phospholipids. *Chemistry and Physics of Lipids*, 157(1):12–20, 2009.

[306] Advanced Chemistry Development, Inc., Toronto, ON, Canada. Acd/chemsketch version 2021.1.0. `www.acdlabs.com`, 2021.

[307] Julian Michalowsky, Johannes Zeman, Christian Holm, and Jens Smiatek. A polarizable martini model for monovalent ions in aqueous solution. *The Journal of Chemical Physics*, 149(16):163319, 2018.

[308] Julian Michalowsky, Lars V Schäfer, Christian Holm, and Jens Smiatek. A refined polarizable water model for the coarse-grained martini force field with long-range electrostatic interactions. *The Journal of Chemical Physics*, 146(5):054501, 2017.

[309] WF Drew Bennett and D Peter Tieleman. Water defect and pore formation in atomistic and coarse-grained lipid membranes: pushing the limits of coarse graining. *Journal of Chemical Theory and Computation*, 7(9):2981–2988, 2011.

[310] Hyung-June Woo and Anders Wallqvist. Spontaneous buckling of lipid bilayer and vesicle budding induced by antimicrobial peptide magainin 2: a coarse-grained simulation study. *The Journal of Physical Chemistry B*, 115(25):8122–8129, 2011.

[311] Alessia Centi, Arghya Dutta, Sapun H Parekh, and Tristan Bereau. Inserting small molecules across membrane mixtures: Insight from the potential of mean force. *Biophysical Journal*, 118(6):1321–1332, 2020.

[312] Jad Eid, Hafez Razmazma, Alia Jraij, Ali Ebrahimi, and Luca Monticelli. On calculating the bending modulus of lipid bilayer membranes from buckling simulations. *The Journal of Physical Chemistry B*, 124(29):6299–6311, 2020.

[313] Roberto Menichetti, Kurt Kremer, and Tristan Bereau. Efficient potential of mean force calculation from multiscale simulations: solute insertion in a lipid membrane. *Biochemical and Biophysical Research Communications*, 498(2):282–287, 2018.

[314] Roberto Menichetti, Kiran H Kanekal, and Tristan Bereau. Drug–membrane permeability across chemical space. *ACS Central Science*, 5(2):290–298, 2019.

[315] Roberto Menichetti, Kiran H Kanekal, Kurt Kremer, and Tristan Bereau. In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force. *The Journal of Chemical Physics*, 147(12):125101, 2017.

[316] Hossein Nejat Pishkenari, Mohammad Reza Barzegar, and Ali Taghibakhshi. Study and simulation of nanoparticle translocation through cell membrane. *Iranian Journal of Science and Technology, Transactions of Mechanical Engineering*, pages 1–22, 2019.

[317] Isabel S Kleinwächter, Stefanie Pannwitt, Alessia Centi, Nadja Hellmann, Eckhard Thines, Tristan Bereau, and Dirk Schneider. The bacteriostatic activity of 2-phenylethanol derivatives correlates with membrane binding affinity. *Membranes*, 11(4):254, 2021.

[318] Djurre H de Jong, Gurpreet Singh, WF Drew Bennett, Clement Arnarez, Tsjerk A Wassenaar, Lars V Schafer, Xavier Periole, D Peter Tieleman, and Siewert J Marrink. Improved parameters for the martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1):687–697, 2013.

[319] Juan Felipe Franco-Gonzalez, Javier Ramos, Victor L Cruz, and Javier Martinez-Salazar. Exploring the dynamics and interaction of a full erbb2 receptor and trastuzumab-fab antibody in a lipid bilayer model using martini coarse-grained force field. *Journal of Computer-Aided Molecular Design*, 28(11):1093–1107, 2014.

[320] Lea Thøgersen, Birgit Schiøtt, Thomas Vosegaard, Niels Chr Nielsen, and Emad Tajkhorshid. Peptide aggregation and pore formation in a lipid bilayer: a combined coarse-grained and all atom molecular dynamics study. *Biophysical Journal*, 95(9):4337–4347, 2008.

[321] Anna L Duncan, Alan J Robinson, and John E Walker. Cardiolipin binds selectively but transiently to conserved lysine residues in the rotor of metazoan atp synthases. *Proceedings of the National Academy of Sciences*, 113(31):8687–8692, 2016.

[322] Yifei Qi, Helgi I Ingólfsson, Xi Cheng, Jumin Lee, Siewert J Marrink, and Wonpil Im. Charmm-gui martini maker for coarse-grained simulations with the martini force field. *Journal of Chemical Theory and Computation*, 11(9):4486–4494, 2015.

[323] Gerd Olofsson and Emma Sparr. Ionization constants pka of cardiolipin. *PloS One*, 8(9):e73040, 2013.

[324] Murugappan Sathappa and Nathan N Alder. The ionization properties of cardiolipin and its variants in model bilayers. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1858(6):1362–1372, 2016.

[325] Lee Serpas, Bridget Milorey, Leah A Pandiscia, Anthony W Addison, and Reinhard Schweitzer-Stenner. Autoxidation of reduced horse heart cytochrome c catalyzed by cardiolipin-containing membranes. *The Journal of Physical Chemistry B*, 120(48):12219–12231, 2016.

[326] Djurre H De Jong, Svetlana Baoukina, Helgi I Ingólfsson, and Siewert J Marrink. Martini straight: Boosting performance using a shorter cutoff and gpus. *Computer Physics Communications*, 199:1–7, 2016.

[327] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[328] Tijana Milenković and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:CIN–S680, 2008.

[329] Tobias Fink and Jean-Louis Reymond. Virtual exploration of the chemical universe up to 11 atoms of c, n, o, f: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *Journal of Chemical Information and Modeling*, 47(2):342–353, 2007.

[330] ChemAxon. Calculator plugin of marvin 17.28.0, 2017.

[331] Peter Ertl. An algorithm to identify functional groups in organic molecules. *Journal of Cheminformatics*, 9(1):1–7, 2017.

[332] RDKit: Open-source cheminformatics. `http://www.rdkit.org`. [Online; accessed 03-May-2021].

[333] Peter Ertl, Eva Altmann, and Jeffrey M McKenna. The most common functional groups in bioactive molecules and how their popularity has evolved over time. *Journal of Medicinal Chemistry*, 63(15):8408–8418, 2020.

[334] Bernadette Mohr, Kirill Shmilovich, Isabel Kleinwächter, Dirk Schneider, Andrew L. Ferguson, and Tristan Bereau. Supporting data for: "Data-driven discovery of cardiolipin-selective small molecules by computational active learning", November 2021.

[335] Dmitry Malyshka, Leah A Pandiscia, and Reinhard Schweitzer-Stenner. Cardiolipin containing liposomes are fully ionized at physiological ph. an ft-ir study of phosphate group ionization. *Vibrational Spectroscopy*, 75:86–92, 2014.

[336] Patrice X Petit, Pauline Dupaigne, Fabrizio Pariselli, François Gonzalvez, François Etienne, Carole Rameau, and Sophie Bernard. Interaction of the alpha-helical h6 peptide from the pro-apoptotic protein tbid with cardiolipin. *The FEBS Journal*, 276(21):6338–6354, 2009.

[337] Jacek Zielonka, Joy Joseph, Adam Sikora, Micael Hardy, Olivier Ouari, Jeannette Vasquez-Vivar, Gang Cheng, Marcos Lopez, and Balaraman Kalyanaraman. Mitochondria-targeted triphenylphosphonium-based compounds: syntheses, mechanisms of action, and therapeutic and diagnostic applications. *Chemical Reviews*, 117(15):10043–10120, 2017.

[338] Matteo Aldeghi, Vytautas Gapsys, and Bert L de Groot. Accurate estimation of ligand binding affinity changes upon protein mutation. *ACS Central Science*, 4(12):1708–1718, 2018.

[339] Bin Liu Liu, editor. *Conjugated Polymers for Biological and Biomedical Applications*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, May 2018.

[340] Rajib Mondal, Sangwon Ko, and Zhenan Bao. Fused aromatic thienopyrazines: structure, properties and function. *Journal of Materials Chemistry*, 20(47):10568, 2010.

[341] Luyao Lu, Tianyue Zheng, Qinghe Wu, Alexander M. Schneider, Donglin Zhao, and Luping Yu. Recent Advances in Bulk Heterojunction Polymer Solar Cells. *Chemical Reviews*, 115(23):12666–12731, December 2015.

[342] Yonglan Liu, Dong Zhang, Baiping Ren, Xiong Gong, Lijian Xu, Zhang-Qi Feng, Yung Chang, Yi He, and Jie Zheng. Molecular simulations and understanding of antifouling zwitterionic polymer brushes. *Journal of Materials Chemistry B*, 8(17):3814–3828, 2020.

[343] Yonglan Liu, Dong Zhang, Yijing Tang, Yanxian Zhang, Xiong Gong, Shaowen Xie, and Jie Zheng. Machine learning-enabled repurposing and design of antifouling polymer brushes. *Chemical Engineering Journal*, 420:129872, 2021.

[344] Rachael A Mansbach and Andrew L Ferguson. Coarse-grained molecular simulation of the hierarchical self-assembly of π-conjugated optoelectronic peptides. *The Journal of Physical Chemistry B*, 121(7):1684–1706, 2017.

[345] Rachael A Mansbach and Andrew L Ferguson. Patchy particle model of the hierarchical self-assembly of π-conjugated optoelectronic peptides. *The Journal of Physical Chemistry B*, 122(44):10219–10236, 2018.

[346] Kirill Shmilovich, Sayak Panda, Anna Stouffer, John Tovar, and Andrew Ferguson. Hybrid computational-experimental data-driven design of self-assembling pi-conjugated peptides. *ChemRxiv preprint: 10.33774/chemrxiv-2021-l42ch*, 2021.

[347] Harald Oberhofer, Karsten Reuter, and Jochen Blumberger. Charge Transport in Molecular Materials: An Assessment of Computational Methods. *Chemical Reviews*, 117(15):10319–10357, August 2017.

[348] Ilhan Yavuz, Blanton N. Martin, Jiyong Park, and K. N. Houk. Theoretical Study of the Molecular Ordering, Paracrystallinity, And Charge Mobilities of Oligomers in

Different Crystalline Phases. *Journal of the American Chemical Society*, 137(8):2856–2866, March 2015.

[349] Ganna Gryn'ova, Kun-Han Lin, and Clémence Corminboeuf. Read between the Molecules: Computational Insights into Organic Semiconductors. *Journal of the American Chemical Society*, 140(48):16370–16386, December 2018.

[350] R A Marcus. Chemical and Electrochemical Electron-Transfer Theory. *Annual Review of Physical Chemistry*, 15(1):155–196, October 1964.

[351] Yuri A. Berlin, Geoffrey R. Hutchison, Pawel Rempala, Mark A. Ratner, and Josef Michl. Charge Hopping in Molecular Wires as a Sequence of Electron-Transfer Reactions. *The Journal of Physical Chemistry A*, 107(19):3970–3980, May 2003.

[352] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[353] Koichi Momma and Fujio Izumi. ıt VESTA3 for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of Applied Crystallography*, 44(6):1272–1276, December 2011.

[354] Christopher A Reynolds, Jonathan W Essex, and W Graham Richards. Atomic charges for variable molecular conformations. *Journal of the American Chemical Society*, 114(23):9075–9079, 1992.

[355] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, and Peter A Kollman. Application of resp charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society*, 115(21):9620–9631, 2002.

[356] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian 16 Revision C.01, 2016.

[357] Alan W Sousa Da Silva and Wim F Vranken. Acpype-antechamber python parser interface. *BMC Research Notes*, 5(1):1–8, 2012.

[358] David Case, D.S. Cerutti, Thomas Cheatham, Thomas Darden, Robert Duke, T.J. Giese, H. Gohlke, Andreas Götz, D. Greene, N. Homeyer, Saeed Izadi, Andriy Kovalenko, Tai-Sung Lee, S. LeGrand, Pengfei Li, C. Lin, J. Liu, Tyler Luchko, Ray Luo, and P.A. Kollman. Amber 2017, university of california, san francisco, 04 2017.

[359] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[360] U Essman, L Perera, ML Berkowitz, T Darden, H Lee, and LG Pedersen. A smooth particle mesh ewald potential. *J. Chem. Phys*, 103:8577–8592, 1995.

[361] Herdeline Ann Mallari Ardoña. *MOLECULAR ENGINEERING STRATEGIES FOR THE DEVELOPMENT OF ENERGY-TRANSPORTING CONJUGATED SYSTEMS TOWARDS BIOELECTRONICS*. PhD thesis, Johns Hopkins University, 2017.

[362] Herdeline Ann M Ardoña, Emily R Draper, Francesca Citossi, Matthew Wallace, Louise C Serpell, Dave J Adams, and John D Tovar. Kinetically controlled coassembly of multichromophoric peptide hydrogelators and the impacts on energy transport. *Journal of the American Chemical Society*, 139(25):8685–8692, 2017.

[363] Krešimir Molčanov, Valentina Milašinović, and Biserka Kojić-Prodić. Contribution of different crystal packing forces in π-stacking: from noncovalent to covalent multicentric bonding. *Crystal Growth & Design*, 19(10):5967–5980, 2019.

[364] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum Mechanical Continuum Solvation Models. *Chemical Reviews*, 105(8):2999–3094, August 2005.

[365] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, January 1988.

[366] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *Journal of Chemical Physics*, 98(7):5648–5652, April 1993.

[367] D.M. de Leeuw, M.M.J. Simenon, A.R. Brown, and R.E.F. Einerhand. Stability of n-type doped conducting polymers and consequences for polymeric microelectronic devices. *Synthetic Metals*, 87(1):53–59, February 1997.

[368] Riccardo Di Pietro, Daniele Fazzi, Tom B. Kehoe, and Henning Sirringhaus. Spectroscopic Investigation of Oxygen- and Water-Induced Electron Trapping and Charge Transport Instabilities in n-type Polymer Semiconductors. *Journal of the American Chemical Society*, 134(36):14877–14889, September 2012.

[369] Steven G. Bratsch. Standard Electrode Potentials and Temperature Coefficients in Water at 298.15 K. *Journal of Physical and Chemical Reference Data*, 18(1):1–21, January 1989.

[370] Omar López-Estrada, Humberto G. Laguna, Cihuapilli Barrueta-Flores, and Carlos Amador-Bedolla. Reassessment of the Four-Point Approach to the Electron-Transfer Marcus–Hush Theory. *ACS Omega*, 3(2):2130–2140, February 2018.

[371] Jean-Luc Brédas, David Beljonne, Veaceslav Coropceanu, and Jérôme Cornil. Charge-Transfer and Energy-Transfer Processes in π-Conjugated Oligomers and Polymers: A Molecular Picture. *Chemical Reviews*, 104(11):4971–5004, November 2004.

[372] Subhajyoti Chaudhuri, Svante Hedström, Dalvin D. Méndez-Hernández, Heidi P. Hendrickson, Kenneth A. Jung, Junming Ho, and Victor S. Batista. Electron Transfer Assisted by Vibronic Coupling from Multiple Modes. *Journal of Chemical Theory and Computation*, 13(12):6000–6009, December 2017.

[373] Björn Baumeier, James Kirkpatrick, and Denis Andrienko. Density-functional based determination of intermolecular charge transfer properties for large-scale morphologies. *Physical Chemistry Chemical Physics*, 12(36):11103, 2010.

[374] Jesse T. E. Quinn, Jiaxin Zhu, Xu Li, Jinliang Wang, and Yuning Li. Recent progress in the development of n-type organic semiconductors for organic field effect transistors. *Journal of Materials Chemistry C*, 5(34):8654–8681, 2017.

[375] Ute Zschieschang, Tatsuya Yamamoto, Kazuo Takimiya, Hirokazu Kuwabara, Masaaki Ikeda, Tsuyoshi Sekitani, Takao Someya, and Hagen Klauk. Organic Electronics on Banknotes. *Advanced Materials*, 23(5):654–658, February 2011.

[376] Tsuyoshi Sekitani, Tomoyuki Yokota, Kazunori Kuribara, Martin Kaltenbrunner, Takanori Fukushima, Yusuke Inoue, Masaki Sekino, Takashi Isoyama, Yusuke Abe, Hiroshi Onodera, and Takao Someya. Ultraflexible organic amplifier with biocompatible gel electrodes. *Nature Communications*, 7(1):11425, September 2016.

[377] Gaetano Ricci, Sofia Canola, Yasi Dai, Daniele Fazzi, and Fabrizia Negri. Impact of Fluoroalkylation on the n-Type Charge Transport of Two Naphthodithiophene Diimide Derivatives. *Molecules*, 26(14):4119, July 2021.

[378] Luke R Fleet, James Stott, Byron Villis, Salahud Din, Michele Serri, Gabriel Aeppli, Sandrine Heutz, and Arokia Nathan. Self-assembled molecular nanowires for high-performance organic transistors. *ACS Applied Materials & Interfaces*, 9(24):20686–20695, 2017.

[379] Yan Zhou, Ting Lei, Lei Wang, Jian Pei, Yong Cao, and Jian Wang. High-performance organic field-effect transistors from organic single-crystal microribbons formed by a solution process. *Advanced Materials*, 22(13):1484–1487, 2010.

[380] Kirill Shmilovich, Yifan Yao, John D. Tovar, Howard E. Katz, André Schleife, and Andrew L. Ferguson. Supplemental data for "computational discovery of high charge mobility self-assembling π-conjugated peptides", 2022.

[381] Ayori Mitsutake, Yoshiharu Mori, and Yuko Okamoto. Enhanced sampling algorithms. In *Biomolecular Simulations*, pages 153–195. Springer, 2013.

[382] Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta, Gen. Subj.*, 1850(5):872–877, 2015.

[383] Yuko Okamoto. Generalized-ensemble algorithms: Enhanced sampling techniques for monte carlo and molecular dynamics simulations. *J. Mol. Graphics Modell.*, 22(5):425–439, 2004.

[384] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[385] David E Shaw, Ron O Dror, John K Salmon, JP Grossman, Kenneth M Mackenzie, Joseph A Bank, Cliff Young, Martin M Deneroff, Brannon Batson, Kevin J Bowers, et al. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the conference on high performance computing networking, storage and analysis*, pages 1–11, 2009.

[386] Trygve Helgaker, Poul Jorgensen, and Jeppe Olsen. *Molecular Electronic-Structure Theory*. John Wiley & Sons, 2014.

[387] Donald Allan McQuarrie and John Douglas Simon. *Physical Chemistry: A Molecular Approach*, volume 1. University science books Sausalito, CA, 1997.

[388] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1. Elsevier, 2001.

[389] John P Bergsma, Peter H Berens, Kent R Wilson, Donald R Fredkin, and Eric J Heller. Electronic spectra from molecular dynamics: a simple approach. *J. Phys. Chem.*, 88(3):612–619, 1984.

[390] Bertrand Guillot. A molecular dynamics study of the far infrared spectrum of liquid water. *J. Chem. Phys.*, 95(3):1543–1551, 1991.

[391] J. J. Salacuse, A. R. Denton, and P. A. Egelstaff. Finite-size effects in molecular dynamics simulations: Static structure factor and compressibility. i. theoretical method. *Phys. Rev. E*, 53:2382–2389, Mar 1996.

[392] Neil E Moe and MD Ediger. Calculation of the coherent dynamic structure factor of polyisoprene from molecular dynamics simulations. *Phys. Rev. E*, 59(1):623, 1999.

[393] Chunxia Chen, Praveen Depa, Janna K Maranas, and Victoria Garcia Sakai. Comparison of explicit atom, united atom, and coarse-grained simulations of poly (methyl methacrylate). *J. Chem. Phys.*, 128(12):124906, 2008.

[394] Arantxa Arbe, Fernando Alvarez, and Juan Colmenero. Neutron scattering and molecular dynamics simulations: Synergetic tools to unravel structure and dynamics in polymers. *Soft Matter*, 8(32):8257–8270, 2012.

[395] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[396] Kirill Shmilovich, Marc Stieffenhofer, Nicholas E. Charron, and Moritz Hoffmann. Supporting data for "temporally coherent backmapping of molecular trajectories from coarse-grain to atomistic resolution", 2022. (accessed May 11, 2022).

[397] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019. Submission date: 25 Mar 2019, DOI:10.48550/arXiv.1903.10145.

[398] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100(2):020603, 2008.

[399] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, 1977.

[400] Omar Valsson and Michele Parrinello. Variational approach to enhanced sampling and free energy calculations. *Phys. Rev. Lett.*, 113(9):090601, 2014.

[401] João Marcelo Lamim Ribeiro, Pablo Bravo, Yihang Wang, and Pratyush Tiwary. Reweighted autoencoded variational bayes for enhanced sampling (rave). *J. Chem. Phys.*, 149(7):072301, 2018.

[402] Wenwei Zheng, Mary A Rohrdanz, and Cecilia Clementi. Rapid exploration of configuration space with diffusion-map-directed molecular dynamics. *J. Phys. Chem. B*, 117(42):12769–12776, 2013.

[403] Eliodoro Chiavazzo, Roberto Covino, Ronald R Coifman, C William Gear, Anastasia S Georgiou, Gerhard Hummer, and Ioannis G Kevrekidis. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proc. Natl. Acad. Sci. U. S. A.*, 114(28):E5494–E5503, 2017.

[404] Wei Chen and Andrew L Ferguson. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.*, 39(25):2079–2102, 2018.

[405] Gregory R Bowman, Daniel L Ensign, and Vijay S Pande. Enhanced modeling via network theory: Adaptive sampling of markov state models. *J. Chem. Theory Comput.*, 6(3):787–794, 2010.

[406] Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.*, 13(7):e1005659, 2017.

[407] Feliks Nüske, Hao Wu, Jan-Hendrik Prinz, Christoph Wehmeyer, Cecilia Clementi, and Frank Noé. Markov state models from short non-equilibrium simulations—analysis and correction of estimation bias. *J. Chem. Phys.*, 146(9):094104, 2017.

[408] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[409] Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E Charron, Gianni De Fabritiis, Frank Noé, and Cecilia Clementi. Machine learning of coarse-grained molecular dynamics force fields. *ACS Cent. Sci.*, 5(5):755–767, 2019.

[410] Ignasi Buch, Matt J Harvey, Toni Giorgino, David P Anderson, and Gianni De Fabritiis. High-throughput all-atom molecular dynamics simulations using distributed computing. *J. Chem. Inf. Model.*, 50(3):397–403, 2010.

[411] S Doerr and G De Fabritiis. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J. Chem. Theory Comput.*, 10(5):2064–2069, 2014.

[412] Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.*, 148(24):241703, 2018.

[413] Feliks Nuske, Bettina G Keller, Guillermo Pérez-Hernández, Antonia SJS Mey, and Frank Noé. Variational approach to molecular kinetics. *J. Chem. Theory Comput.*, 10(4):1739–1752, 2014.

[414] Francesca Vitalini, Antonia SJS Mey, Frank Noé, and Bettina G Keller. Dynamic properties of force fields. *J. Chem. Phys.*, 142(8):02B611_1, 2015.

[415] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134(17):174105, 2011.

[416] Gregory R Bowman, Vijay S Pande, and Frank Noé. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, volume 797. Springer Science & Business Media, 2013.

[417] Benjamin Trendelkamp-Schroer, Hao Wu, Fabian Paul, and Frank Noé. Estimation and uncertainty of reversible markov models. *J. Chem. Phys.*, 143(17):11B601_1, 2015.

[418] Martin K Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. Pyemma 2: A software package for estimation, validation, and analysis of markov models. *J. Chem. Theory Comput.*, 11(11):5525–5542, 2015.

[419] Moritz Hoffmann, Martin Scherer, Tim Hempel, Andreas Mardt, Brian de Silva, Brooke E Husic, Stefan Klus, Hao Wu, Nathan Kutz, Steven L Brunton, et al. Deeptime: A python library for machine learning dynamical models from time series data. *Mach. Learn.: Sci. Technol.*, 3(1):015009, 2021.

[420] Dominik Fritz, Konstantin Koschke, Vagelis A Harmandaris, Nico FA van der Vegt, and Kurt Kremer. Multiscale modeling of soft matter: Scaling of dynamics. *Phys. Chem. Chem. Phys.*, 13(22):10412–10420, 2011.

[421] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *J. Chem. Phys.*, 139(1):07B604_1, 2013.

[422] Hao Wu and Frank Noé. Variational approach for learning markov processes from time series data. *J. Nonlinear Sci.*, 30(1):23–66, 2020.

[423] Frank Noé and Cecilia Clementi. Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.*, 11(10):5002–5011, 2015.

[424] Yaoyi Chen, Andreas Krämer, Nicholas E Charron, Brooke E Husic, Cecilia Clementi, and Frank Noé. Machine learning implicit solvation for molecular dynamics. *J. Chem. Phys.*, 155(8):084101, 2021.

[425] Jonas Köhler, Yaoyi Chen, Andreas Krämer, Cecilia Clementi, and Frank Noé. Force-matching coarse-graining without forces. *arXiv preprint arXiv:2203.11167*, 2022. Submission date: 21 Mar 2022, DOI:10.48550/arXiv.2203.11167.

[426] Luigi Bonati, GiovanniMaria Piccini, and Michele Parrinello. Deep learning the slow modes for rare events sampling. *Proc. Natl. Acad. Sci. U. S. A.*, 118(44):e2113533118, 2021.

[427] Maximilian Topel and Andrew L Ferguson. Reconstruction of protein structures from single-molecule time series. *J. Chem. Phys.*, 153(19):194102, 2020.

[428] Mohammad M Sultan and Vijay S Pande. Automated design of collective variables using supervised machine learning. *J. Chem. Phys.*, 149(9):094106, 2018.

[429] Benjamin Kurt Miller, Mario Geiger, Tess E Smidt, and Frank Noé. Relevance of rotationally equivariant convolutions for predicting molecular properties. *arXiv preprint arXiv:2008.08461*, 2020. Submission date: 19 Aug 2020, DOI:10.48550/arXiv.2008.08461.

[430] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. Submission date: 22 Feb 2018, DOI:10.48550/arXiv.1802.08219.

[431] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.*, 13(1):1–11, 2022.

[432] Hythem Sidky, Wei Chen, and Andrew L Ferguson. Molecular latent space simulators. *Chem. Sci.*, 11(35):9459–9467, 2020.

[433] Daan Frenkel, Berend Smit, and Mark A Ratner. *Understanding molecular simulation: from algorithms to applications*, volume 2. Academic press San Diego, 1996.

[434] E Weinan, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66(5):052301, 2002.

[435] Peter G Bolhuis, David Chandler, Christoph Dellago, and Phillip L Geissler. Transition path sampling: Throwing ropes. *Annu. Rev. Phys. Chem*, 53:291–318, 2002.

[436] Rosalind J Allen, Chantal Valeriani, and Pieter Rein Ten Wolde. Forward flux sampling for rare event simulations. *J. Phys. Condens. Matter*, 21(46):463102, 2009.

[437] Arthur F Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.*, 78(20):3908, 1997.

[438] TP Straatsma and HJC Berendsen. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *J. Chem. Phys.*, 89(9):5876–5886, 1988.

[439] Zahra Shamsi, Kevin J Cheng, and Diwakar Shukla. Reinforcement learning based adaptive sampling: Reaping rewards by exploring protein conformational landscapes. *J. Phys. Chem. B*, 122(35):8386–8395, 2018.

[440] Frank Noé. Machine learning for molecular dynamics on long timescales. In *Machine learning meets quantum physics*, pages 331–372. Springer, 2020.

[441] Yusuke Naritomi and Sotaro Fuchigami. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *J. Chem. Phys.*, 134(6):02B617, 2011.

[442] Christian R Schwantes and Vijay S Pande. Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9. *J. Chem. Theory Comput.*, 9(4):2000–2009, 2013.

[443] Christian R Schwantes and Vijay S Pande. Modeling molecular kinetics with tica and the kernel trick. *J. Chem. Theory Comput.*, 11(2):600–608, 2015.

[444] Wei Chen, Hythem Sidky, and Andrew L Ferguson. Nonlinear discovery of slow molecular modes using state-free reversible vampnets. *J. Chem. Phys.*, 150(21):214114, 2019.

[445] Hythem Sidky, Wei Chen, and Andrew L Ferguson. High-resolution markov state models for the dynamics of trp-cage miniprotein constructed over slow folding modes identified by state-free reversible vampnets. *J. Phys. Chem. B*, 123(38):7999–8009, 2019.

[446] Michael S Jones, Brennan Ashwood, Andrei Tokmakoff, and Andrew L Ferguson. Determining sequence-dependent dna oligonucleotide hybridization and dehybridization mechanisms using coarse-grained molecular simulation, markov state models, and infrared spectroscopy. *J. Am. Chem. Soc.*, 143(42):17395–17411, 2021.

[447] John D Chodera, William C Swope, Frank Noé, Jan-Hendrik Prinz, Michael R Shirts, and Vijay S Pande. Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures. *J. Chem. Phys.*, 134(24):06B612, 2011.

[448] Hao Wu, Antonia SJS Mey, Edina Rosta, and Frank Noé. Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states. *J. Chem. Phys.*, 141(21):12B629_1, 2014.

[449] Hao Wu, Fabian Paul, Christoph Wehmeyer, and Frank Noé. Multiensemble markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci. U.S.A.*, 113(23):E3221–E3230, 2016.

[450] Edina Rosta and Gerhard Hummer. Free energies from dynamic weighted histogram analysis using unbiased markov state model. *J. Chem. Theory Comput.*, 11(1):276–285, 2015.

[451] Pratyush Tiwary and Michele Parrinello. From metadynamics to dynamics. *Phys. Rev. Lett.*, 111(23):230602, 2013.

[452] Rodrigo Casasnovas, Vittorio Limongelli, Pratyush Tiwary, Paolo Carloni, and Michele Parrinello. Unbinding kinetics of a p38 map kinase type ii inhibitor from metadynamics simulations. *J. Am. Chem. Soc.*, 139(13):4780–4788, 2017.

[453] Jesús A Izaguirre, Christopher R Sweet, and Vijay S Pande. Multiscale dynamics of macromolecules using normal mode langevin. In *Biocomputing 2010*, pages 240–251. World Scientific, 2010.

[454] Jan-Hendrik Prinz, John D Chodera, Vijay S Pande, William C Swope, Jeremy C Smith, and Frank Noé. Optimal use of data in parallel tempering simulations for the construction of discrete-state markov models of biomolecular dynamics. *J. Chem. Phys.*, 134(24):06B613, 2011.

[455] Pantelis R Vlachas, Julija Zavadlav, Matej Praprotnik, and Petros Koumoutsakos. Accelerated simulations of molecular systems through learning of effective dynamics. *J. Chem. Theory Comput.*, 18(1):538–549, 2021.

[456] Stefan Klus, Feliks Nüske, Péter Koltai, Hao Wu, Ioannis Kevrekidis, Christof Schütte, and Frank Noé. Data-driven model reduction and transfer operator approximation. *J. Nonlinear Sci.*, 28(3):985–1010, 2018.

[457] Péter Koltai, Hao Wu, Frank Noé, and Christof Schütte. Optimal data-driven estimation of generalized markov state models for non-equilibrium dynamics. *Computation*, 6(1):22, 2018.

[458] Frank Noé and Feliks Nuske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model Simul.*, 11(2):635–655, 2013.

[459] Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[460] David S Watkins. *The matrix eigenvalue problem: GR and Krylov subspace methods*. SIAM, 2007.

[461] Vijay S Pande, Kyle Beauchamp, and Gregory R Bowman. Everything you wanted to know about markov state models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

[462] Christoph Wehmeyer, Martin K Scherer, Tim Hempel, Brooke E Husic, Simon Olsson, and Frank Noé. Introduction to markov state modeling with the pyemma software [article v1.0]. *Living J. Comp. Mol. Sci.*, 1(1):5965, 2019.

[463] Brooke E Husic and Vijay S Pande. Markov state models: From an art to a science. *J. Am. Chem. Soc.*, 140(7):2386–2396, 2018.

[464] Marco Sarich, Frank Noé, and Christof Schütte. On the approximation quality of markov state models. *Multiscale Model Simul*, 8(4):1154–1177, 2010.

[465] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nat. Commun*, 9(1):1–11, 2018.

[466] David DL Minh and John D Chodera. Optimal estimators and asymptotic variances for nonequilibrium path-ensemble averages. *J. Chem. Phys.*, 131(13):134110, 2009.

[467] Daniel M Zuckerman and Thomas B Woolf. Dynamic reaction paths and rates through importance-sampled stochastic dynamics. *J. Chem. Phys.*, 111(21):9475–9484, 1999.

[468] Thomas B Woolf. Path corrected functionals of stochastic trajectories: towards relative free energy and reaction coordinate calculations. *Chem. Phys. Lett.*, 289(5-6):433–441, 1998.

[469] Daniel M Zuckerman and Thomas B Woolf. Efficient dynamic importance sampling of rare events in one dimension. *Phys. Rev. E*, 63(1):016702, 2000.

[470] Christof Schütte, Adam Nielsen, and Marcus Weber. Markov state models and molecular alchemy. *Mol. Phys.*, 113(1):69–78, 2015.

[471] WF Van Gunsteren and HJC Berendsen. Algorithms for brownian dynamics. *Mol. Phys.*, 45(3):637–647, 1982.

[472] Giovanni Bussi and Michele Parrinello. Accurate sampling using langevin dynamics. *Phys. Rev. E*, 75(5):056707, 2007.

[473] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. Langevin equation with colored noise for constant-temperature molecular dynamics simulations. *Phys. Rev. Lett.*, 102(2):020601, 2009.

[474] N Goga, AJ Rzepiela, AH De Vries, SJ Marrink, and HJC Berendsen. Efficient algorithms for langevin and dpd dynamics. *Journal of Chemical Theory and Computation*, 8(10):3637–3649, 2012.

[475] Benedict Leimkuhler and Charles Matthews. Robust and efficient configurational molecular sampling via langevin dynamics. *J. Chem. Phys.*, 138(17):05B601_1, 2013.

[476] Josh Fass, David A Sivak, Gavin E Crooks, Kyle A Beauchamp, Benedict Leimkuhler, and John D Chodera. Quantifying configuration-sampling error in langevin simulations of complex molecular systems. *Entropy*, 20(5):318, 2018.

[477] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(5):826–843, 2011.

[478] Jim Pfaendtner and Massimiliano Bonomi. Efficient sampling of high-dimensional free-energy landscapes with parallel bias metadynamics. *J. Chem. Theory Comput.*, 11(11):5062–5067, 2015.

[479] Arushi Prakash, Christopher D Fu, Massimiliano Bonomi, and Jim Pfaendtner. Biasing smarter, not harder, by partitioning collective variables into families in parallel bias metadynamics. *J. Chem. Theory Comput.*, 14(10):4985–4990, 2018.

[480] Alexey Onufriev, Donald Bashford, and David A Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct. Funct. Genet.*, 55(2):383–394, 2004.

[481] A. D. Jr. MacKerell, D. Bashford, M. Bellott, R. L. Jr. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[482] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.*, 100(9):L47–L49, 2011.

[483] Andrew L Ferguson. Bayeswham: A bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method. *J. Comput. Chem.*, 38(18):1583–1605, 2017.

[484] Jordane Preto and Cecilia Clementi. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.*, 16(36):19181–19191, 2014.

[485] Behrooz Hashemian and Marino Arroyo. Topological obstructions in the way of data-driven collective variables. *J. Chem. Phys.*, 142(4):044102, 2015.

[486] Peter G Bolhuis, Christoph Dellago, and David Chandler. Reaction coordinates of biomolecular isomerization. *Proc. Natl. Acad. Sci. U.S.A.*, 97(11):5877–5882, 2000.

[487] Imre Jákli, Svend J Knak Jensen, Imre G Csizmadia, and András Perczel. Variation of conformational properties at a glance. true graphical visualization of the ramachandran surface topology as a periodic potential energy surface. *Chem. Phys. Lett.*, 547:82–88, 2012.

[488] David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7(23):3910–3916, 2005.

[489] Ulrich HE Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281(1-3):140–150, 1997.

[490] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141–151, 1999.

[491] Yuji Sugita and Yuko Okamoto. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.*, 329(3-4):261–270, 2000.

[492] Hiroaki Fukunishi, Osamu Watanabe, and Shoji Takada. On the hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.*, 116(20):9058–9067, 2002.

[493] Giovanni Bussi. Hamiltonian replica exchange in gromacs: a flexible implementation. *Mol. Phys.*, 112(3-4):379–384, 2014.

[494] Grégoire A Gallet and Fabio Pietrucci. Structural cluster analysis of chemical reactions in solution. *J. Chem. Phys.*, 139(7):074101, 2013.

[495] Silvio Pipolo, Mathieu Salanne, Guillaume Ferlat, Stefan Klotz, A Marco Saitta, and Fabio Pietrucci. Navigating at will on the water phase diagram. *Phys. Rev. Lett.*, 119(24):245701, 2017.

[496] Joseph E Basconi and Michael R Shirts. Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations. *J. Chem. Theory Comput.*, 9(7):2887–2899, 2013.

[497] DJ Bicout and Attila Szabo. Electron transfer reaction dynamics in non-debye solvents. *J. Chem. Phys.*, 109(6):2325–2338, 1998.

[498] Yihang Wang and Pratyush Tiwary. Understanding the role of predictive time delay and biased propagator in rave. *J. Chem. Phys.*, 152(14):144102, 2020.

[499] Luca Donati, Marcus Weber, and Bettina G Keller. A review of girsanov reweighting and of square root approximation for building molecular markov state models. *J. Math. Phys.*, 63(12):123306, 2022.

[500] Frank Noé. Machine learning for molecular dynamics on long timescales. *arXiv preprint arXiv:1812.07669*, 2018.

[501] Cecilia Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Current opinion in structural biology*, 18(1):10–15, 2008.

[502] Marissa G Saunders and Gregory A Voth. Coarse-graining methods for computational biology. *Annual review of biophysics*, 42:73–93, 2013.

[503] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical reviews*, 116(14):7898–7936, 2016.

[504] Young C Kim and Gerhard Hummer. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *Journal of molecular biology*, 375(5):1416–1433, 2008.

[505] Kathryn A Scott, Peter J Bond, Anthony Ivetac, Alan P Chetwynd, Syma Khalid, and Mark SP Sansom. Coarse-grained md simulations of membrane protein-bilayer self-assembly. *Structure*, 16(4):621–630, 2008.

[506] Joshua Lequieu, Andrés Córdoba, Joshua Moller, and Juan J De Pablo. 1cpn: A coarse-grained multi-scale model of chromatin. *The Journal of Chemical Physics*, 150(21):215102, 2019.

[507] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, 1975.

[508] Paulo CT Souza, Riccardo Alessandri, Jonathan Barnoud, Sebastian Thallmair, Ignacio Faustino, Fabian Grünewald, Ilias Patmanidis, Haleh Abdizadeh, Bart MH Bruininks, Tsjerk A Wassenaar, et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature methods*, 18(4):382–388, 2021.

[509] Sangjae Seo and Wataru Shinoda. Spica force field for lipid membranes: domain formation induced by cholesterol. *Journal of chemical theory and computation*, 15(1):762–774, 2018.

[510] Rhiju Das and David Baker. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, 77:363–382, 2008.

[511] Wei Han, Cheuk-Kin Wan, Fan Jiang, and Yun-Dong Wu. Pace force field for protein simulations. 1. full parameterization of version 1 and verification. *Journal of chemical theory and computation*, 6(11):3373–3389, 2010.

[512] Andrzej Koliński et al. Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, 51, 2004.

[513] Aram Davtyan, Nicholas P Schafer, Weihua Zheng, Cecilia Clementi, Peter G Wolynes, and Garegin A Papoian. Awsem-md: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *The Journal of Physical Chemistry B*, 116(29):8494–8503, 2012.

[514] John M Jumper, Nabil F Faruk, Karl F Freed, and Tobin R Sosnick. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. *PLoS computational biology*, 14(12):e1006342, 2018.

[515] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deepcg: Constructing coarse-grained models via deep neural networks. *The Journal of chemical physics*, 149(3):034101, 2018.

[516] Shriram Chennakesavalu, David J Toomer, and Grant M Rotskoff. Ensuring thermodynamic consistency with invertible coarse-graining. *The Journal of Chemical Physics*, 158(12):124126, 2023.

[517] Xinqiang Ding and Bin Zhang. Contrastive learning of coarse-grained force fields. *Journal of Chemical Theory and Computation*, 18(10):6334–6344, 2022.

[518] Aleksander EP Durumeric, Nicholas E Charron, Clark Templeton, Félix Musil, Klara Bonneau, Aldo S Pasos-Trejo, Yaoyi Chen, Atharva Kelkar, Frank Noé, and Cecilia Clementi. Machine learned coarse-grained protein force-fields: Are we there yet? *Current Opinion in Structural Biology*, 79:102533, 2023.

[519] Marloes Arts, Victor Garcia Satorras, Chin-Wei Huang, Daniel Zuegner, Marco Federici, Cecilia Clementi, Frank Noé, Robert Pinsler, and Rianne van den Berg. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. *arXiv preprint arXiv:2302.00600*, 2023.

[520] Andreas Krämer, Aleksander P Durumeric, Nicholas E Charron, Yaoyi Chen, Cecilia Clementi, and Frank Noé. Statistically optimal force aggregation for coarse-graining molecular dynamics. *arXiv preprint arXiv:2302.07071*, 2023.

[521] Sergei Izvekov and Gregory A Voth. A multiscale coarse-graining method for biomolecular systems. *The Journal of Physical Chemistry B*, 109(7):2469–2473, 2005.

[522] William George Noid, Jhih-Wei Chu, Gary S Ayton, Vinod Krishna, Sergei Izvekov, Gregory A Voth, Avisek Das, and Hans C Andersen. The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models. *The Journal of chemical physics*, 128(24):244114, 2008.

[523] M Scott Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of chemical physics*, 129(14):144108, 2008.

[524] Aleksandra E Badaczewska-Dawid, Andrzej Kolinski, and Sebastian Kmiecik. Computational reconstruction of atomistic protein structures from coarse-grained models. *Computational and structural biotechnology journal*, 18:162–176, 2020.

[525] Allison P Heath, Lydia E Kavraki, and Cecilia Clementi. From coarse-grain to all-atom: toward multiscale analysis of protein landscapes. *Proteins: Structure, Function, and Bioinformatics*, 68(3):646–661, 2007.

[526] Leandro E Lombardi, Marcelo A Martí, and Luciana Capece. Cg2aa: backmapping protein coarse-grained structures. *Bioinformatics*, 32(8):1235–1237, 2016.

[527] Srinivasa M Gopal, Shayantani Mukherjee, Yi-Ming Cheng, and Michael Feig. Primo/primona: a coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins: Structure, Function, and Bioinformatics*, 78(5):1266–1281, 2010.

[528] Pilar Brocos, Paola Mendoza-Espinosa, Rolando Castillo, Jaime Mas-Oliva, and Ángel Pineiro. Multiscale molecular dynamics simulations of micelles: coarse-grain for self-assembly and atomic resolution for finer details. *Soft Matter*, 8(34):9005–9014, 2012.

[529] Matías R Machado and Sergio Pantano. Sirah tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics*, 32(10):1568–1570, 2016.

[530] Piotr Rotkiewicz and Jeffrey Skolnick. Fast procedure for reconstruction of full-atom protein models from reduced representations. *Journal of computational chemistry*, 29(9):1460–1465, 2008.

[531] David N Nicholson and Casey S Greene. Constructing knowledge graphs and their biomedical applications. *Computational and structural biotechnology journal*, 18:1414–1428, 2020.

[532] Soojung Yang and Rafael Gómez-Bombarelli. Chemically transferable generative backmapping of coarse-grained proteins. *arXiv preprint arXiv:2303.01569*, 2023.

[533] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[534] Tamas Lazar, Elizabeth Martínez-Pérez, Federica Quaglia, András Hatos, Lucía B Chemes, Javier A Iserte, Nicolás A Méndez, Nicolás A Garrone, Tadeo E Saldaño, Julia Marchetti, et al. Ped in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic acids research*, 49(D1):D404–D411, 2021.

[535] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

[536] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

[537] Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv preprint arXiv:2210.05274*, 2022.

[538] Kevin E Wu, Kevin K Yang, Rianne van den Berg, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *arXiv preprint arXiv:2209.15611*, 2022.

[539] Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

[540] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, pages 2022–12, 2022.

[541] Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F Miller III, and Anima Anandkumar. Dynamic-backbone protein-ligand structure prediction with multiscale generative diffusion models. *arXiv preprint arXiv:2209.15171*, 2022.

[542] John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, et al. Illuminating protein space with a programmable generative model. *bioRxiv*, pages 2022–12, 2022.

[543] Calvin Luo. Understanding ditba ffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[544] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[545] Yihang Wang, Lukas Herron, and Pratyush Tiwary. From data to noise to data for mixing physics across temperatures with generative artificial intelligence. *Proceedings of the National Academy of Sciences*, 119(32):e2203656119, 2022.

[546] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*, 2019.

[547] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[548] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[549] GH Golub and CF Van Loan. Special linear systems. *Matrix computations, 2nd ed. Baltimore: John Hopkins University Press*, 1989.

[550] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[551] Dan Nguyen, Chuanying Chen, B Montgomery Pettitt, and Junji Iwahara. Nmr methods for characterizing the basic side chains of proteins: electrostatic interactions, hydrogen bonds, and conformational dynamics. *Methods in enzymology*, 615:285–332, 2019.

[552] Alexandre Esadze, Da-Wei Li, Tianzhi Wang, Rafael Brüschweiler, and Junji Iwahara. Dynamics of lysine side-chain amino groups in a protein studied by heteronuclear 1h-15n nmr spectroscopy. *Journal of the American Chemical Society*, 133(4):909–919, 2011.

[553] Robin Winter, Frank Noé, and Djork-Arné Clevert. Permutation-invariant variational autoencoder for graph-level representation learning. *arXiv preprint arXiv:2104.09856*, 2021.

[554] Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *International Conference on Artificial Neural Networks*, pages 412–422. Springer, 2018.

[555] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.