

THE UNIVERSITY OF CHICAGO

SYMMETRY ADAPTED MACHINE LEARNING MOLECULAR DESCRIPTORS AND
ACTIVE LEARNING APPROACHES FOR SWITCHABLE SUPERLATTICES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY
NICHOLAS S.M. HERRINGER

CHICAGO, ILLINOIS
AUGUST 2025

ABSTRACT

This thesis presents advancements in computational methodologies to enhance molecular simulations through the development of active learning techniques, machine-learning methods, and robust open-source software engineering with a focus on ML-integration. The research is organized around four primary contributions: (1) computational modeling of self-assembling peptide- π conjugated systems to guide the design of supramolecular nanomaterials, (2) the development and implementation of Permutationally Invariant Networks for Enhanced Sampling (PINES), (3) the high-throughput virtual screening of DNA-functionalized nanoparticles (DFPs) using an active learning approach, and (4) the structured software development of the PINES and ALPineFOREst frameworks.

First, this thesis contributes to the design of tunable peptide- π conjugated nanomaterials by performing molecular dynamics simulations of perylene diimide (PDI)-peptide conjugates and quantifying their self-assembly behavior. My simulations of VEVAG and VEVA systems revealed distinct stacking geometries linked to subtle differences in amino acid composition. These computational results validated experimental findings and offered mechanistic insight into how sequence variation alters supramolecular organization and optoelectronic properties.

Second, we introduce PINES as a novel autoencoder-based iterative collective variable (CV) discovery method with on-the-fly biasing for enhanced sampling in molecular dynamics simulations. While many CV discovery methods have been developed, few are symmetry-adapted to address permutation of molecular labels. By leveraging the Permutation Invariant Vector (PIV) as a symmetry-adapted molecular featurization, the CVs discovered by PINES are inherently invariant under translation, rotation, and permutation and utilize a data-driven process to describe the highest variation dynamics of the system. The effectiveness of PINES is validated on three simple molecular systems, each demonstrating an advantage gained from using PINES over other existing methods. We first apply PINES to an LJ fluid of

13 Argon atoms and show its ability to approximate global observables like potential energy while remaining general purpose and data-driven, unlike other symmetry adapted methods such as SOAP. We then use PINES to investigate Na^+ and Cl^- pairing in explicit water, a benchmark system which has been heavily studied, where we confirm that our calculated free energy surfaces (FES) are in strong agreement with literature and we highlight the additional sampling of high energy solvent states furnished by PINES CVs relative to traditional CVs which exclude water explicitly. Finally, we test PINES on a solvated alkane chain and find that PINES recovers FES in great agreement with a previous computational study using handcrafted molecular descriptors to probe the wetting/dewetting transitions of the alkane [1] as well as that PINES uniquely identifies a weaker CV associated with the fraction of helical pitch in the alkane chain conformation.

Third, this thesis investigates DNA-functionalized nanoparticles (DFPs) as tunable materials exhibiting dynamic hybridization behaviors. We design a novel active learning protocol utilizing multi-state and multi-thermal free-energy calculations from coarse-grained molecular dynamics simulations to evaluate and score each simulated system with a proxy for “switchability.” We generate a labeled dataset of system features and corresponding switchability scores to train Gaussian Process surrogate models with bespoke kernel functions, then we employ Thompson sampling to select the next round of molecular candidates using Bayesian optimization. We demonstrate how even vast molecular design spaces can be traversed efficiently using active learning approaches to high throughput virtual screening.

Finally, this thesis emphasizes robust software development practices through the creation and deployment of modular, extensible frameworks: the C++ PINES method for symmetry-adapted, iterative CV discovery and the ALPineFOREst Python library for streamlined active learning cycles. Attention to software architecture, maintainability, and open-source accessibility ensures these tools are available and adaptable for broader scientific applications. This attention towards utility for the scientific community has already resulted in the use of

PINES in several ongoing research projects within our group, investigating topics as disparate as silicon-carbide vacancies and water structure in relation to the Mpemba effect, as well as the expected immediate integration of ALPineFOREst into the Center for the Synthesis of Science Across Scales (CSSAS) Energy Frontier Research Center (EFRC) workflow for DFP and other active-learning related applications.

Together, these contributions exemplify the integration of computational chemistry, machine learning, and research software engineering to drive innovation in molecular simulations, facilitating more efficient and insightful exploration of complex chemical systems as well as development of user-focused, community-based scientific software.

To Mom & Dad, for showing me the way.

To Greg, for caring about my work.

To Chris, for always believing in me.

To Eric, Matthias, and Andy for their kindness and mentorship.

Most of all, to my fiancée Tegan, for standing by my side on this journey and making my
days worth living. Our next adventure is yours, my love.

"But I am very poorly today & very stupid & I hate everybody & everything. One lives only to make blunders."

- *Charles Darwin*

TABLE OF CONTENTS

LIST OF FIGURES	x
ACKNOWLEDGMENTS	xviii
1 INTRODUCTION	1
1.1 Computationally Guided Tuning of Sequence-Dependent Twist Angles in Peptide- π Chromophore Stacking	2
1.2 Iterative Discovery and Refinement of Symmetry-Adapted Collective Variables	3
1.3 Applied Active Learning for Switchable Superlattices	4
1.4 Shifting from Application to Tool-Making	6
2 COMPUTATIONALLY GUIDED TUNING OF PEPTIDE-CONJUGATED PERYLENE DIIMIDE SELF-ASSEMBLY	9
2.1 Abstract	9
2.2 Introduction	10
2.3 Methods	13
2.3.1 Experimental Methods	13
2.3.2 Computational Methods	16
2.4 Results and Discussion	21
2.5 Conclusion	37
2.6 Acknowledgements	38
3 PERMUTATIONALLY INVARIANT NETWORKS FOR ENHANCED SAMPLING (PINES): DISCOVERY OF MULTI-MOLECULAR AND SOLVENT-INCLUSIVE COLLECTIVE VARIABLES	39
3.1 Abstract	39
3.2 Introduction	40
3.3 Methods	47
3.3.1 Molecular Enhanced Sampling with Autoencoders (MESA)	47
3.3.2 Permutation Invariant Vector (PIV) Featurization	50
3.3.3 Autoencoder network design and training	54
3.3.4 Enhanced sampling along discovered CVs using parallel bias metadynamics	56
3.3.5 Convergence assessment	57
3.3.6 All-atom molecular dynamics simulations	58
3.3.7 13-particle Ar cluster	59
3.3.8 NaCl ion pair in water	60
3.3.9 Aqueously solvated n -pentatetracontane ($C_{45}H_{92}$) polymer chain	61
3.4 Results and Discussion	62
3.4.1 Self-assembly of a 13-particle Ar cluster	63
3.4.2 Association/dissociation of a NaCl ion pair in water	70

3.4.3	Hydrophobic collapse of an n -pentatetracontane ($C_{45}H_{92}$) polymer chain in water	78
3.5	Conclusions	85
3.6	Acknowledgements	87
3.7	Appendix: Supporting Information	87
4	ACTIVE LEARNING FOR OPTIMAL THERMORESPONSIVE DNA-BASED SWITCH- ABLE SUPERLATTICES	93
4.1	Abstract	93
4.2	Introduction	94
4.3	Methods	97
4.3.1	Thermoresponsive DFP Mechanism	97
4.3.2	CG-DNA Simulations	100
4.3.3	Evaluating System Fitness	101
4.3.4	Surrogate Modeling with Bespoke Molecular Similarity Kernel	105
4.3.5	Bayesian Optimization	106
4.4	Results and Discussion	109
4.5	Conclusions	111
4.6	Acknowledgements	112
5	FROM PINES TO ALPINEFOREST: SEEING THE FOREST FOR THE TREES	113
5.1	Reduce and Reuse, Don't Recycle	115
5.1.1	Respect the Environment	115
5.1.2	Quick and Dirty Solution Pollution	116
5.2	PINES	116
5.3	ALPineFOREst	118
5.4	Conclusions	119
6	CONCLUSION	121
6.1	Contributions to the field	121
6.2	Directions for the Future	124
6.2.1	Applications for PINES and Further Developments	124
6.2.2	Cultivating the Forest	127
	REFERENCES	130

LIST OF FIGURES

2.1	Chemical structures of the DAVG-PDI-GVAD and DAIA-PDI-AIAD peptides inspired by computational input.	22
2.2	Transmission electron microscopy (TEM) images of assembled peptides (a, e), UV-Vis (b, f), PL (c, g) and CD (d,h) spectroscopy traces of DAVG (a-d) and DAIA (e-h) peptides in basic (solid pink), neutral (blue short dash), and acidic (green long dash) media. UV-Vis and PL spectra were recorded at 5-7.5 μ M concentration range. CD spectra were recorded at 15-20 μ M concentration range.	24
2.3	Chemical structures of the members of the VEVX1X2-PDI-X2X1VEV peptide libraries.	27
2.4	TEM images of a) VEVAG, and b) VEVA A from acidic solution of concentration 0.2 mg/mL.	28
2.5	UV-Vis signatures of a) VEVAG under basic, neutral, and acidic conditions, b) VEVX1G peptides under acidic conditions, c) VEVA A under basic, neutral, and acidic conditions, and d) VEVX1A peptides under acidic conditions. All the spectra were recorded at 5-7.5 μ M concentration range.	30
2.6	PL signature of a) VEVAG under basic, neutral, and acidic condition. b) VEVA A under basic, neutral, and acidic condition. All the spectra were recorded at 5-7.5 μ M concentration range.	31
2.7	CD signature of a) VEVAG under basic, neutral, and acidic condition. b) VEVA A under basic, neutral, and acidic condition. All the spectra were recorded at 15-20 μ M concentration range.	32
2.8	Potential of mean force (PMF) as a function of signed twist angle (θ) recovered from well-tempered metadynamics simulations of (A) DAVG, (B) DAIA, (C) VEVAG, and (D) VEVA A. Metadynamics simulations consider a dimerized pair of molecules biased along the collective variable defined as the signed twist angle (θ). Annotations indicate the difference in free energy ΔF between the local free energy minima in the $\theta > 0$ and $\theta < 0$ regimes. Standard errors are illustrated by shading and were estimated by block averaging our trajectories into five equally sized contiguous blocks.	34
2.9	Cumulative distribution function (CDF) for the probability of observing a dimer pair at or below a given signed twist angle magnitude $ \theta $ for (A) DAVG vs. DAIA and (B) VEVAG vs. VEVA A. Shading represents standard errors estimated by block averaging where each trajectory is split into five equally sized contiguous blocks. The placement of the smaller Gly residue in the X_2 position immediately adjacent to the PDI core leads to a higher observed probability of smaller magnitude twist angles, whereas its replacement with a bulkier Ala residue leads to a higher prevalence of larger magnitude twist angles.	35

2.10	Potential of mean force (PMF) as a function of dimer aromatic core COM separation for (A) DAVG and DAIA, and (B) VEVAG and VEVA extracted from umbrella sampling simulations. Shading represents pointwise standard errors estimated from three independently run sets of umbrella sampling simulations for each system. The arbitrary zero of free energy is defined for the non-interacting system at large COM separation, $ \mathbf{d} $. Annotations indicate the difference in the dimerization free energies between the two peptide systems $\Delta\Delta\beta F$. The presence of a proximate Gly residue in the X_2 position immediately adjacent to the PDI core results in a more favorable free energy of dimerization than the presence of an Ala residue in the same position.	36
3.1	Schematic illustration of the Permutationally Invariant Network for Enhanced Sampling (PINES) approach for data-driven CV discovery and enhanced sampling. (a) Molecular simulation trajectories provide training data for CV discovery. The process is initialized with unbiased or biased simulations to provide an initial first sampling of phase space. Subsequent rounds are trained over biased simulations generated by enhanced sampling in the learned CVs from the previous round. (b) The trajectories are featurized and passed to a maximum mean discrepancy-Wasserstein autoencoder (MMD-WAE) to perform unsupervised nonlinear dimensionality reduction into a low-dimensional latent space spanned by the learned CVs. To stabilize the CV discovery process and help disentangle the latent space, we employ batch normalization between all layers with exception of the final layer. (c) After training the network, we extract the CVs from the bottleneck layer between the encoder and decoder. We find that it is useful to first pass the output of the bottleneck layer through a principal components analysis (PCA) whitening transformation to help align the discovered manifold with the highest variance directions for efficient parallel bias metadynamics (PBMetaD) sampling. The learned CVs are represented as the functional composition $\xi = h \circ f(\mathbf{r}) = h(\mathbf{z})$, where $\mathbf{z} = f(\mathbf{r})$ is the result of applying the featurization f – in this work a PIV featurization – to the Cartesian coordinates $\mathbf{r}$ of the system, and $\xi = h(\mathbf{z})$ is the action of the MMD-WAE encoder plus PCA whitening transform upon the input featurization that generates a projection into the learned CVs. The derivative of the CVs with respect to the Cartesian coordinates of atom i is required for propagation of biasing forces in the CVs to biasing forces on the atoms, and comes from simple application of the chain rule to the function composition that is computed in practice by automatic differentiation, $\nabla_{\mathbf{r}_i}\xi = (\partial h/\partial \mathbf{z})\nabla_{\mathbf{r}_i}\mathbf{z}$. (d) The learned CVs are implemented within PBMetaD calculations to drive sampling in the configurational phase space spanned by these CVs, completing one full iteration of PINES. The biased simulation trajectories from Round n are then passed forward as the training trajectories for CV discovery within Round $(n + 1)$ to determine new CVs and conduct new enhanced sampling calculations. The iterative process is terminated when the learned CVs stabilize, we observe no further exploration of phase space, and the reweighted free energy landscape converges.	49

- 3.2 Schematic illustration of calculation of the PIV featurization of a *NaCl* ion pair in water. (a) Snapshot of an instantaneous state of the *NaCl* ion pair and the two most proximate water molecules. All molecular renderings in this work are constructed using Visual Molecular Dynamics (VMD) [2]. (b) We convert the Cartesian coordinates of the atomic system into a distance matrix recording all pairwise distances a_{ij} between all pairs of atoms i and j . The symmetry of the distance matrix means we need only record the upper triangle and is arranged into a block structure by atom class such that indistinguishable atoms are grouped together. Since the number of pairwise distances increases quadratically with the number of atoms, we may also choose to restrict the atoms over which we construct the distance matrix by, for example, omitting all but the alpha-carbons in a protein or disregarding the far-field solvent molecules around a solute. In the interests of space, we present here the abridged distance matrix where we have colored each block or group of (indistinguishable) pairwise distances: Na-Cl (green), Na-O (red), Na-H (white), Cl-O (orange), and Cl-H (blue). (c) The elements of the distance matrix a_{ij} are transformed by a rational switching function to produce the adjacency matrix with elements v_{ij} . The parameters of the switching function are tuned to present the largest dynamic range and sensitivity to configurational changes in the short- to medium-ranged field. It also has the desirable property of scaling the pairwise distances to a $(0, 1)$ range. We choose r_0 in each block (r_0^{block}) to be the mean of the first and second peaks of the RDF, set $d_0 = 0$ and $m = 2n$, and tune n such that $v_{ij} \approx 0.9$ at the a_{ij} value of the first peak and $v_{ij} \approx 0.1$ at the a_{ij} value of the second peak in the RDF. (d) The elements within each block of the adjacency matrix are placed in non-descending order and concatenated to form PIV. 53
- 3.3 Application of PINES to a 13-particle Ar system in vacuum at 50 K. (a) Each column shows a 2D embedding of the round-by-round accrued sampling projected into the converged CVs ξ_0 , ξ_1 , and ξ_2 at the terminal Round 5. The unbiased simulation quite thoroughly samples the thermally accessible phase space, yet each biased round advances the frontier into unsampled regions on the periphery of the explored region. (b) Corresponding free energy surfaces (FES) computed by reweighting of the biased simulation data collected in the current round of sampling. (c) Cosine similarity between corresponding CVs in successive rounds. The high cosine similarity value indicates rapid convergence of the iterative CV discovery and enhanced sampling cycles within the five rounds of PINES. 65

3.4	Bottom-up and top-down analysis of terminal learned CVs for the 13-particle Ar system. (a) Split plots show the signed contribution from each PIV element to each CV $\{\xi_0, \xi_1, \xi_2\}$ as a bar chart on top and as a cumulative sum of the signed contributions on the bottom. The magnitude of contributions on the bar chart emphasize the importance of individual PIV elements while the slopes in different regions of the scatter plot illuminate the concerted effect of atom groups on each CV. (b-e) Projections of the terminal Round 5 sampling data into the terminal CVs and colored by physically interpretable candidate variables. The potential energy, which is known a priori to be the putative CV for a Lennard-Jones fluid, correlates well with ξ_0 . Selected snapshots are presented to show the smooth transition across ξ_0 from the vapor-like phase (red, orange) to the condensed phase (green), and finally to the highly ordered solid-like phase (blue). The gradients in ξ_0 and ξ_1 with respect to the average and maximum pairwise distances indicates that PINES has learned these as important variables in driving assembly/disassembly.	67
3.5	Reweight free energy surfaces from the terminal round of enhanced sampling for the 13-particle Ar system projected into physically interpretable order parameters. (a) A 1D FES projected into the average pairwise interatomic distance. Representative snapshots show that the global free energy minimum to be occupied by a solid-like, filled icosahedron with an average pairwise distance of 0.62 nm. (b) A 2D FES projected into the potential energy and maximum pairwise interatomic distance. Representative snapshots illustrate the solid-like global free energy minimum and liquid-like and vapor-like phases.	69
3.6	Application of PINES to a NaCl ion pair in water at 300 K. (a) Each column shows a 2D embedding of the round-by-round accrued sampling projected into the converged CVs ξ_0 , ξ_1 , and ξ_2 at the terminal Round 5. The unbiased simulation quite thoroughly samples the thermally accessible phase space, yet each biased round advances the frontier into unsampled regions on the periphery of the explored region. Rounds 1-3 show broader exploration of the CV space in half or less of the simulation time compared to the unbiased run and indicate that the CVs are accelerating sampling of novel states. Rounds 4-5, despite being much longer than the previous biased runs, provide diminishing returns in the exploration of new areas of CV space. (b) Corresponding free energy surfaces (FES) computed by reweighting of the biased simulation data collected in the current round of sampling. (c) Cosine similarity between corresponding CVs in successive rounds. The high cosine similarity values between all neighboring rounds indicate that the CVs discovered in the initial unbiased round are not substantially altered by learning over the biased data and the PINES protocol rapidly converges. . . .	72

3.7	Bottom-up and top-down analysis of terminal learned CVs for the NaCl ion pair in water. (a) Split plots show the signed contribution from each PIV element to each CV $\{\xi_0, \xi_1, \xi_2\}$ as a bar chart on top and as a cumulative sum of the signed contributions on the bottom. The magnitude of contributions on the bar chart emphasize the importance of individual PIV elements while the slopes in different regions of the scatter plot illuminate the concerted effect of atom groups on each CV. (b-f) Projections of the terminal Round 5 sampling data into the terminal CVs and colored by physically interpretable candidate variables. ξ_0 is well-correlated with the average Cl-O and Cl-H distance, ξ_1 with the average Na-O and Na-H distance, and ξ_2 with the Na-Cl distance. Three distinct color bands are apparent in panel (f) corresponding to the contact pair (blue), the solvent-shared (green), and disassociated (red) states (cf. Fig. 3.8c).	75
3.8	Rewighted free energy surfaces from the terminal round of enhanced sampling for the NaCl ion pair in water projected into physically interpretable order parameters. (a) A 1D FES projected into the Na-Cl distance. The landscape clearly illustrates the contact pair ($r_{\text{Na-Cl}} = 0.26$ nm), solvent-shared ($r_{\text{Na-Cl}} = 0.51$ nm), and disassociated ($r_{\text{Na-Cl}} \gtrsim 0.6$ nm) states. The free energy barrier lying between the contact pair and solvent-shared states is $(4.9 \pm 0.1) k_B T$ as measured from the contact pair, and involves a collective rearrangement of solvent shell water molecules to accommodate the separating ion pair [3, 4]. Uncertainties in the free energy correspond to standard errors estimated by five-fold block averaging. (b) A 2D FES projected into the Na-Cl distance and the solvent coordination number of the Na ion n_{Na} . The contact pair comprises multiple solvation states of the Na ion. Transitions to the solvent separated pair appear to proceed most readily through the $n_{\text{Na}} = 5$ contact pair state.	77
3.9	Application of PINES to an <i>n</i> -pentatetracontane ($C_{45}H_{92}$) chain in water at 300 K. (a) Each column shows a 2D embedding of the round-by-round accrued sampling projected into the converged CVs ξ_0 , ξ_1 , and ξ_2 at the terminal Round 7. Rounds 1-4 show significantly increased exploration of novel states compared to the unbiased run and indicate that the CVs are accelerating sampling of unexplored regions of phase space. In Rounds 5-6 the sampling of the CV space begins to saturate and suggests that the CVs have converged. (b) Corresponding free energy surfaces (FES) computed by reweighting of the biased simulation data collected in the current round of sampling. (c) Cosine similarity between corresponding CVs in successive rounds. The CVs change significantly round-to-round over Rounds 1-4 before converging in Rounds 5-7.	80

- 3.10 Bottom-up and top-down analysis of terminal learned CVs for the n -pentatetracontane chain in water. (a) Split plots show the signed contribution from each PIV element to each CV $\{\xi_0, \xi_1, \xi_2\}$ as a bar chart on top and as a cumulative sum of the signed contributions on the bottom. The magnitude of contributions on the bar chart emphasize the importance of individual PIV elements while the slopes in different regions of the scatter plot illuminate the concerted effect of atom groups on each CV. (b-d) Projections of the terminal Round 7 sampling data into the terminal CVs and colored by physically interpretable candidate variables. ξ_0 and ξ_1 are correlated with the chain radius of gyration, R_g , and the number of waters, N , in the first solvation shell of the chain. We highlight representative configurations on the ξ_0 - ξ_1 surface corresponding to the extended (red), collapsed (blue), and partially collapsed (green) structures. ξ_2 is correlated with the helical pitch f_{helix} . We highlight representative configurations on the ξ_1 - ξ_2 surface corresponding to planar rings (blue) and corkscrews (red). 82
- 3.11 Reweighted free energy surfaces from the terminal round of enhanced sampling for the n -pentatetracontane chain in water. (a) A 1D FES projected into the number of waters N in the first solvation shell of the chain. A collapsed state with $N \approx 130$ is more stable than the fully extended state with $N \approx 200$ by approximately $4 k_B T$. Uncertainties in the free energy correspond to standard errors estimated by five-fold block averaging. (b) A 2D FES projected into the number of waters N in the first solvation shell of the chain and the radius of gyration R_g . The global free energy minimum at $(N \approx 130, R_g \approx 0.57 \text{ nm})$ contains the hydrophobically collapsed chain. 84
- 3.12 Switching function parameterization process from pairwise RDFs. (a) The Na-Cl pair correlation function is used to determine the r_0 parameter of the switching function, which is chosen to be the midpoint between the first two peaks of $g(r)$. This process naturally embeds transition regions close to r_0 on the switching function curve. (b) The second parameter, n , controls the steepness of the switching function curves with larger values producing steeper curves. n is chosen as the lowest positive integer which embeds the first $g(r)$ peak at a value greater than or equal to 0.9 and embeds the second $g(r)$ peak at a value less than or equal to 0.1. Four distinct system states are labeled along the Na-Cl pair correlation function and Na-Cl switching function, showing the conversion of pairwise distances to PIV values and the separation of states along the (0,1) range. 88
- 3.13 Example of the fraction of variance explained (FVE) plots used to determine the appropriate dimensionality of the latent space. In line with prior work, we identify a knee in the FVE plot that indicates the optimal number of bottleneck nodes. This example shows the intrinsic dimensionality identification for the network trained on the aggregated data set and is found to be $d = 3$ 89

3.14	Analysis of convergence and converged FES for 13 Argon atoms in an entropically dominated $5 \times 5 \times 5nm^3$ box. (a) The probability distribution, and (b) log probability distribution, for the $5 \times 5 \times 5nm^3$ box show that the gas-like state is highly favored compared to the liquid-like and solid-like phases. Volume effects are particularly impactful on finite size systems and the increase in the box size represents a shift to a system dominated by entropy rather than enthalpy. (c) The deposited gaussian height along each CV as a function of time. It can be seen that all three CVs are sufficiently converged as evidenced by the severe drop in the deposition height by the end of the trajectory. (d,e) The 1D (d) and 2D (e) FES show a global energy minimum for the gas-like state and a metastable liquid-like state.	90
3.15	Analysis of transferability of learned CVs for the 13 atom Argon system to out-of-training temperature regimes. (a,b) The respective 1D and 2D FES for the system at 25K. The solid-like phase is strongly favored and the learned CVs continue to effectively facilitate psuedo-phase transtions and exploration of the liquid-like and gas-like phases. (c,d) The respective 1D and 2D FES for the system at 100K. The gas-like phase is strongly favored and the learned CVs continue to effectively facilitate psuedo-phase transtions and exploration of the liquid-like and solid-like phases.	91
3.16	Analysis of transferability of learned CVs for the NaCl system to out-of-training charge regimes as divalent ions. (a,b) The respective 1D and 2D FES for a 25 ns unbiased simulation of the divalent system. The increased charges on the ions prohibit exploration of configurations outside of the contact pair. (c,d) The respective 1D and 2D FES for a 25 ns PBMetaD simulation of the divalent system along the PINES-identified CVs. The transition barrier separating the contact pair and solvent-shared pair states is $\approx 30x$ larger compared to the monovalent system, however acceleration along the PINES CVs is sufficient to traverse the transition barrier. (e,f) The respective 1D and 2D FES for a 25 ns MetaD simulation along the ion distance of the divalent system. While an improvement on the exploration of the unbiased simulation, the ion distance is not a sufficient CV to efficiently traverse the large transition barrier.	92
4.1	The proposed novel switching mechanism for DFP systems. In (a) we show a dimer representation of the DFPs and note the opposing entropic/enthalpic trends for transitioning between two states. In (b) we represent the coarsening of the DFPs with semi-opaque circles corresponding to the bi-length functionalized DNA, which we use to depict the process of (c) DFP lattice expansion and (d) compression.	99
4.2	We show a shortcoming of the forcefield to account for directionality in DNA binding and the resulting parallel (a) and anti-parallel (b) hybridization modes. In (c) we show that unphysical parallel binding is only prevalent at short DFP separations and that we only use DFP separations with $>80\%$ allowed hybridization.	102

4.3	In (a) we demonstrate leveraging system thermodynamics to modulate the DFP FES as a function of temperature and our corresponding proxy evaluation of $\Delta\Delta G$ for switchability. In (b) we provide an example from our active learning campaign for this process.	104
4.4	The KPC similarity space shows strong organization into meaningfully local regions from system features, and predicted $\Delta\Delta G$ values illuminate the impact of the features on the system's switchability for the initial dataset (a) and the first round of active learning (b).	107
4.5	ALPineFOREst as a framework for connecting active learning python libraries for MD campaigns.	108
4.6	Violin plots by round of simulated candidates. The campaign is ongoing and will include further rounds, currently shown as whitespace without violin plots. Final results from this work are expected to be published in August 2025	110

ACKNOWLEDGMENTS

This work, and the long journey behind it, wouldn't have been possible without the support of so many people.

First and most importantly, thank you to my fiancée, Tegan. She's been by my side through all six years of this PhD—through COVID, through the daily grind of research, through all the highs and lows. She brought joy and meaning to a process that can often feel slow and lonely, and she gave up things she wanted to do so I could keep going. I'm incredibly lucky, and I can't wait to see where her next chapter takes us.

To my parents—thank you for teaching me to love learning. As professors yourselves, you showed me what it looked like to dedicate your life to asking questions and sharing knowledge. You made this path feel possible. You each played formative roles in my interest in chemistry through the science chats Dad and I had on neighborhood walks and examples of chemistry in cooking like the Leidenfrost effect that Mom pointed out. And during grad school, your empathy, born of experience, helped more than you probably know.

To my brother Greg, our weekly phone calls were one of the most grounding things in my life. You always made the effort to stay connected, and even read through my first paper with me over the course of several months just to understand what I was working on. That meant the world to me. In the final push, when things got especially chaotic, your support really helped carry me through.

To my oldest brother Chris—thank you for believing in me, even when I didn't.

To the mentors who shaped my path: Siva Dasetty, thank you for helping me grow so much as a researcher early on in grad school, your friendship and support meant a great deal to me. Eric Wassinger, thank you for forever igniting my enthusiasm for chemistry research and for encouraging me to follow wherever my curiosity led me. Matthias Heyden, thank you for building the foundation of both my computational research skills and my confidence in myself as a researcher. And of course to my PI Andy Ferguson, thank you for your

incredible support over the entirety of my PhD. Your expertise and kindness are renowned for good reason, and I'm deeply grateful for the empathy you showed me during my struggles with my mental health. I feel truly lucky to have worked with you and for the researcher that you have helped me to become.

To the collaborators and colleagues I was lucky to work with: Max Gruschka, thank you so much for your dedication to getting me across the finish line in these last few months. It's been a pleasure to work with you and I know that you have a bright academic journey ahead of yourself. Dan McKeen, I'm thankful for our weekly meetings, they were often the highlight of my week. Our science collaboration was great, but I think we both know that trading book recommendations was truly invaluable. Diya Gandhi and Junhee Lee, you two were such fantastic undergrads to work with on PINES and our time together was truly one of the highlights of my time here.

CHAPTER 1

INTRODUCTION

Molecular simulations serve as a cornerstone of modern physical chemistry, biophysics, and materials science, offering an unparalleled window into the structure and dynamics of systems at atomic resolution. From elucidating the thermodynamics of protein folding to predicting the behavior of novel materials under extreme conditions, simulation methods enable both mechanistic insight and predictive modeling in regimes that are often inaccessible to experimental techniques. Their flexibility and generality have led to widespread adoption across disciplines, solidifying their role as indispensable tools in scientific discovery.

Despite these strengths, molecular simulations face persistent challenges. Chief among them is the difficulty of interrogating the vast conformational and compositional spaces that govern system behavior. The computational cost of reaching relevant time and length scales grows rapidly with system size and complexity, and brute-force approaches to exhaustive sampling are rarely feasible. This limitation motivates a growing interest in developing methods that maximize the scientific return from limited computational resources. In particular, machine learning techniques offer a compelling avenue for identifying hidden patterns in simulation data, constructing reduced representations of system dynamics, and guiding the selection of informative new simulations.

This dissertation explores a suite of machine learning approaches developed to accelerate molecular discovery and enhance sampling in simulation-based workflows. Chapter 2 presents a study of peptide- π -conjugated perylene diimide (PDI) systems in which I contributed all-atom molecular simulations of the VEVAG and VEVA sequences. These simulations guided experimental interpretation by revealing the structural determinants of supramolecular assembly and photophysical behavior as a function of amino acid sequence. Chapter 3 focuses on methods for learning collective variables (CVs) that capture the essential degrees of freedom governing a system’s evolution, with an emphasis on discovering symmetry-adapted

representations that respect permutational invariances. Chapter 4 turns to the challenge of efficiently navigating molecular design spaces using active learning and Bayesian optimization, enabling the targeted discovery of high-performing systems. Chapter 5 reflects on the broader computational ecosystem, highlighting the role of modular software development in enabling reproducibility, extensibility, and integration across diverse machine learning pipelines. Together, these chapters contribute to a more efficient, principled, and scalable approach to scientific modeling in molecular simulation.

1.1 Computationally Guided Tuning of Sequence-Dependent Twist Angles in Peptide- π Chromophore Stacking

In this collaboration on a study of perylene diimide (PDI)–peptide conjugates, which self-assemble into supramolecular architectures with tunable optoelectronic properties[5, 6]. I performed and analyzed all-atom molecular dynamics simulations of self-assembling synthetic π -conjugated peptides using GROMACS[7] and accelerated with well-tempered metadynamics from PLUMED[8]. These simulations provided essential mechanistic insight into how subtle variations in peptide sequence near the π -conjugated core influence molecular packing geometry and free energy landscapes of dimerization.

This project exemplifies how computational modeling can complement experimental characterization, enabling predictive design of functional materials. Through metadynamics and umbrella sampling, I quantified the relative stacking geometries, twist angles, and dimerization free energies associated with π -conjugated peptide dimer association. These results helped validate the hypothesis that the identity of residues adjacent to the PDI chromophore governs assembly mode, either promoting tight, cofacial packing or favoring more loosely associated aggregates with distinct optical signatures.

This work introduces two unifying themes in my PhD: molecular simulation and enhanced sampling. Simultaneously this work provides a good example of how simple nuances in a

staggeringly vast chemical space can affect unexpected changes to system structures and properties. In this work we simply validate experimental results from our collaborators and seek to elucidate design rules for oligopeptide systems, however following chapters explore the final theme of my dissertation: machine learning approaches to leverage the blessings of molecular richness in the face of the curse of dimensionality.

1.2 Iterative Discovery and Refinement of Symmetry-Adapted Collective Variables

The efficiency of molecular simulations is often determined both by the quality of the force field or the integration algorithm and by the ability to focus sampling on the most relevant degrees of freedom. Enhanced sampling methods aim to overcome kinetic bottlenecks by biasing the system along a small number of carefully chosen collective variables (CVs), which ideally capture the essential coordinates that drive rare events or transitions of interest. However, identifying such CVs is a longstanding challenge which is complicated by the “chicken-and-egg” nature of the problem: good sampling requires good CVs, but discovering good CVs requires good sampling.

This circular dependency has led to the development of various iterative methods [9, 10, 11, 12] that alternate between simulation and CV discovery. Early approaches used trial-and-error or human intuition to select likely order parameters, often relying on structural similarity, chemical intuition, or known reaction coordinates. More recent methods have introduced principled approaches to CV discovery using techniques from dimensionality reduction, information theory, or statistical mechanics. However, many of these approaches suffer from significant limitations: some produce CVs that are not symmetry-adapted and thus fail to generalize across equivalent configurations; others generate CVs that, while mathematically optimal, are difficult to interpret or rationalize in terms of classical chemistry or physics. Conversely, some CV discovery efforts err in the opposite direction by overcon-

straining the space of CVs to maintain interpretability, thereby excluding non-intuitive but physically meaningful directions of motion.

Symmetry-adapted CV discovery methods attempt to bridge this gap by learning representations that are invariant under relevant physical symmetries, such as permutation, translation, or rotation. This invariance not only improves generalization and sampling efficiency but also aligns the learned CVs with physically meaningful descriptors. Nevertheless, even symmetry-adapted methods are not immune to the chicken-and-egg dilemma. Many require an initial simulation that explores enough of configuration space to provide meaningful training data, which is rarely guaranteed in systems with large energy barriers or complex landscapes.

Iterative refinement frameworks aim to resolve this tension by bootstrapping between biased simulation and updated CV discovery. A biased simulation, even if performed with a suboptimal set of initial CVs, can provide better coverage of relevant configurations, which can in turn be used to learn improved CVs for the next round of sampling. Over successive iterations, this cycle can converge to a set of CVs that capture the dominant slow modes or high variance directions and enables efficient sampling of the system’s free energy landscape.

In this work I demonstrate a novel data-driven iterative approach which extends high-quality CV-based enhanced sampling to previously inaccessible molecular systems. I highlight the benefit of discovering symmetry-adapted CVs and emphasize how they not only serve as robust, transferable CVs but extend the ability to thoroughly describe the dynamics of solvated molecular structures.

1.3 Applied Active Learning for Switchable Superlattices

The growing demand for smart, programmable materials has placed renewed emphasis on the development of responsive materials, systems whose physical properties can be tuned or reconfigured in response to external stimuli such as temperature, pH, light, or chemical

environment. These materials offer the potential for bespoke functionality across a range of applications, from medical applications [13, 14] and molecular sensing [15] to reconfigurable electronics [16] and optical metamaterials [16, 17]. Central to this vision is the ability to precisely control structural and thermodynamic behavior at the nanoscale, enabling the targeted design of materials with properties that are not only tailored to specific applications but are also dynamically tunable.

DNA-functionalized particles (DFPs) represent a particularly promising platform for building such responsive materials. By grafting single-stranded DNA to nanoparticle surfaces, one can program the interactions between particles via Watson-Crick base pairing, controlling not only the identity of binding partners but also the strength and reversibility of the interactions. The resulting systems can self-assemble into ordered superlattices with customizable lattice constants, symmetries, and packing motifs [16, 17, 18, 19]. This makes DFPs a compelling system for the rational design of tunable crystalline materials, where the emergent properties of the superlattice, such as photonic bandgaps or mechanical stiffness, are sensitive to microscopic design choices like DNA sequence, strand density, or spacer length.

However, the design space of DFP systems is vast. Combinatorial variations in DNA sequence, strand length, grafting density, particle size, and ionic environment create a large combinatorial candidate space, only a small fraction of which are likely to yield desirable material properties. Exhaustive exploration of this landscape via brute-force simulation or experimental synthesis is computationally and logistically infeasible. Thus, a major challenge in materials design is not only to discover systems with high performance but to do so efficiently, minimizing the number of systems that must be evaluated while still capturing the most promising candidates, in other words a classic instance of the “needle in a haystack” problem.

To address this challenge, this dissertation explores the use of active learning as a prin-

cipled strategy for navigating high-dimensional molecular design spaces. Active learning methods use machine learning models trained on existing data to identify the most informative next points to sample, typically those that maximize predictive uncertainty, expected improvement, or other acquisition objectives. These techniques can be applied to both virtual screening, where candidate systems are selected for simulation, and physical screening, where experimental synthesis and measurement are guided by model-based recommendations. By selectively choosing which systems to investigate next, active learning enables more efficient discovery of high-performing materials with significantly reduced computational or experimental burden.

1.4 Shifting from Application to Tool-Making

Over the course of this work, my focus gradually shifted from applying machine learning methods to specific molecular simulation problems toward the broader challenge of creating computational tools that others could use and extend. While machine learning has become increasingly influential in the physical sciences, its practical adoption often lags behind its potential. Many promising techniques remain out of reach for researchers who lack the expertise, time, or computational infrastructure to implement and adapt them to their own systems. Even experienced computational scientists may find it difficult to incorporate machine learning into their workflows, given the fast pace of methodological development, the complexity of hyperparameter tuning, and the lack of standardized implementations that balance flexibility with usability.

These challenges are exacerbated by the fact that no single machine learning model or architecture is universally optimal. The success of a method is often highly dependent on the structure of the data, the scale of the problem, and the specific physical questions being asked. As such, even with a general awareness of available techniques, researchers are often left without clear guidance on which method to use or how to adapt a promising method to

the constraints of their particular domain. I experienced firsthand the considerable growing pains associated with using, adapting, and connecting existing ML based methods across a multitude of software. Simple "mom-and-pop" code implementations failed to scale and scripts attempting to massage data into suitable shapes for off-the-shelf methods convoluted debugging workflows, resulting in frequent research setbacks.

One way to address this gap is to integrate machine learning directly into research software infrastructure, making advanced methods more accessible, adaptable, and efficient. This integration is most impactful when guided by three key principles. First, user-friendly operation ensures that non-experts can adopt tools quickly and with minimal friction. This includes well-documented interfaces, meaningful default settings, and intuitive workflows. Second, modular design reflects the inherently interchangeable nature of machine learning components, allowing different models, kernels, or acquisition strategies to be swapped in and out without requiring a complete rewrite of the surrounding infrastructure. Modular architectures also support community-driven development, where users can contribute new capabilities while benefiting from shared maintenance and support. Third, efficient implementation ensures that tools scale with available computational resources, making them practical for large simulations and large datasets.

The creation of PINES and ALPineFOREst serves my lab mates and the greater scientific community by removing unnecessary obstacles to research progress. ALPineFOREst, for example, allows other researchers to bypass the substantial challenges I faced in developing an active learning framework for assessing functionalized nanoparticle systems, enabling future members of the CSSAS EFRC to immediately begin investigating new systems. PINES has already garnered interest for use in highly symmetric molecular environments and is currently being utilized for multiple projects within the group.

This dissertation embraces these principles in the development of the PINES method and ALPineFOREst python library, two modular, ML-integrated software tools designed to

enable data-driven modeling and active learning in molecular simulation. The shift from standalone application to reusable infrastructure not only maximizes the long-term impact of each methodological contribution, but also lowers the barrier for others to build on, adapt, and apply machine learning methods to new scientific questions.

CHAPTER 2

COMPUTATIONALLY GUIDED TUNING OF PEPTIDE-CONJUGATED PERYLENE DIIMIDE SELF-ASSEMBLY

The following chapter is adapted from work originally published in Langmuir. I conducted and analyzed molecular dynamics simulations of the self-assembling π -conjugated peptides. These simulations provided insight into the driving force for self-assembly and intermolecular stacking orientation via the dimer COM-distance PMF and the dimer twist angle PMF respectively.

S.S. Panda, K. Shmilovich, N.S.M. Herringer, N.J. Marin, A.L. Ferguson, and J.D. Tovar “Computationally guided tuning of peptide-conjugated perylene diimide self-assembly” Langmuir 37 28 8594-8606 (2021) [<https://doi.org/10.1021/acs.langmuir.1c01213>]

2.1 Abstract

Peptide- π conjugated materials are important for bio-interfacing charge-transporting applications due to their aqueous compatibility and formation of long-range π -electron networks. Perylene diimides (PDIs), well-established charge-transporting pi-systems, can self-assemble in aqueous solutions when conjugated with amino acids. In this work, we leveraged computational guidance from our previous work to access two different self-assembled architectures from PDI-amino acid conjugates. Furthermore, we expanded the design rule to other sequences to learn that the closest amino acids to the π -core have a significant effect on the photophysical properties of the resulting assemblies. By simply altering glycine to

alanine at the closest residue position, we observed significantly different electronic properties as revealed through UV-vis, photoluminescence and circular dichroism spectroscopies. Accompanying molecular dynamics simulation revealed two distinct types of self-assembled architectures: co-facial structures when the smaller glycine residue is at the closest residue position to the π -core versus rotationally-shifted structures when glycine is substituted for the larger alanine. This study illustrates the use of tandem computation and experiment to unearth and understand new design rules for supramolecular materials, and exposes a modest amino acid substitution as a means to predictably modulate the supramolecular organization and engineer the photophysical properties of π -conjugated peptidic materials.

2.2 Introduction

Self-assembling π -conjugated materials have emerged as an attractive platform to create functional nanoscale electronic materials from molecular building blocks.[20, 21, 22, 23, 24] The gradual growth in the field of bionanotechnology encouraged researchers to look into biocompatible π -conjugated materials with electroactive properties.[25, 26, 27, 28, 29] Due to their ability to form supramolecular architectures through a range of non-covalent interactions, peptides were coupled with π -chromophores to form ordered structures, thus achieving long-range conduction pathways through π -electronic regions in aqueous conditions.[30, 31, 32] The resulting π -conjugated peptide materials can be used to construct field-effect transistors, solar cells or photoconductors based on their ability to transport charges.[33, 34, 35, 36] Incorporation of π -conjugated groups within the peptidic segments introduces quadrupolar interactions that represent substantial perturbations from natural peptide systems. Hence, a delicate balance between non-covalent interactions is necessary to predictably design the optoelectronic outcome of peptide- π conjugates.

Differing electronic interactions that arise from the quadrupolar interactions of π -chromophores in hierarchical supramolecular structures can lead to different optoelectronic and

electrical properties. [37, 38] Supramolecular electronic materials are particularly versatile compared to polymeric and single-crystal π -conjugated systems because their macroscopic properties are easily tunable via relatively small changes in the choice of monomeric building blocks.[20, 21] Peptides offer a powerful route to such tuning because their side-chains can influence a vast range of properties based on local differences in hydrophobicity, acidity, or polarity; hence, sequence variation is often exploited to achieve a diverse range of peptidic π -electron structures and functions. For example, we previously observed excitonic or excimeric interactions in supramolecular peptide- π conjugated systems that were dictated by the side-chain hydrophobicity of the constituent amino acids.[39] The chirality of individual amino acids (D- or L- configurations) also serves as a point of variation to tune electronic properties at the supramolecular level.[40, 41, 42] Moreover, the positional isomerism of different peptide sequences imparts variation of macroscopic and microscopic properties.[43] All of these types of structure variations are readily accessible due to the ease of solid-phase peptide synthesis (SPPS), thus achieving rapid variation of morphological and optoelectronic properties of the self-assembled structures.

Perylene diimides (PDIs) were investigated extensively due to their excellent optoelectronic and electrical properties.[44, 45, 46, 47] The high electron affinity and relatively efficient charge transport properties make PDIs useful n-type materials in different organic electronic devices such as field-effect transistors,[48] solar cells[49] and light-emitting diodes[50]. Their macroscopic properties can be tuned easily by variation of the functionalizing group at the imide nitrogen position and on the ‘bay’ regions of the perylene, which can lead to significant differences in the orientation and separation (both longitudinal and transverse direction) of inter-chromophore packing and subsequently in the observed optoelectronic and electrical properties.[51] PDIs were coupled with single amino acids at both ends and their impact on optoelectronic properties was investigated.[52, 53, 54, 55] For example, Adams and co-workers investigated optoelectronic properties of several peptide-PDI conjugates in solu-

tion and gel form by systematically varying the amino acids with aliphatic and aromatic side chains.[53, 54] Polypeptide-PDI conjugates have been also studied previously.[5, 6, 56, 57, 58] Hodgkiss and co-workers identified potentially three different states of PDI chromophores within the aggregates: monomeric, H-aggregated and π -aggregated that shows a broad, featureless signal similar to liquid crystalline forms.[58, 5] These broad featureless absorption signals were explained as a mixture of H-type and J-type aggregates by Frauenrath and co-workers.[57] However, a predictable trend of optoelectronic properties of peptide-PDI conjugates based on the amino acid sequences remains elusive.

Recently, Thurston and Ferguson conducted molecular dynamics simulations of self-assembled peptides with D-X1-X2-X3 sequences conjugated to a PDI core.[59] They developed a key metric ' α ' which corresponds to the ratio of 'aligned' oligopeptides in a cluster to the 'associated' oligopeptides in a cluster. Two peptides are said to be 'associated' if the distance between any two of their atoms is less than 0.5 nm. Similarly, peptides were designated as 'aligned' if the distance between the atoms in the π -core is less than 0.5 nm. This metric provides a way to quantify the good geometric alignment of the π -cores within the oligopeptide self-assembly. New peptides DAVG-PDI-GVAD ($\alpha = 0.64 \pm 0.07$) and DAIA-PDI-AIAD ($\alpha = 0.58 \pm 0.05$) were identified to have high cofacial intermolecular alignment, but until now had not been synthesized and characterized experimentally. In this report, we describe the optoelectronic properties of these two peptides, showing two very different outcomes within self-assembled nanomaterials derived from them. To further challenge this finding, we report here a small library of PDI-peptide conjugates that vary by the nature of the amino acid residues closest to the PDI central core, and we present data that indicates a comparable optoelectronic dichotomy can be engineered by design into PDI peptide nanomaterials. This research highlights the power of synergistic feedback loops between computational and experimental peptide nanomaterial design in the context of realizing advanced optoelectronic function.

2.3 Methods

2.3.1 *Experimental Methods*

General Information N-Methyl-2-pyrrolidone (NMP) was obtained from Advanced ChemTech. Pyridine was obtained from Alfa-Aesar. Dichloromethane (DCM) was 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'-tetramethyluronium hexafluorophosphate (HBTU) was purchased from Oakwood Products Inc. Wang resin (preloaded with amino acid) and Fmoc-protected amino acids were obtained from Advanced Chem Tech. All other reagents and starting materials were obtained from Sigma-Aldrich and were used as received.

General Solid-Phase Peptide Synthesis (SPPS) All peptides were synthesized using the standard Fmoc solid-phase technique with Wang resin preloaded with Fmoc-protected valine (Wang-Val, 0.700 mmol/g). 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 45-90 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 as needed. The general procedure for amino acid coupling was repeated for each additional amino acid until the desired peptide sequence was obtained. Last amino acid was deprotected using piperidine solution leaving a -NH₂ terminated resin.

General On-Resin Dimerization Procedure Following a previously reported literature, [60] -NH₂ terminated resin (0.10 mmol) and 3,4,9,10-Perylenetetracarboxylic dianhydride (PDA) (0.012 g, 0.031 mmol) were transferred to the Schlenk flask equipped with a reflux condenser. The mixture was dried under vacuum and was suspended in pyridine (5.0 mL) thereafter. This suspension was heated to 65 °C followed by the addition of diisopropyl ethyl amine (0.25 mL, 1.5 mmol) and reaction was further heated to 135 °C for 10-12 hours. A second portion of PDA (0.008 g, 0.02 mmol) was added along with a second portion of pyridine (3.0 mL). Reaction was maintained at 135 °C and continued for an additional 10-12 hours. Resin was cooled to room temperature, filtered and washed with MeOH. The resin was then suspended in pyridine (6.0 mL), heated to 65 °C followed by the addition of diisopropyl ethyl amine (0.25 mL, 1.5 mmol) and further heated to 135 °C for 6 hours. Resin was cooled to room temperature.

General Peptide Cleavage and Workup Procedure Following dimerization, the resin was returned to the peptide chamber and again subjected to a standard wash cycle: 1 × H₂O, 1 × 2-Propanol, 1 × CH₃CN, 2 × NMP, 2 × methanol, and 2 × DCM. The resin was treated with 14.25 mL of trifluoroacetic acid, 375 μL of water, and 375 μL of triisopropylsilane for 2 h. The peptide solution was filtered from the resin beads, washed 3 times with DCM, and concentrated by evaporation under reduced pressure. The crude peptide was then precipitated from solution with 60-80 mL of diethyl ether and isolated through centrifugation. The resulting pellet was triturated with diethyl ether to yield crude product, which was dissolved in approximately 15-20 mL of water and 50 μL of potassium hydroxide (1 M) and lyophilized.

Characterization details of peptide-pi molecules can be found in the supporting information.

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.

Reverse-Phase HPLC HPLC purification was performed on an Agilent 1100 series (semi-preparative/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.

UV-vis and Photoluminescence UV-vis spectra were obtained using a Varian Cary 50 Bio UV-vis spectrophotometer. Photoluminescence spectra were obtained using a PTi Photon Technology International Fluorometer (QuantaMaster 40) with a 75 W Ushio Xenon short arc lamp and operated with Felix32 Version 1.2 software. Spectroscopic samples were prepared by diluting the peptide solution to the appropriate concentration (concentration mentioned in spectra captions) in Millipore water. Neutral condition refers to the sample only in Millipore water at appropriate concentration. The pH was then adjusted by adding 15 μL of 1M KOH for basic solutions and 40 μL of 1M HCl for acidic solutions.

Circular Dichroism (CD) D spectra were obtained using a Jasco J-810 spectropolarimeter. Spectroscopic samples were prepared by diluting the peptide solution to the appropriate concentration (concentration mentioned in spectra captions) in Millipore water. Neutral condition refers to the sample only in Millipore water at appropriate concentration. The pH was then adjusted by adding 25 μL of 1M KOH for basic solutions and 60 μL of 1M HCl for acidic solutions.

Dynamic Light Scattering (DLS) DLS spectra were obtained using a Zetasizer Nano-ZS90 (Malvern Instruments). Spectroscopic samples were prepared by diluting the peptide solution to the appropriate concentration (concentration mentioned in spectra captions) in

Millipore water. Neutral condition refers to the sample only in Millipore water at appropriate concentration. The pH was then adjusted by adding 15 μ L of 1M KOH for basic solutions and 40 μ L of 1M HCl for acidic solutions.

Attenuated total reflection Fourier transform infrared (ATR-FTIR) ATR-FTIR spectra were obtained on lyophilized acidic peptide solutions using a Thermo Scientific Nicolet iD5 ATR-IR.

Transmission Electron Microscopy (TEM) Imaging was performed on a FEI Technai 12 TWIN transmission electron microscope equipped with an SIS Megaview III CCD digital camera at an accelerating voltage of 100 kV. Acidic solutions (0.02 wt %) of each peptide were prepared by placing the samples in a closed container with a vial of conc. HCl opened within, which initiated the assembly process by diffusion of HCl vapor to the sample. The peptides were pipetted (drop of 1 mg/mL solution) onto 200 mesh copper grids coated with carbon and incubated for 5 min at 25 °C. Excess solution was wicked off by touching the side of the grid to filter paper. The samples were then stained with a 2% uranyl acetate solution, and excess moisture was wicked off. The grid was allowed to dry in air before imaging.

2.3.2 Computational Methods

All-atom molecular modeling All molecular dynamics simulations of the DAVG, DAIA, VEVAG, and VEVA peptides were performed using GROMACS 2019.2 simulation suite[7] patched with the PLUMED v. 2.5.2[61] enhanced sampling libraries. Each π -conjugated peptide was modeled using the AMBER99SB-ILDN[62] forcefield with all-atom topologies generated using the AnteChamber PYthon Parser interfAcE (ACPYPE)[63] wrapper for ANTECHAMBER (AmberTools17.0) (D.A. Case, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D.Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermel-

stein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman (2017), AMBER 2017, University of California, San Francisco.). The ionizable Asp and Glu side chains and two C-termini carboxylates were prepared in a fully protonated state corresponding to low-pH (pH < 1) conditions. Partial charges were obtained using the Restrained Electrostatic Potential (RESP) method[64, 65] with calculations performed in Gaussian16[66] at the B3LYP/6-31G(d) level of theory. Water solvent was modeled using the TIP3P water model[67]. The leapfrog algorithm was used to numerically integrate forward Newton’s equations of motion with a timestep of 2 fs.[68] Simulations were performed in three-dimensional cuboidal simulation boxes with periodic boundary conditions. The LINCS algorithm[69] was used to constrain all covalent bonds with hydrogen atoms. Lennard-Jones interactions were smoothly shifted to zero at a 1.0 nm cutoff and particle mesh Ewald (PME)[70] with a real space cutoff of 1.0 nm and a Fourier grid spacing of 0.16 nm were used to treat Coulombic interactions. The real space cutoff and Fourier grid spacing values are optimized during runtime for performance. A Parrinello-Rahman barostat[71] and velocity rescaling thermostat[72] were used to control pressure and temperature, respectively. Initial system configurations were subjected to steepest descent energy minimization to eliminate forces larger 10 kJ/mol.nm and initial velocities assigned by sampling from a Maxwell-Boltzmann distribution at 300 K. Systems were then equilibrated under 100 ps of simulation in the NVT ensemble at 300 K followed by 100 ps of simulation in the NPT ensemble at 300 K and 1 bar.

Well-tempered metadynamics calculations The potential of mean force (PMF) as a function of the signed twist angle θ between a pair of oligopeptides within a molecular dimer were estimated using well-tempered metadynamics[73], following a protocol that we have previously reported[40, 74]. In brief, we perform well-tempered metadynamics along the single collective variable (CV) defined as the signed twist angle θ between the PDI cores of

a dimerized pair of molecules:

$$\theta = \arccos(\mathbf{m}_1 \cdot \mathbf{m}_2) \cdot \text{sgn}((\mathbf{m}_1 \times \mathbf{m}_2) \cdot \mathbf{d}) \quad (2.1)$$

Here, $\mathbf{m}_1$ and $\mathbf{m}_2$ are vectors that connect the terminal nitrogen atoms of the PDI core of arbitrarily labeled molecule 1 and molecule 2, respectively, and $\mathbf{d}$ is a vector that points from the center of mass (COM) of the PDI core of molecule 1 to molecule 2. The magnitude of the angle between the two PDI cores is captured in the first term $\arccos(\mathbf{m}_1 \cdot \mathbf{m}_2)$ of eq 1 that is endowed with a positive or negative sign by the second term $\text{sgn}((\mathbf{m}_1 \times \mathbf{m}_2) \cdot \mathbf{d})$ that captures that handedness of the dimer configuration via the sign of the projection of $\mathbf{d}$ onto the cross product $\mathbf{m}_1 \times \mathbf{m}_2$. A positive signed twist angle ($\theta > 0$) corresponds to M-type (left-handed) chirality, while a negative signed twist angle ($\theta < 0$) corresponds to P-type (left-handed) chirality. Sampling is focused on the range $\theta \in [-1.25, 1.25]$ radians by applying a harmonic restraint with force constant $k_\theta = 1000 \text{ kJ/mol.rad}^2$ if θ departs this range. To maintain the integrity of the dimer, an additional harmonic restraint with force constant $k_d = 1000 \text{ kJ/mol.nm}^2$ was applied to the COM distance $|\mathbf{d}|$ between the aromatic cores of the two molecules if $|\mathbf{d}| > 0.6 \text{ nm}$ for the smaller DAVG system and $|\mathbf{d}| > 0.7 \text{ nm}$ for the larger DAIA, VEVA, and VEVAG systems. Each dimer was initialized with $|\mathbf{d}| = 0.45 \text{ nm}$ and $\theta = +10^\circ$ and solvated within a $7 \times 6 \times 4 \text{ nm}^3$ simulation box with TIP3P water.[67] Periodic boundary conditions were implemented in all three dimensions. The box was designed to be sufficiently large to accommodate at least 1.50 nm of spacing between any atom in the dimer pair and box edges such that there are no spurious interactions between periodic images of the dimer.

Well-tempered metadynamics simulations were performed using the PLUMED enhanced sampling libraries[61] patched to GROMACS 2019.2[7]. Calculations were performed independently for each system and trajectory snapshots saved every 10 ps for analysis. Each metadynamics production run used the signed twist angle θ as the biasing CV and deposited

Gaussians with initial height of 1.2 kJ/mol and width 0.2 radians every 1 ps using a bias factor of $\gamma = 20.0$ for the DAVG, DAIA, and VEVAG systems, and $\gamma = 22.5$ for the VEVA system. We define convergence by ensuring that both (1) the height of the Gaussians deposited by the well-tempered metadynamics falls to approximately zero ($< \sim 0.05$ kJ/mol) indicating that the bias potential has converged, and (2) collective sampling along the signed twist angle θ becomes approximately uniform, defined by the criterion that the fraction of configurations with $\theta < 0$ and $\theta > 0$ is within $\sim 5\%$. Convergence was achieved in the DAVG and DAIA systems after 7 μ s of simulation, and in the VEVGA and VEVA systems after 6 μ s and 5.61 μ s, respectively.

The well-tempered metadynamics potentials of mean force are used to estimate the cumulative probability distribution (CDF) in the absolute magnitude of the signed twist angle $CDF(|\theta|)$. This function measures the integrated probability of observing the system with a twist angle of magnitude $|\theta|$ or smaller. Given the unbiased probability density function $P(\theta)$, the probability of observing a dimer at or below a given twist angle magnitude is given by:

$$CDF(|\theta|) = \frac{\int_{-\theta}^{\theta} P(\theta^*)}{\int_{-\pi/2}^{\pi/2} P(\theta^*)} \approx \frac{\sum_{i=0}^N w_i \text{ if } |\theta_i| \leq |\theta| \text{ and } |\mathbf{d}_i| \leq d_{cut} \text{ otherwise } 0}{\sum_{i=0}^N w_i \text{ if } |\theta_i| \leq 1.25 \text{ and } |\mathbf{d}_i| \leq d_{cut} \text{ otherwise } 0} \quad (2.2)$$

where the second expression represents the numerical estimate computed from our well-tempered metadynamics runs in which we approximate the integrals as sums. The index i iterates over the N simulation snapshots saved over the course of the metadynamics run, and w_i is the weight associated with each frame under reweighting to the unbiased ensemble. The summations are restricted to the volume of phase space over which the angular and COM harmonic restraints are not active, corresponding to $|\theta_i| < 1.25$ and $|\mathbf{d}_i| < d_{cut}$, where $d_{cut} = 0.6$ nm for DAVG and $d_{cut} = 0.7$ nm for DAIA, VEVA, and VEVAG, to remain within

our definition of a dimer pair defined during our metadynamics simulation. It is useful to observe that in the case of an unbiased simulation there would be no externally applied bias and each weight $w_i = \frac{1}{N}$ in the summation and eq 2 reduces to counting the fraction of configurations with $|\theta_i| < |\theta|$.

Umbrella sampling calculations Umbrella sampling was performed in the COM distance between a pair of peptides in order to estimate the dimerization potential of mean force curves quantifying the free energy change for the assembly of an equilibrium dimer. Calculations were performed using GROMACS 2019.2 and the data analyzed using the g_wham module implementing the Weighted Histogram Analysis Method (WHAM)[75] and applying the entropic correction[7, 76]. Dimers were initialized at the signed twist angle θ corresponding to the global minimum of the free energy landscapes in θ computed by well-tempered metadynamics (see Well-tempered metadynamics calculations) and then subjected to energy minimization and relaxation (see All-atom molecular modeling). In all cases, the angle θ deviated no more than 15° from its initial value after performing the relaxation. The box dimensions were chosen to be $7.5 \times 7.5 \times 7.5 \text{ nm}^3$ for the VEVAG and VEVA molecules and $12 \times 12 \times 12 \text{ nm}^3$ for the DAVG and DAIA molecules. In each case the box was sufficiently large that there were no spurious interactions between periodic images of the molecules through the periodic boundaries even at the largest values of $|\mathbf{d}|$ considered. A steered molecular dynamics simulation was then performed to pull the dimerized pair apart to a distance of $|\mathbf{d}| \text{ 3.5 nm}$ by applying a force constant of 1000 kJ/mol.nm^2 to the COM distance $|\mathbf{d}|$. Frames extracted from this pulling trajectory were used to initialize umbrella sampling calculations at particular values of $|\mathbf{d}|$. We stratified sampling into 16 umbrella sampling windows at a spacing of 0.2 nm over the range $|\mathbf{d}| = 0.3\text{-}3.3 \text{ nm}$. Each umbrella simulation was performed in the NPT ensemble at 300 K and 1 bar and was first equilibrated for 100 ps before being subjected to a 10 ns production run. Harmonic umbrella potentials were applied with spring constants $k_u = 1000 \text{ kJ/mol.nm}^2$ and centered on the initial value of $|\mathbf{d}|$

within the window. As necessary, additional umbrella windows were run to fill in any gaps in the overlap between the umbrella window histograms and the PMF is recalculated with g_wham. Uncertainties are estimated by performing three independent sets of umbrella sampling simulations for each system.

2.4 Results and Discussion

Validation of the computational guidance. We prepared the two sequences revealed from computational screening to have high co-facial PDI stacking, corresponding to a high degree of geometric overlap of the PDI π -cores rather than an optimal geometry that would favor electrical transport or exciton migration. The sequence space screened in this initial assessment was constrained to four amino acid residues within the peptide flanking wings, with the terminal position fixed as aspartic acid (D) in order to impart pH sensitivity to the assembly process. The peptides are effectively ionized and dissolved at basic pH, although there is computational[77] and experimental evidence of molecular pre-association/aggregation (see Figures S23-25). The assembly into tape/fibril-like nanomaterials is initiated upon protonation of the terminal carboxylates at acidic pH. In these determined sequences, it is notable that both DAIA and DAVG have a relatively small residue closest to the PDI core (A or G) and a larger hydrophobic group in the next position (I or V). Using standard Fmoc-based solid-phase peptide synthesis in conjunction with our past methods to incorporate the PDI π -system (see experimental section/SI), PDI cores were substituted with the DAIA and DAVG sequences (Figure 1). Our experimental and computational investigation revealed the two distinct types of self-assembled architectures from these PDI-peptide conjugates.

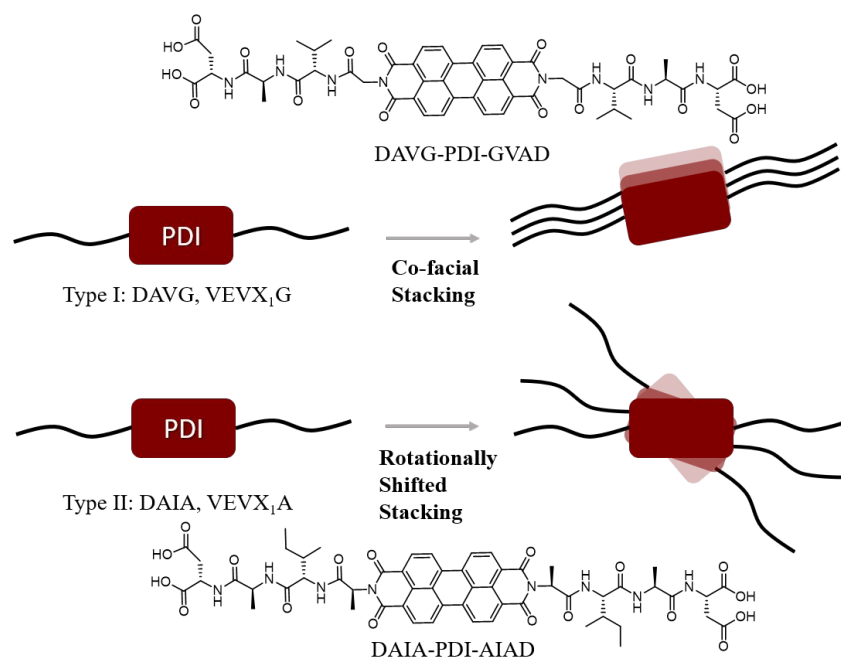


Figure 2.1: Chemical structures of the DAVG-PDI-GVAD and DAIA-PDI-AIAD peptides inspired by computational input.

These peptides form 1-D tape-like nanostructures when exposed to acidic aqueous environments in line with other related molecules (Figure 2), and the photophysical responses were probed in basic (ca. pH $\approx$ 12.5), neutral (ca. pH $\approx$ 7), and acidic (ca. pH $\approx$ 1.2) media to observe the extent of self-association at different pH. The UV-Vis responses of DAVG-PDI under basic and neutral conditions showed signatures typical of PDI chromophores: two distinct peaks at 535 nm and 502 nm corresponding to 0-0 and 0-1 transitions, respectively. The ratio of 0-0/0-1 peak is about 0.61 and 0.70 in basic and neutral medium similar to the ratio observed for other PDI-conjugated materials in aqueous media.[78, 79] These band ratio values are in between the ratio expected for PDI-monomer (>1.4) and stacked dimers or aggregates (<0.6).[78, 80] However, acidic environments resulted in a broad absorption profile with new peaks arising around 450 nm, 585 nm and 625 nm (Figure 2). A distinct color change from red to black was also observed upon going from basic to acidic media. Previously, these types of shifts in absorption maxima were explained by formation of liquid crystalline-type materials,[5, 81, 82] J-aggregates[83, 84] or mixture of H- and J-aggregates. However, similar broadening and characteristic shifts of absorption maxima have also been reported in perylene monoimide systems due to charge transfer and Frenkel state mixing.[85, 86, 87] This mixing allows for extended exciton separation and an elongated charge transfer band. DAIA showed a similar response to DAVG in the neutral medium, but the clear separation between 0-0 and 0-1 peaks was not as pronounced in the basic medium. The ratio of 0-0/0-1 peak is about 0.89 and 0.86 respectively for basic and neutral medium, which is higher than the ratio observed for DAVG indicating presence of more monomeric structure.[79, 80] The spectral response of DAIA under acidic conditions was vastly different from DAVG, maintaining the general signature observed under basic conditions for both DAVG and DAIA, which is the prototypical UV-Vis signature of PDI chromophores. This relatively broader and unstructured absorption spectrum with reduced extinction coefficient is hallmark of H-type assembled structure in PDI-conjugated molecules in aqueous

solution.[80, 85] We classify these two different photophysical responses within the acidic aggregates as Type-I (DAVG-PDI) and Type-II (DAIA-PDI) for the rest of the article.

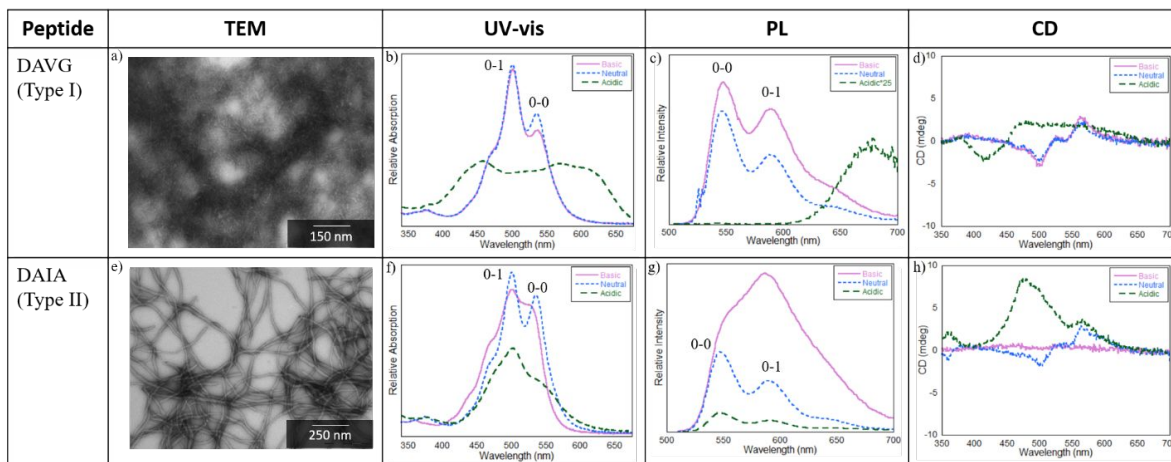


Figure 2.2: Transmission electron microscopy (TEM) images of assembled peptides (a, e), UV-Vis (b, f), PL (c, g) and CD (d,h) spectroscopy traces of DAVG (a-d) and DAIA (e-h) peptides in basic (solid pink), neutral (blue short dash), and acidic (green long dash) media. UV-Vis and PL spectra were recorded at 5-7.5 μ M concentration range. CD spectra were recorded at 15-20 μ M concentration range.

PL spectroscopy of DAVG-PDI in neutral and basic media showed three distinct peaks around 540 nm, 590 nm and 640 nm. The peak around 640 nm was previously assigned to H-aggregate formation whereas the 540 nm and 590 nm peaks can be assigned to the 0-0 and 0-1 transitions, respectively.[84] The basic medium showed higher fluorescence intensity as compared to neutral solution due to the greater degree of molecular dissolution. The acidic solution of DAVG formed a broad low-energy excimer like signal ($\lambda_{max} \approx 675$ nm) which was observed previously in liquid-crystalline or J-aggregate materials.[81, 82, 83, 84] DAIA showed a similar response to DAVG in neutral media. However, in basic media, the 0-1 transition, which had lower intensity in neutral condition as compared to the 0-0 transition, showed a prominent rise, and the peak around 640 nm was not observed which may be buried in the red-tail of the 0-1 peak. It is possible that a subtle change in aggregation occurred in basic media which was also evident from the slightly different spectral shapes apparent in the UV-Vis spectra. However, the higher intensity of signal in basic medium can again be attributed to the disassembly of aggregates. The acidic assembled DAIA sample showed high-energy ‘exciton’ like vibronic signatures similar to the neutral and basic solutions, in stark contrast to assembled DAVG.

CD spectroscopy of DAVG-PDI showed weak Cotton effects in basic and neutral media around 0-0 and 0-1 transition similar to other PDI-oligopeptide assemblies, indicating weak exciton-coupled states.[56] Acidic media presented different signatures: low-energy absorption bands which suggested that the molecules were in a chiral environment, and a weakly bisignate higher-energy 450 nm band which indicated weak exciton coupling. At basic pH, DAIA showed no characteristic signal indicating no interaction between the chromophores. However, it showed a weak Cotton effect in neutral solution as observed by a crossover at the wavelength coinciding with the absorption maxima. This shows similarity to DAVG in neutral medium indicating that pi-systems are weakly exciton-coupled in neutral media forming similar types of assemblies. Further lowering the pH showed no crossover around the

absorption maxima, however the intensity of signal was increased indicating the molecules are held in a chiral environment.

Expansion and generalization of the mined sequences. The molecular simulations of Thurston and Ferguson predicted high alignment propensities for both DAVG-PDI and DAIA-PDI, but the observed photophysical profiles reported herein showed vastly different responses. We hypothesized that residues closer to the π -core may have significant influence on the assembly process and the resulting photophysical response. Thus, we extended our investigation to a related series of peptide-PDI conjugates inspired by the initial computational guidance based on the general $VEVX_1X_2$ sequence tagged to the PDI core (Figure 3). This peptide was chosen based on our prior work that showed the VEVAG pentapeptide to be a good consensus sequence to form peptide- π -peptide conjugates when appended to a variety of chromophores, all of which displayed good aqueous solubility at basic pH and 1-D assembly fidelity at acidic pH. We varied the steric bulk of the side-chains of amino acids at the X_1 position using glycine (G, R = H), alanine (A, R = CH₃), valine (V, R = CH(CH₃)₂) and phenylalanine (F, R = CH₂Ph). Attempts to vary the steric bulk at the X_2 position were met with some difficulties due to the proximity to the PDI core: we could only achieve the addition of glycine (G, R = H) and alanine (A, R = CH₃). We found previously that more hydrophobic residues formed closely coupled aggregate states due to the higher van der Waals-type interaction among residues outcompeting aqueous solvation.[39] However, these more hydrophobic residues can also lead to the increased steric crowding within these assemblies. These systematic variations of ‘steric bulk’ at different positions of the pentapeptide sequence provide some guidance about the effect of amino acids on internal structure and optoelectronic properties of the PDI-based assemblies.

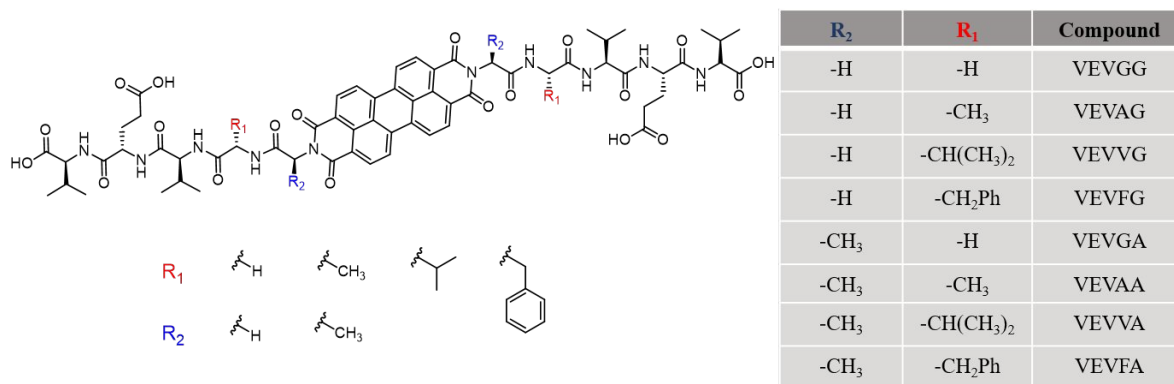


Figure 2.3: Chemical structures of the members of the VEVX1X2-PDI-X2X1VEV peptide libraries.

1-D tapelike nanostructures with high aspect ratios were visualized from all of the VEVX1X2 peptides when aqueous basic solutions (0.2 mg/mL) were exposed to HCl vapor within a closed chamber. In general, the nanostructures were micrometers in length and comparable to other peptide- π -peptide conjugates reported previously. The switching of amino acids in the X1 and X2 positions did not have any significant impact on the overall one-dimensionality of the supramolecular nanostructures; however, the width of the nanostructures varied significantly. For example, VEVAG presented widths of 8.8 ± 0.9 nm while VEVA A had widths of 14.5 ± 0.7 nm. These two peptides are shown as representative examples of nanostructures in Figure 4.

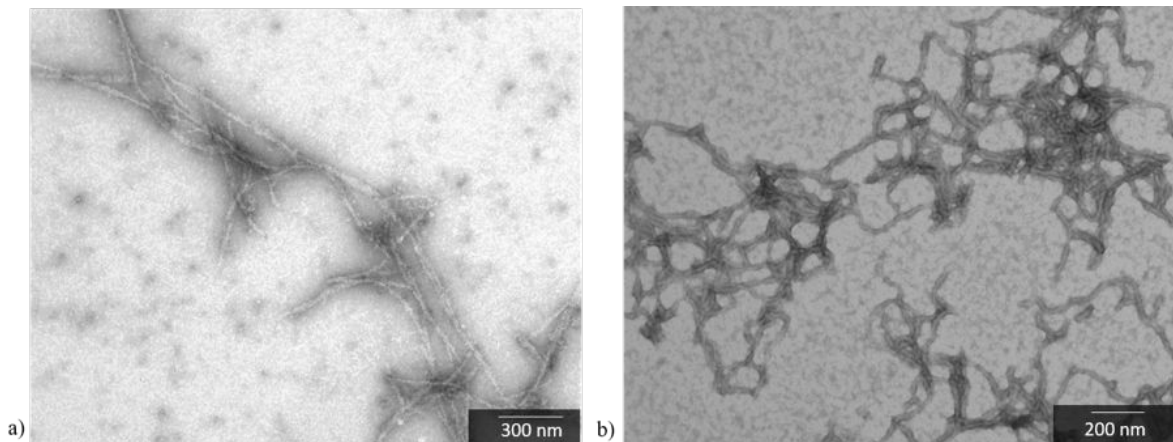


Figure 2.4: TEM images of a) VEVAG, and b) VEVA A from acidic solution of concentration 0.2 mg/mL.

VEVX1G peptides recorded similar photophysical responses to that of DAVG (Type I). UV-Vis spectroscopy showed two distinct peaks at 535 and 502 nm under basic and neutral media (representative example in Figure 5a). However, the absorption recorded from acidic solution showed broadening for all the peptides similar to the Type-I response but with slight variation in intensity and magnitude. VEVGG showed absorption maxima around 480 nm and 560 nm and the broadening was lesser than the other peptides of this family. VEVAG and VEVVG showed absorption maxima around 450 nm and relatively broader signals. VEVFG showed a maximum around 465 nm and additional peaks were observed at 530 nm and 560 nm (Figure 5b). The photophysical responses from VEVX1A peptides matched that from DAIA (Type II). The representative absorption profile of VEVA is shown in Figure 5c. Generally, two peaks (i.e. 535 nm corresponding to 0-0 transition and 502 nm corresponding to 0-1 transition) were observed for this family of peptides regardless of the condition, but the intensity varied significantly. In basic media, the 0-0/0-1 peak intensity ratio increased indicative of aggregate dissolution. Acidic solutions showed higher weight towards the 0-1 peak indicating H-type assembly processes for all the peptides in this family.

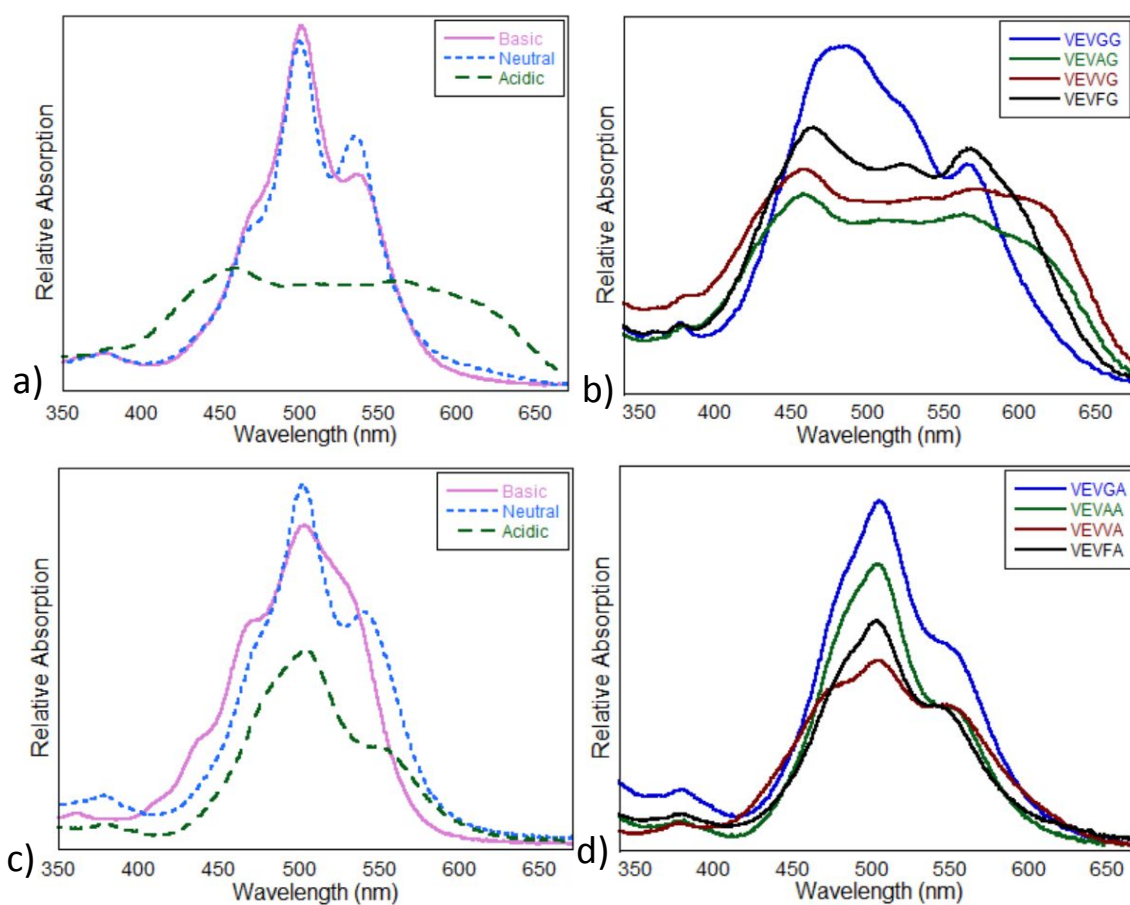


Figure 2.5: UV-Vis signatures of a) VEVAG under basic, neutral, and acidic conditions, b) VEVX1G peptides under acidic conditions, c) VEVA A under basic, neutral, and acidic conditions, and d) VEVX1A peptides under acidic conditions. All the spectra were recorded at 5-7.5 μM concentration range.

Generally, all VEVX1G peptides showed similar PL signals to that of Type I DAVG-PDI. Emissive responses with clear vibronic structures were observed under basic and neutral conditions whereas acidic solutions showed broad emissive signals. (VEVAG as a representative example is shown in Figure 6a). VEVAG showed more of a quenched signal at basic pH as compared to the neutral medium which is indicative of salt-driven assembly process at higher pH. Similarly, VEVX1A peptides showed similar emission responses to Type II DAIA-PDI as anticipated by the UV-vis responses. Upon lowering the pH, signal quenching was observed but the peak positions remained unchanged, which is typical of weakly-coupled H-type aggregates. VEVAA PL is shown as a representative example in Figure 6b. Here, a new excimer peak started to arise (around 675 nm) reminiscent of what was observed for Type-I peptides, but the other features mirror those of the Type II DAIA peptides. This low energy excimeric peak was absent for VEVGA but more pronounced for VEVVA and VEVFA. These data indicate the distinctive nature of variation of steric bulk with respect to particular position in the peptidic backbone of peptide-PDI conjugates.

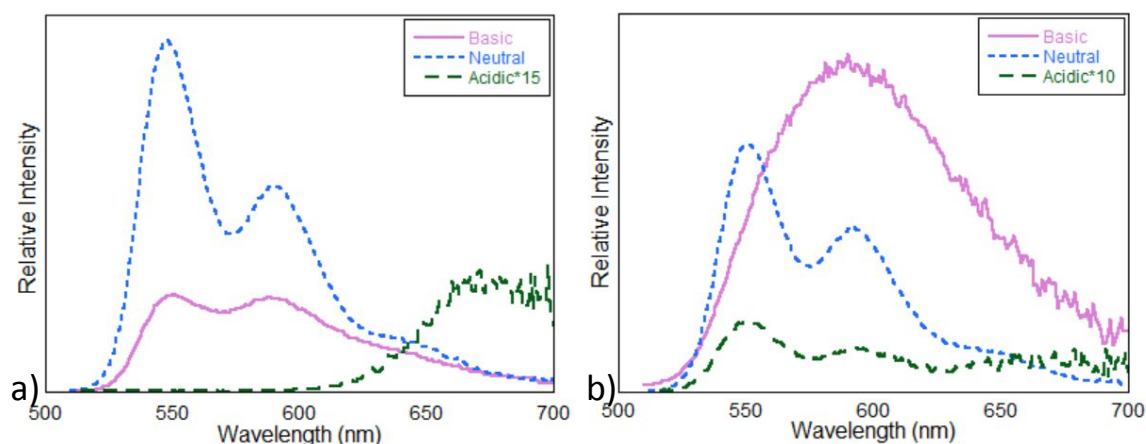


Figure 2.6: PL signature of a) VEVAG under basic, neutral, and acidic condition. b) VEVAA under basic, neutral, and acidic condition. All the spectra were recorded at 5-7.5 μM concentration range.

VEVX1G peptides show relatively weak Cotton effects under basic and neutral conditions, reminiscent of the DAVG Type I peptide. CD spectroscopy revealed uniform profiles under acidic media for all the peptides: low-energy absorption bands suggesting that the molecules are in chiral environments, and high-energy bands (e.g. 450-480 nm) that were exciton coupled as observed via signal crossover at an absorption maxima (for a representative example, see Figure 7a). VEVGG showed a relatively stronger Cotton effect in acidic medium which is expected as it also shows different signature in UV-Vis spectroscopy from the other peptides in the same family. VEVX1A peptides recorded similar signatures to the Type-II DAIA peptide. Peptides in basic media showed no CD signature indicating a lack of interaction between chromophores or other local chirality. Upon lowering the pH to neutrality, Cotton effects were observed for the low-energy band i.e. around 500 nm. The high-energy CD features indicated that the chromophores are in chiral environment but not specifically exciton coupled. Further lowering the pH to acidic, the intensity of Cotton effect increased indicating strong exciton coupling in between chromophores (see Figure 7b for a representative example).

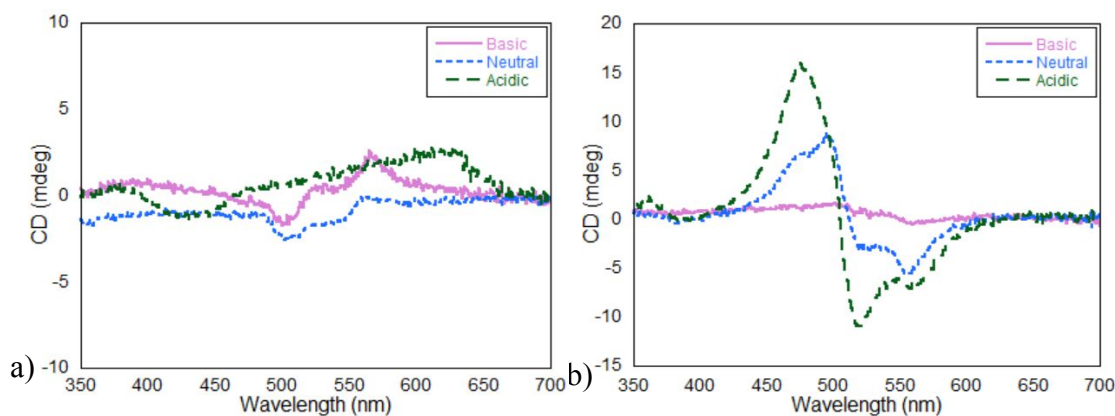


Figure 2.7: CD signature of a) VEVAG under basic, neutral, and acidic condition. b) VEVAA under basic, neutral, and acidic condition. All the spectra were recorded at 15-20 μ M concentration range.

Molecular modeling of pi-peptide packing and stability. To gain insight into the molecular structure and dynamics of the self-assembled aggregates, we performed all-atom molecular dynamics simulations of the DAVG, DAIA, VEVAG, and VEVA systems under acidic conditions where experimental assembly into self-assembled nanoaggregates was observed. We adopt a reductionist posture in which we study the behavior of a dimer pair as representative of the behavior of molecules with larger multibody aggregates. In doing so we sacrifice explicit representation of multibody effects for a minimal system whose thermally accessible configurational phase space we can exhaustively sample using enhanced simulation techniques to obtain converged estimates of dimer structure and stability (see Computational Methods). We have previously adopted this approach to gain molecular-level understanding of the determinants of supramolecular chirality and demonstrated good agreement between the computational prediction and experimental measurements using circular dichroism.[40, 74]

We report in Figure 8 the potential of mean force (PMF) as a function of signed twist angle between the two members of the molecular dimer for each of the DAVG, DAIA, VEVAG, and VEVA peptide systems. The signed twist angle is defined as the relative angle subtended between the two PDI cores and is defined mathematically in eq. 1. A signed twist angle of $\theta = 0$ corresponds to cofacial parallel stacking, a positive signed twist angle ($\theta > 0$) to M-type (left-handed) chirality, and a negative signed twist angle ($\theta < 0$) to P-type (left-handed) chirality. The PMFs for each of the four systems exhibit two-well character, with local free energy minima in the $\theta > 0$ (M-type, left-handed) and $\theta < 0$ (P-type, right-handed) portions of the range. The relative depths of these free energy minima expose the relative preferences for left- versus right-handed packing of the dimer. Our results identify very little preference in chirality for DAVG and VEVAG, with the free energy difference between the two wells lying at or below the 1 kT level of thermal fluctuations – where k is Boltzmann’s constant and T is the absolute temperature, and $\beta = \frac{1}{kT}$ – with $\Delta\beta F = (0.14 \pm 0.04)$ kT and $\Delta\beta F = (0.98 \pm 0.14)$ kT for DAVG and VEVAG, respectively. Conversely, our calculations do

predict a weak thermodynamic preference for left-handed (M-type) supramolecular chirality for DAIA ($\Delta\beta F = (3.5 \pm 0.4)$ kT) and VEVA ($\Delta\beta F = (1.5 \pm 0.3)$ kT). These results suggest that a (left-handed) chiral preference is associated with the placement of a bulkier Ala residue in the X_2 position adjacent to the PDI core.

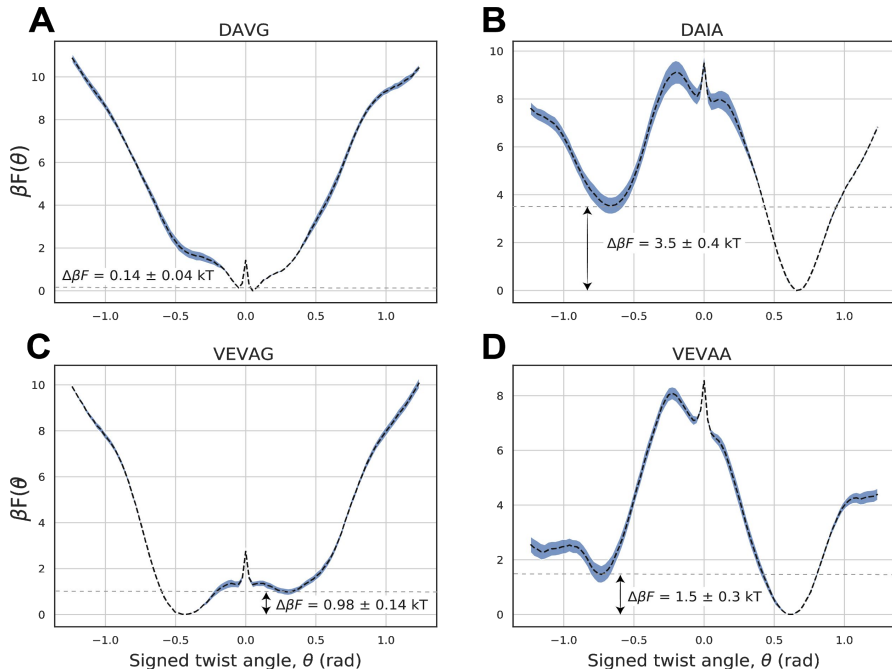


Figure 2.8: Potential of mean force (PMF) as a function of signed twist angle (θ) recovered from well-tempered metadynamics simulations of (A) DAVG, (B) DAIA, (C) VEVAG, and (D) VEVA. Metadynamics simulations consider a dimerized pair of molecules biased along the collective variable defined as the signed twist angle (θ). Annotations indicate the difference in free energy ΔF between the local free energy minima in the $\theta > 0$ and $\theta < 0$ regimes. Standard errors are illustrated by shading and were estimated by block averaging our trajectories into five equally sized contiguous blocks.

To further explore this observation, we report in Figure 9 the cumulative distribution functions in the absolute magnitude of the signed twist angle for each of the four systems. This quantity reports the equilibrium probability of observing a dimer with a twist angle of magnitude or less. In comparing DAVG vs. DAIA and VEVAG vs. VEVAA, we observe that the presence of a smaller Gly residue in the X_2 position immediately adjacent to the PDI core results in a higher prevalence of smaller magnitude twist angles compared to the presence of a bulkier Ala residue that leads to an elevated preference for larger magnitude twist angles.

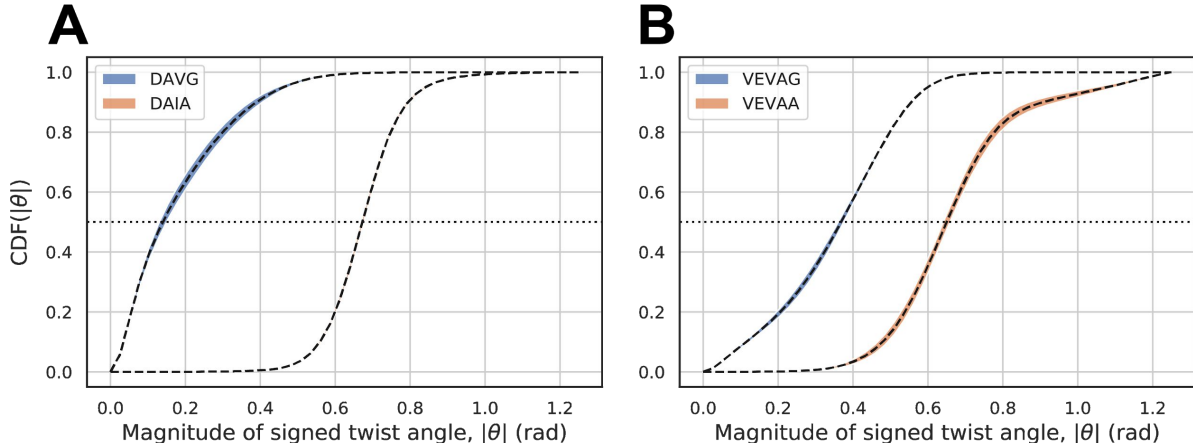


Figure 2.9: Cumulative distribution function (CDF) for the probability of observing a dimer pair at or below a given signed twist angle magnitude $|\theta|$ for (A) DAVG vs. DAIA and (B) VEVAG vs. VEVAA. Shading represents standard errors estimated by block averaging where each trajectory is split into five equally sized contiguous blocks. The placement of the smaller Gly residue in the X_2 position immediately adjacent to the PDI core leads to a higher observed probability of smaller magnitude twist angles, whereas its replacement with a bulkier Ala residue leads to a higher prevalence of larger magnitude twist angles.

To assess the thermodynamic driving force for self-assembly, we report in Figure 10 the dimerization potential of mean force curves for the four systems. In all cases we observe dimerization to be spontaneous, with a free energetic driving force for dimerization of approximately $\Delta\beta F \approx 25\text{-}35$ kT. This favorable driving force is consistent with experimental observations of spontaneous self-assembly under low-pH conditions. Comparing DAVG vs. DAIA and VEVAG vs. VEVAA, we observe that the peptides possessing a Gly residue in the X_2 position immediately adjacent to the PDI core possess a $\Delta\Delta\beta F \approx 5\text{-}10$ kT more favorable dimerization free energy compared to those with a bulkier Ala residue in this position.

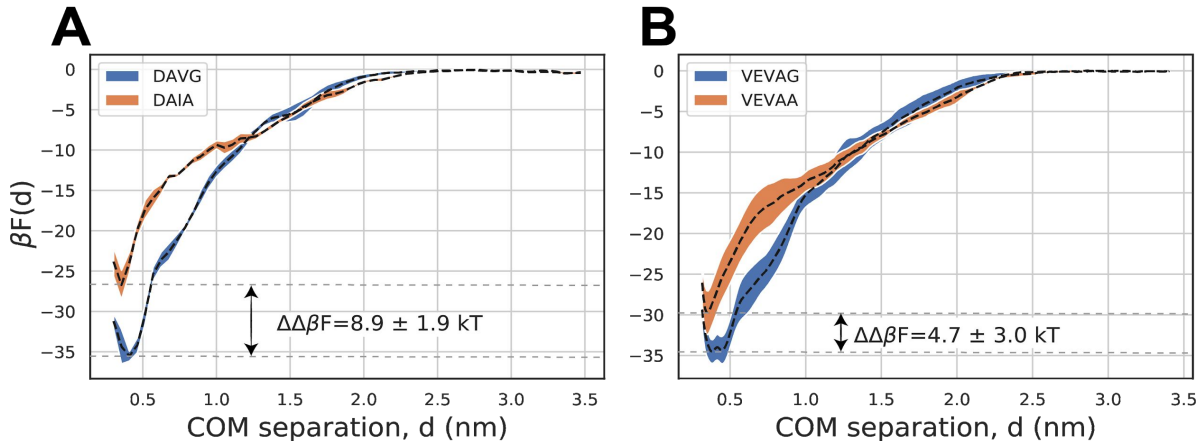


Figure 2.10: Potential of mean force (PMF) as a function of dimer aromatic core COM separation for (A) DAVG and DAIA, and (B) VEVAG and VEVAA extracted from umbrella sampling simulations. Shading represents pointwise standard errors estimated from three independently run sets of umbrella sampling simulations for each system. The arbitrary zero of free energy is defined for the non-interacting system at large COM separation, $|d|$. Annotations indicate the difference in the dimerization free energies between the two peptide systems $\Delta\Delta\beta F$. The presence of a proximate Gly residue in the X_2 position immediately adjacent to the PDI core results in a more favorable free energy of dimerization than the presence of an Ala residue in the same position.

Taken together, the higher prevalence of smaller twist angles and more favorable dimerization free energy for the DAVG/VEVAG systems compared to DAIA/VEVAA suggest that the systems possessing the smaller Gly residue in the X_2 position produce aggregates that are more cofacially aligned and more strongly bound than those possessing the bulkier Ala residue in the X_2 position that are more twisted and less strongly associated. These results are consistent with experimental UV-Vis and PL spectra wherein the DAVG/VEVAG systems exhibited Type I (i.e., J-aggregate and/or liquid-crystalline) photophysical behavior upon acidification, whereas DAIA/VEVAA exhibited Type II (i.e., weakly coupled H-aggregate) behavior (Table 1, Fig. 4,5). These calculations provide a molecular-level rationale for the two different classes of observed behavior as a function of the steric bulk of the amino acid residue in the X_2 position: bulkier residues induce weaker association and reduced cofacial packing. Figure 1 summarizes the influence of amino acid residues to the formation of self-assembled architecture of PDI-peptide conjugates.

2.5 Conclusion

We designed a series of peptide-PDI-peptide materials that shows a broad range of absorption and emission properties as effected by small changes in the composition of the amino acid sequence. This is the first example of systematic tuning of hydrophobicity and size in polypeptide-conjugated PDI molecules to attain distinct photophysical responses in aqueous solution. We established that lower hydrophobicity/size of the amino acids (i.e. glycine) closest to the PDI core drives the formation of Type-I structures which can be J-aggregates or liquid-crystalline type materials or a mixture of both. Small changes in hydrophobicity from a side chain of $-H$ to $-CH_3$ results in Type-II structures which can be interpreted as weakly coupled H-type aggregates. All-atom molecular dynamics simulations provide molecular-level corroboration that this small change in the hydrophobic side-chain of the core-adjacent amino acid residue can substantially alter the aggregated structures by mod-

ulating the relative orientation between chromophores and the binding strength between neighboring peptides within the supramolecular aggregate. This study provides new insight about how different electronic couplings between PDI chromophores can be controllably engineered by modest changes in the amino acid sequence and provides a platform to exploit and study previously inaccessible PDI electronic states in aqueous solution. Using initial computational insight followed by tandem experimental and computational analysis, we identified a general structure/function relationship between amino acid sequences and photophysical properties in PDI-polypeptide conjugates that can be used in future to design new aqueous self-assembled electronic materials.

2.6 Acknowledgements

This research is based upon work supported by the National Science Foundation’s Designing Materials to Revolutionize and Engineer our Future (DMREF) program (Grant Nos. DMR-1728947 and DMR-1841807). K.S. is supported by the National Science Foundation’s Graduate Research Fellowship (Grant No. DGE-1746045). This work was completed in part with resources provided by the University of Chicago Research Computing Center. We gratefully acknowledge computing time on the University of Chicago high-performance GPU-based cyberinfrastructure (Grant No. DMR-1828629).

CHAPTER 3

PERMUTATIONALLY INVARIANT NETWORKS FOR ENHANCED SAMPLING (PINES): DISCOVERY OF MULTI-MOLECULAR AND SOLVENT-INCLUSIVE COLLECTIVE VARIABLES

The following chapter is adapted from work originally published in the *Journal of Chemical Theory and Computation*.

N.S.M. Herringer, S. Dasetty, D. Gandhi, J. Lee, and A.L. Ferguson “Permutationally Invariant Networks for Enhanced Sampling (PINES): Discovery of multi-molecular and solvent-inclusive collective variables” J. Chem. Theory Comput. 20 178-198 (2024) [[https://doi.org/ 10.1021/acs.jctc.3c00923](https://doi.org/10.1021/acs.jctc.3c00923)]

3.1 Abstract

The typically rugged nature of molecular free energy landscapes can frustrate efficient sampling of the thermodynamically relevant phase space due to the presence of high free energy barriers. Enhanced sampling techniques can improve phase space exploration by accelerating sampling along particular collective variables (CVs). A number of techniques exist for data-driven discovery of CVs parameterizing the important large scale motions of the system. A challenge to CV discovery is learning CVs invariant to symmetries of the molecular system, frequently rigid translation, rigid rotation, and permutational relabeling of identical particles. Of these, permutational invariance has proved a persistent challenge in frustrating the data-driven discovery of multi-molecular CVs in systems of self-assembling particles and solvent-inclusive CVs for solvated systems. In this work, we integrate Permutation Invari-

ant Vector (PIV) featurizations with autoencoding neural networks to learn nonlinear CVs invariant to translation, rotation, and permutation, and perform interleaved rounds of CV discovery and enhanced sampling to iteratively expand sampling of configurational phase space and obtain converged CVs and free energy landscapes. We demonstrate the Permutationally Invariant Network for Enhanced Sampling (PINES) approach in applications to the self-assembly of a 13-atom Argon cluster, association/dissociation of a NaCl ion pair in water, and hydrophobic collapse of a $C_{45}H_{92}$ *n*-pentatetracontane polymer chain. We make the approach freely available as a new module within the PLUMED2 enhanced sampling libraries.

3.2 Introduction

Classical molecular dynamics (MD) simulations present a powerful technique to understand and engineer the molecular-level behaviors of chemical, materials, and biological systems.[88, 89] Rugged free energy landscapes are a generic feature of many molecular systems [90] wherein high free energy barriers several multiples or more of the thermal energy can trap unbiased simulation trajectories and result in incomplete sampling of the thermally accessible phase space, failure to resolve the relevant metastable states, and unconverged thermodynamic averages [91, 92]. Collective variable (CV)-based biasing techniques are a class of non-Boltzmann sampling techniques designed to ameliorate the sampling problem by accelerating barrier crossing along particular CVs [93]. A panoply of CV biasing enhanced sampling techniques exist [94, 10, 91, 92], including, for example, umbrella sampling[95], adaptive biasing force[96], and metadynamics[97]. The techniques differ in the numerical details of how the biasing forces are applied, but the success of all of these techniques relies upon the identification of appropriate CVs along which to drive sampling. Robert Zwanzig’s observation regarding the projection operator formalism is apposite [98]: “Statistical mechanics does not tell us what the relevant variables are. This is our choice. If we choose

well, the results may be useful; if we choose badly, the results (while still formally correct) will probably be useless.” Generally, we wish to select CVs coincident with the large scale and/or slow motions of the system so that we can enhance sampling along the conformational degrees of freedom important to the system dynamics. We also wish to judiciously select a parsimonious number of CVs so as to enable efficient sampling and the opportunity for thermodynamically converged sampling within the CV subspace. CV discovery is fundamentally a problem of dimensionality reduction in the space of the Cartesian coordinates defining the configurational phase space of the system, and these objective functions – high-variance CVs for large scale motions or maximally autocorrelated CVs for slow motions – underpin the preponderance of systematic CV discovery techniques using variants of dimensionality reduction algorithms [99, 10, 94].

Autoencoders – artificial neural networks (ANNs) comprising an encoding of high-dimensional observations into a low-dimensional latent bottleneck followed by decoding of the latent space back up to the high-dimensional observational space – have proven particularly well-suited to CV discovery applications [100, 101]. As universal function approximators capable of unsupervised learning and automatic differentiation, these architectures can be trained over unlabeled simulation trajectories to discover nonlinear CVs that are explicit and differentiable functions of the atomic coordinates [94]. These favorable properties mean that it is not necessary to know the metastable states of the system *a priori* (i.e., unlabeled trajectory data is sufficient) and the differentiable CVs can be modularly incorporated into standard enhanced sampling algorithms by propagating biasing forces on the CVs to those on the atoms using the chain rule [10, 9]. In terms of high-variance CV discovery, we previously introduced Molecular Enhanced Sampling with Autoencoders (MESA) as an approach that interleaves successive rounds of CV learning and biased sampling to simultaneously discover CVs coincident with the large-scale dynamical motions of the system and enhance sampling along these coordinates [10, 9]. The process is iterated until we achieve

convergence in phase space exploration and a set of CVs that globally parameterize the thermally relevant configurational subspace. Tiwary and co-workers developed reweighted autoencoded variational Bayes for enhanced sampling (RAVE)[11] as an approach similar to MESA that focusses on the discovery of a single CV that is highly correlated with a physically interpretable order parameter. Belkacemi, Gkeka, Lelièvre, and Stoltz proposed free energy biasing and iterative learning with autoencoders (FEBILAE) as a technique very closely related to MESA but which reweights the biased simulation data prior to each new round of CV discovery to eliminate the effect of non-Boltzmann sampling in the learned CVs [102]. Ensing and co-workers developed the FABULOUS approach that shares similarities to both MESA and RAVE as well as earlier work by Ma and Dinner [103, 104]. This approach essentially employs only the decoding block of an autoencoder to identify CVs capable of accurate reconstruction of atomic coordinates. Lubber and co-workers developed deep autoencoder neural network (DAENN)[105] and a framework called Deep Learning for Collective Variables (DeepCV)[106] implementing it to iteratively learn CVs of chemical reactions from limited training configurations covering only the reactant state using eXtended Social Permutation Invariant (xSPRINT) molecular representation. In terms of slow CV discovery, a number of unsupervised autoencoder-based architectures have also found success, including the time-lagged autoencoders (TAE) of Wehmeyer and Noé [107], variational dynamics encoders (VDE) of Pande and co-workers [108, 109, 110], state predictive information bottleneck (SPIB) of Wang and Tiwary [111], and our state-free reversible VAMPNets (SRVs) [112, 113, 114] also known as Deep-TICA [115]. In a supervised learning modality, Bolhuis and co-workers employed augmented/extended autoencoders that also take in labeled information on the molecular committor function to learn CVs capable of accurate reconstruction of both the atomic coordinates and progression along a reactive path [116].

An important consideration in CV discovery is the determination of symmetry-adapted CVs that respect the underlying symmetries of the molecular system [117, 118]. The most

commonly encountered symmetries are translation, rotation, and permutation arising from the invariance of a molecular system to rigid translation and/or rotation and the permutational relabeling of identical and indistinguishable particles. Failure to respect these symmetries in the learned CVs is at best data inefficient and at worst can lead to fundamental failure modes associated with the artificial distinction of physically identical system configurations. In general, CVs for enhanced sampling should be invariant to these symmetries, although there are cases where covariance may be desirable [119, 120, 121]. Symmetry adaptation is typically enforced by the choice of featurization of the system and/or within the neural network architecture. The learned CVs are relatively easily made invariant to translation and rotation by representing the system using internal coordinates (i.e., bonds, angles, torsions) that do not depend on the arbitrary location and orientation of the Cartesian coordinate system [117, 122, 123] or by data augmentation [10, 124].

Permutational invariance has proven more challenging to eliminate, and this has frustrated CV discovery and biased sampling in common and important applications such as self-assembling systems of identical particles or solutes immersed in a bath of identical solvent molecules. Good methods do exist, however, to handle permutational invariance. At the level of the network architecture, Winter, Noé, and Clevert developed a permutation-invariant graph autoencoder (PIGAE) for graph reconstruction by employing an explicit permuter that learns an explicit permutation matrix for each input graph, although the quartic scaling in reordering operations during each training step makes this approach expensive [125]. Similarly, Huang and co-workers [126] developed a GraphVAMPnets based approach that utilizes graph neural networks to respect permutation and rotational symmetries of particles in self-assembling systems by enforcing identical node embeddings for such symmetric particles and learn the slow CVs from the resulting graph embeddings using VAMPnets. Prudente, Acioli, and Soares Neto [127] and Nguyen and Le [128] developed special purpose neural networks for the fitting of potential energy surfaces of small molecules

in which the first layer of the network was modified to respect permutational invariance, although generalization and scaling to larger systems is challenging. A cheaper and more scalable approach is to represent the system to the network via permutationally-invariant featurizations [117]. Provided these featurizations can be expressed as explicit functions of the particle coordinates, the derivatives of the CVs learned by the neural network with respect to the Cartesian coordinates can be constructed from the chain rule via automatic differentiation and/or analytical derivatives. Permutational invariance is typically enforced by summing particle-centric basis functions over all identical particles or a permutationally-invariant ordering. Examples of the former class include permutation-invariant polynomials (PIPs) generated by summing monomial functions of internal coordinates over all like-particle orderings [129, 130], atom-centered symmetry functions (ACSF) that achieve permutational invariance by a similar summation procedure [131, 132], many-body tensor representations (MBTR) that define scalar geometry functions (e.g., atomic number, distance, angle) over small body-order numbers of atoms and arranges these into permutationally-invariant distribution functions for each class of interaction [133], solvent signature vectors that use Gaussian kernel functions to quantify the distribution of solvent molecules around the solute [134], solvent shell featurizations that quantify the number of water molecules in nested shells around the solute [135, 136], and smooth overlap of atomic orbitals (SOAP) that generates the partial power spectra of an expansion of spherical harmonic representations of the atomic-centered smoothed density fields summed over all identical atoms [137]. In general, the exponential scaling in the number of permutational orderings associated with the summations for each of these approaches must be controlled by defining a cutoff in the range and/or order of the interactions considered [117]. Ordering approaches typically offer a computationally cheaper means to achieve permutational symmetry adaptation [117]. Examples of this class include the bag-of-bonds (BoB) featurization based on a concatenation of the ordered (inverse) pairwise distances for each class of bond scaled by the product of the nuclear charges

[138] and social permutation invariant (SPRINT) coordinates constructed from the leading eigenvector of a graph adjacency matrix [139, 140]. Of particular interest in this work is the Permutation Invariant Vectors (PIV) featurization developed by Pietrucci and co-workers [141, 142, 140]. This representation can be conceived of as a concatenation of instantaneous radial distribution functions for classes of atom pairs in the system that, by virtue of distance-based ordering, is invariant to like-particle permutation and, through the use of exclusively pairwise distances, also translation and rotation. Physically motivated, bespoke order parameters engineered to preserve permutational invariance have also been developed, although these can be less generically extensible to arbitrary molecular systems. A trivial strategy to treat the permutational invariance of a solvent bath is to simply ignore the solvent coordinates and assume that the solvent effects are sufficiently strongly imprinted on the solute conformational dynamics that they can be driven implicitly by biasing solute-centric CVs [143]. Patel, Varilly, Chandler, and Garde developed the indirect umbrella sampling (INDUS) approach to drive water density fluctuations in user-defined control volumes and have used this technique to drive sampling of wetting/dewetting transitions [144, 1]. Zou *et al.* employed CVs based on coordination numbers, Steinhardt bond order parameters, interfacial water density, intermolecular angles, and pair orientational entropy, and combined these within the SPIB framework to drive sampling of crystal nucleation and polymorph formation [145]. Rizzi *et al.* combined CVs describing solvent coordination number and solvent structure around a binding pocket with the deep linear discriminant analysis (Deep-LDA) approach to incorporate water coordinates in driving sampling of host-guest interactions [146].

The primary contribution of this work is to present an approach for the integrated data-driven discovery of translationally, rotationally, and permutationally-invariant CVs in systems containing identical particles and their use within enhanced sampling calculations to accelerate phase space exploration. We term this approach Permutationally Invariant Net-

works for Enhanced Sampling (PINES). Our approach employs a PIV representation as input for an autoencoder to perform interleaved nonlinear CV discovery and CV biasing enhanced sampling using parallel bias metadynamics [147]. We draw inspiration for the synthesis of permutationally invariant featurizations with ANNs from the deep potential molecular dynamics (DeePMD) approach of Zhang *et al.* that employs an ordered distance featurization passed to an ANN to learn a potential energy function for molecular dynamics [123] and the PIP-NN approach of Jiang and Guo that performs a similar feat using a permutation-invariant polynomial (PIP) featurization [148]. We eschew the use of physically-motivated, hand-crafted descriptors in favor of a more generic, transferable, and system agnostic featurization strategy, but observe that the incorporation of system-specific knowledge may be straightforwardly achieved by augmenting the descriptor list with any additional features provided they are explicit and differentiable functions of the particle Cartesian coordinates. The PINES paradigm is also modular in the sense that different featurizations may be employed to adapt to different symmetries of the system (e.g., translation, rotation, permutation, inversion, chiral) and different ANNs may be used to perform different learning tasks (e.g., unsupervised CV discovery, supervised learning of potential energy surfaces). We demonstrate PINES in applications to three molecular systems in which proper accounting for particle permutational invariance is of paramount concern: (i) self-assembly of a 13-atom *Ar* cluster, (ii) assembly/disassembly of a *NaCl* ion pair in water, and (iii) hydrophobic collapse of a $C_{45}H_{92}$ *n*-pentatetracontane polymer chain. We make the approach freely available as a new user-friendly module at <https://github.com/Ferg-Lab/pines.git> that can be patched with PLUMED2 [8] enhanced sampling libraries.

The structure of the remainder of this paper is as follows. In Sec. 4.3, we introduce the PINES method for translationally, rotationally, and permutationally-invariant CV discovery and enhanced sampling using parallel bias metadynamics. This is followed by a description of the PIV featurization, PINES training procedure, and the enhanced sampling molecular

dynamics simulations. In Sec. 4.4, we demonstrate applications of PINES in discovering CVs for enhanced sampling to self-assembly of 13-atom *Ar* cluster, association/disassociation of *NaCl* ion pair in water, and hydrophobic collapse of $C_{45}H_{92}$ polymer chain, and compare their free energy landscapes to prior work. Finally, in Sec. 4.5 we present our conclusions and perspectives for future work.

3.3 Methods

3.3.1 *Molecular Enhanced Sampling with Autoencoders (MESA)*

PINES can be conceived as a permutationally-invariant extension to the Molecular Enhanced Sampling with Autoencoders (MESA) approach that we previously developed for data-driven CV discovery and enhanced sampling [10, 9]. The key innovation of this approach is to use autoencoding ANNs to perform unsupervised, nonlinear dimensionality reduction of high-dimensional molecular simulation trajectories to discover low-dimensional CV parameterizations spanning the high-variance collective motions of the system. By virtue of the ANN architecture and automatic differentiation through the network, the functional dependence upon the first derivatives with respect to the Cartesian coordinates of the particles in the system are explicitly available, which enables propagation of biases in the CVs to forces on atoms and permits the learned CVs to be seamlessly incorporated into any off-the-shelf CV biasing enhanced sampling approach. The “chicken-and-egg problem” of data-driven CV discovery observes that CV discovery requires training trajectories spanning the thermally relevant phase space, and trajectories spanning the thermally relevant phase space require good CVs to drive sampling [149]. MESA solves this problem by interleaving successive rounds of CV discovery and enhanced sampling until there is no additional round-to-round enhancement in exploration of phase space or change in number or character of the learned CVs. The first round of the CV discovery process may be initialized using short unbiased

trajectories, simulated annealing, or biased trajectories in intuited or physically-motivated CVs. Subsequent rounds of the process are conducted by performing enhanced sampling in data-driven CVs discovered by the ANN. A schematic illustration of the PINES approach with parallel bias metadynamics as the CV biasing enhanced sampling approach is illustrated in Fig. 3.1.

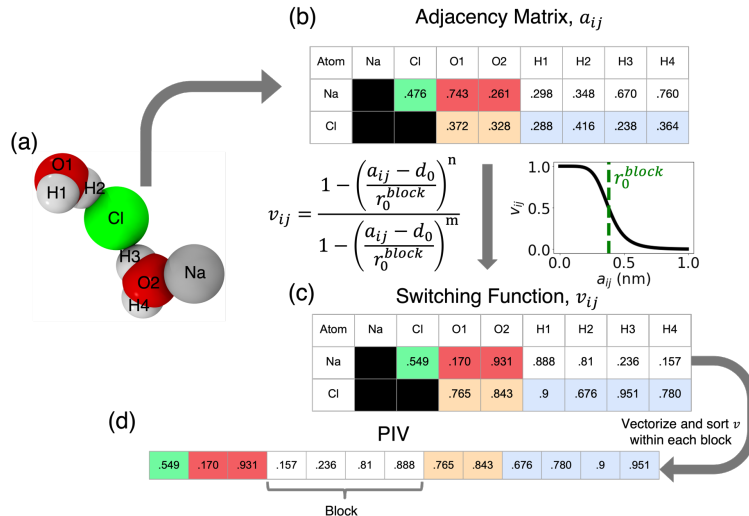


Figure 3.1: Schematic illustration of the Permutationally Invariant Network for Enhanced Sampling (PINES) approach for data-driven CV discovery and enhanced sampling. (a) Molecular simulation trajectories provide training data for CV discovery. The process is initialized with unbiased or biased simulations to provide an initial first sampling of phase space. Subsequent rounds are trained over biased simulations generated by enhanced sampling in the learned CVs from the previous round. (b) The trajectories are featurized and passed to a maximum mean discrepancy-Wasserstein autoencoder (MMD-WAE) to perform unsupervised nonlinear dimensionality reduction into a low-dimensional latent space spanned by the learned CVs. To stabilize the CV discovery process and help disentangle the latent space, we employ batch normalization between all layers with exception of the final layer. (c) After training the network, we extract the CVs from the bottleneck layer between the encoder and decoder. We find that it is useful to first pass the output of the bottleneck layer through a principal components analysis (PCA) whitening transformation to help align the discovered manifold with the highest variance directions for efficient parallel bias metadynamics (PBMetaD) sampling. The learned CVs are represented as the functional composition $\xi = h \circ f(\mathbf{r}) = h(\mathbf{z})$, where $\mathbf{z} = f(\mathbf{r})$ is the result of applying the featurization f – in this work a PIV featurization – to the Cartesian coordinates $\mathbf{r}$ of the system, and $\xi = h(\mathbf{z})$ is the action of the MMD-WAE encoder plus PCA whitening transform upon the input featurization that generates a projection into the learned CVs. The derivative of the CVs with respect to the Cartesian coordinates of atom i is required for propagation of biasing forces in the CVs to biasing forces on the atoms, and comes from simple application of the chain rule to the function composition that is computed in practice by automatic differentiation, $\nabla_{\mathbf{r}_i} \xi = (\partial h / \partial \mathbf{z}) \nabla_{\mathbf{r}_i} \mathbf{z}$. (d) The learned CVs are implemented within PBMetaD calculations to drive sampling in the configurational phase space spanned by these CVs, completing one full iteration of PINES. The biased simulation trajectories from Round n are then passed forward as the training trajectories for CV discovery within Round $(n + 1)$ to determine new CVs and conduct new enhanced sampling calculations. The iterative process is terminated when the learned CVs stabilize, we observe no further exploration of phase space, and the reweighted free energy landscape converges.

3.3.2 *Permutation Invariant Vector (PIV) Featurization*

PINES employs a Permutation Invariant Vector (PIV) [141, 142, 140] representation of the molecular system to permit discovery of translationally, rotationally, and permutationally-invariant CVs for enhanced sampling. The computational graph underpinning PINES is simply a PIV featurization of the system followed by a regularized autoencoder for data-driven CV discovery (Fig. 3.1a,b). In this sense, it is very similar to MESA but employs a PIV representation of the molecular system. Importantly, gradients can be backpropagated end-to-end through the autoencoder encoder and PIV featurizer to permit calculation of the gradients in the CVs required for enhanced sampling calculations.

The PIV featurization can be conceived as a concatenation of instantaneous pair correlation functions specifying the ordered set of pairwise distances between the chemically identical and indistinguishable atoms of each class (e.g., all water O atoms, all Na^+ ions, etc.) [141, 142, 140]. Since the featurization is constructed from pairwise distances it is naturally invariant to rigid translation and rotation, and by virtue of the inherent ordering of pairwise distances within an instantaneous pair correlation function, it is also invariant to permutations of the labels attached to the indistinguishable atoms of each class. Since the PINES network is trained to reconstruct the PIV representation of the system, the learned CVs are also inherently invariant to translation, rotation, and permutation. Importantly, the PIV representation of a system is complete in the sense that if all pairwise distances are incorporated into the PIV featurization, rigid graph theory dictates that the Cartesian coordinates of the particles constituting the system can be unambiguously reconstructed up to translation, rotation, and permutation [150, 151].

PIVs were introduced by Pietrucci and co-workers [141, 142, 140] and have been used, among other applications, to efficiently drive sampling of different phases of water [141]. We illustrate in Fig. 3.2 the construction of a PIV featurization for a $NaCl$ ion pair in water. The PIV featurization proceeds in three steps. First, we compute the distance matrix with

elements a_{ij} recording the Euclidean pairwise distances between atoms i and j . We arrange the elements of the matrix into a block structure where each block contains the set of pairwise distances between two atom classes (i.e., groups of chemically indistinguishable atoms). Within each block the ensemble of pairwise distances are unchanged under permutational relabeling of the chemically indistinguishable atoms within each class. Second, we pass the pairwise distances through a rational switching function to produce the adjacency matrix,

$$v_{ij} = \frac{1 - \left(\frac{a_{ij}-d_0}{r_0}\right)^n}{1 - \left(\frac{a_{ij}-d_0}{r_0}\right)^m} \quad (3.1)$$

where the hyperparameter d_0 is minimum interparticle distance, r_0 is the interparticle distance which lies on the midpoint of the switching function curve (Fig. 3.2b), and n and m are non-negative numbers which control the steepness of the switching function curve and thus determine the resolution of the various pairwise interactions. This transformation has the two desirable effects of scaling the distances to the interval $(0, 1)$ and producing values close to unity for short distances, values close to zero for large distances, and a switching range between these two limits over which the transformation has the most sensitivity to changes in a_{ij} . This ensures that the v_{ij} values passed as input to the encoder are appropriately scaled to $(0, 1)$ range and, by tuning the parameters of the switching function, permits us to focus the dynamic range of v_{ij} upon the physical range of distances over which we wish the network to pay the most attention. In principle, other switching functions[152] could be applied as well in place of the rational function (Eq. 1) for computing v_{ij} . Typically, we set the parameters in Eq. 1 based on a radial distribution function (RDF) collected over a short unbiased trial simulation of the system in its stable state (Fig. S1). We choose r_0 to be the mean of the first and second peaks of the RDF, set $d_0 = 0$ and $m = 2n$ and tune n such that $v_{ij} \approx 0.9$ at the a_{ij} value of the first peak and $v_{ij} \approx 0.1$ at the a_{ij} value of the second peak in RDF. This ensures that the v_{ij} values are most responsive to particles moving between

the first two shells of their radial distribution where it is typically most important to highlight structural differences between various system configurations and become insensitive in the far field [140]. When appropriately tuned in this fashion, the PIV featurization show large inter-class differences between qualitatively different system configurations and small intra-class differences for configurations in the same structural class [140]. The important dynamical ranges for the systems studied in this work were relatively narrow such that we were able to use relatively steep switching functions that give good resolution over the region of interest. For larger dynamic ranges – such as one might encounter in large or complex self-assembling systems – one would typically use a longer-ranged switching function so as not to lose resolution in the flat regions of the switch with small gradients. Such a long-ranged switching function, however, offers reduced resolving power over its range due to the gradual nature of the switch. A superior, albeit more computationally expensive, solution would be to use multiple steep switching functions for each pairwise interaction tuned to be responsive over a range of concentric distance shells, wherein each switch provides high resolution over a local range and is arranged to partially overlap with its neighboring switching functions, thereby providing complete coverage of the relevant dynamical range. Third, we sort the v_{ij} values within each block of the adjacency matrix in non-descending order and concatenate them into a single vector – the eponymous PIV. The sorting operation is the critical step in introducing permutational invariance.

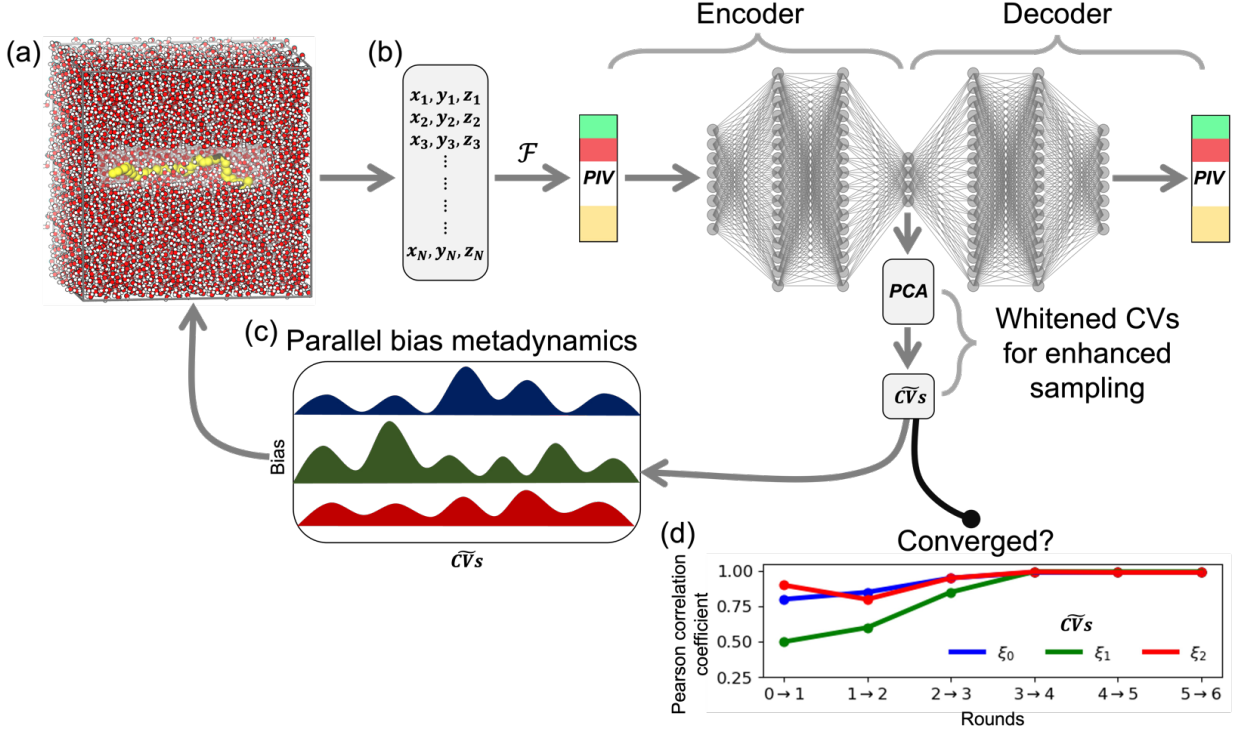


Figure 3.2: Schematic illustration of calculation of the PIV featurization of a *NaCl* ion pair in water. (a) Snapshot of an instantaneous state of the *NaCl* ion pair and the two most proximate water molecules. All molecular renderings in this work are constructed using Visual Molecular Dynamics (VMD) [2]. (b) We convert the Cartesian coordinates of the atomic system into a distance matrix recording all pairwise distances a_{ij} between all pairs of atoms i and j . The symmetry of the distance matrix means we need only record the upper triangle and is arranged into a block structure by atom class such that indistinguishable atoms are grouped together. Since the number of pairwise distances increases quadratically with the number of atoms, we may also choose to restrict the atoms over which we construct the distance matrix by, for example, omitting all but the alpha-carbons in a protein or disregarding the far-field solvent molecules around a solute. In the interests of space, we present here the abridged distance matrix where we have colored each block or group of (indistinguishable) pairwise distances: Na-Cl (green), Na-O (red), Na-H (white), Cl-O (orange), and Cl-H (blue). (c) The elements of the distance matrix a_{ij} are transformed by a rational switching function to produce the adjacency matrix with elements v_{ij} . The parameters of the switching function are tuned to present the largest dynamic range and sensitivity to configurational changes in the short- to medium-ranged field. It also has the desirable property of scaling the pairwise distances to a (0, 1) range. We choose r_0 in each block (r_0^{block}) to be the mean of the first and second peaks of the RDF, set $d_0 = 0$ and $m = 2n$, and tune n such that $v_{ij} \approx 0.9$ at the a_{ij} value of the first peak and $v_{ij} \approx 0.1$ at the a_{ij} value of the second peak in the RDF. (d) The elements within each block of the adjacency matrix are placed in non-descending order and concatenated to form PIV.

Importantly, construction of the PIV from the Cartesian coordinates of the system is a deterministic and differentiable operation that permits backpropagation of gradients from the learned CVs to the PIV featurization and then the Cartesian coordinates by automatic differentiation. A well known deficiency of PIVs, and one that is shared by all featurizations that achieve permutational invariance by ordering, is that the featurization can exhibit kinks and its derivative therefore possess discontinuities at locations in the configurational phase space where the ordering of the particles changes [117]. Empirically, we encountered no difficulties or instabilities in either our CV learning process or our enhanced sampling calculations in the learned CVs. It is possible that the discontinuities are sufficiently gentle and/or rare that they do not present any difficulties, but this issue may warrant additional study to identify situations where these singularities may present difficulties.

3.3.3 Autoencoder network design and training

The second component of PINES is an autoencoding neural network that performs unsupervised nonlinear dimensionality reduction of molecular simulation training trajectories to discover low-dimensional embeddings into a small number of high-variance CVs [101, 100, 10, 9]. In this work, we employ a maximum mean discrepancy-Wasserstein autoencoder (MMD-WAE)[153]. The network is trained by minimizing the loss function comprising a mean squared error (MSE) term to encourage high fidelity reconstruction and a maximum mean discrepancy (MMD) loss to regularize the latent space,

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (PIV_i - \widehat{PIV}_i)^2 + E_P[k(x, x')] + E_Q[k(y, y')] - 2E_{P,Q}[k(x, y)], \quad (3.2)$$

where $k(z, z') = \exp(z - z')^2$ is a Gaussian kernel with unit bandwidth, $\frac{1}{N} \sum_{i=1}^N (PIV_i - \widehat{PIV}_i)^2$ is the reconstruction’s element-wise mean squared error, $E_P[k(x, x')] + E_Q[k(y, y')] - 2E_{P,Q}[k(x, y)]$ is the MMD loss, $P(x)$ is the probability distribution of the latent space, and

$Q(y)$ is the target latent space distribution. The MMD loss is equal to zero only when the two distributions are identical. In this work, we choose the target distribution to be a normal distribution with a zero mean and unit variance.

We note that in principle it may be advantageous to sort the elements within each block of the reconstructed PIV into non-descending order prior to calculating the MSE loss. In this way the decoder need only learn the distribution of values within each block instead of also being required to learn the sorting rule. Empirically, we find that the decoder is able to readily learn to produce values in non-descending order and only rarely violates proper sorting within the reconstructed PIV. It appears, therefore, that pre-loss sorting of the reconstructed PIV is unnecessary. Additionally we draw attention to the possibility of using non-uniform weights over the various contributions to the MSE term in order to modulate the importance of reconstruction fidelity for particular interactions in the system [140], although in our work we observed satisfactory performance using uniform weighting.

The encoder and decoder are both implemented as fully-connected feedforward neural networks with three hidden layers (Fig. 3.1b). For all systems considered in this work, the overall network architecture is M -64-32-16- d -16-32-64- M , where M is the dimensionality of the PIV featurization and d is the number of nodes in the bottleneck layer that defines the number of CVs to be discovered. An appropriate value for d is determined each round by training networks with d ranging from 1-8 and selecting the appropriate value of d from a knee in a plot of the fraction of variance explained curve using the L-method of Salvador and Chan [10, 9, 154] (Fig. S2). For the systems studied in this work, we find a consistent latent space dimensionality of $d=3$ over the PINES iterations. For more complicated systems where the first round of sampling was insufficient to expose all of the important collective modes, these may emerge gradually over the PINES iterations and be manifest as an increase in the latent space dimensionality exposed by this analysis. Biomolecular systems are archetypal complex systems in the sense that they tend to evince an emergent large-scale low-dimensionality

due to nonlinear many-body couplings between the atomic degrees of freedom [155, 143, 156, 157, 158, 159, 160]. Nevertheless, the enhanced sampling technique selected for this work, parallel bias metadynamics (cf. Section 2.1.4), scales well to quite large numbers of CVs [147, 161] thereby enabling the application of PINES to systems with high-dimensional latent spaces. We employ *tanh* activation functions between all hidden layers and a sigmoid activation in the final layer of the decoder. Batch normalization is employed between all hidden layers except the layer preceding the reconstructed PIV, which we find to be critical in avoiding saturation and stabilizing CV discovery. We construct and train the network using PyTorch[162] with PyTorch Lightning[163] and train it using backpropagation employing the Adam optimizer[164] with a batch size of 100 and learning rate of 10^{-4} . Training was terminated when the validation loss stopped decreasing employing a patience of 10 epochs. Training was performed on a NVIDIA Quadro RTX 6000 GPU card and, for the systems studied in this work, typically required hundreds of epochs and tens of GPU-minutes. After training is complete, we discard the decoder and pass the output of the bottleneck layer through a principal components analysis (PCA) whitening transformation (Fig. 3.1b) [165]. We find this step useful in disentangling the learned CVs by imposing ordering by variance that helps to both stabilize the round-to-round discovery process and also improves enhanced sampling by orienting the biasing potentials along the orthogonalized PCA subspace. This transformation can be thought of as a simple change of basis in which the dimensionality of the latent space is conserved and the whitened CVs represent a rotation of the manifold to align the axes with the highest variance regions of the embedding.

3.3.4 *Enhanced sampling along discovered CVs using parallel bias metadynamics*

The learned CVs within each round are used to conduct enhanced sampling of configurational space using parallel bias metadynamics (PBMetaD)[147]. The PINES approach is compatible

with any off-the-shelf CV biasing enhanced sampling technique, but we elect to use PBMetaD due to its efficient scaling to large numbers of CVs by the application of multiple simultaneous one-dimensional biases (Fig. 3.1c).

In this work, we never have need to drive sampling of more than three CVs, but the PBMetaD approach has been demonstrated up to several tens of CVs [161]. We couple all learned CVs to a well-tempered variant of PBMetaD enhanced sampling calculation implemented within the PLUMED2 library [8, 166, 147]. The hyperparameters for the PBMetaD calculations are detailed for each of the three systems in Section 3.3.6.

Convergence of the PBMetaD calculations are assessed using two criteria [167]. First, we check that the biasing potential has become quasi-static by assuring that the height of the Gaussian biases deposited over the course of the simulation drop below a particular fraction of the initial height. In early iterations we adopt a less stringent criterion of 10% but in later rounds, we refine this to a more strict criterion of 1%. Second, we monitor the distribution of the CVs and check for diffusive sampling over the full range explored by each CV to verify that the simulation can freely diffuse over the barriers in all coupled coordinates.

3.3.5 *Convergence assessment*

Convergence of PINES is assessed by measuring the dimensionality d and cosine similarity ρ of the CVs learned between successive iterations and the expansion of the explored phase space. Stabilization of d of the MMD-WAE bottleneck layer indicates that no additional dimensions are being explored in successive biasing rounds, a cosine similarity of $\rho > 0.9$ between the CVs discovered in two consecutive rounds – computed by projecting the present round’s simulation data into the CVs learned in the present and previous rounds – is used to assess whether the CVs themselves are approximately constant between successive rounds, and we verify that the boundary of explored phase space is no longer expanding (Fig. 3.1d).

Once all convergence criteria are satisfied, we declare the PINES procedure to be com-

plete. At this point we construct a training data set by collating the biased trajectories from all prior rounds and use this concatenated data set to run a final round of CV discovery and PBMetaD enhanced sampling. For large datasets, we perform uniform sampling across the multidimensional CV space used for the latest round of biasing to generate a smaller, subsampled dataset that is representative of the whole CV space but more amenable to training a terminal network to identify the optimal embedding. These final CVs are used to perform a PBMetaD production run to demonstrate their impact on accelerating exploration of phase space. The unbiased free energy surface is recovered from the terminal PBMetaD biased trajectories by applying the standard Torrie-Valleau sampling weights to each configuration to eliminate the effect of the applied bias and enable projection into arbitrary CVs beyond those in which the biased sampling was conducted [147, 95]. Empirically, we do not observe any significant benefit from applying this reweighting to the biased trajectories in each individual CV discovery round, but this, of course, is possible [102].

3.3.6 *All-atom molecular dynamics simulations*

All-atom molecular dynamics simulations of each of the three systems considered in this work were performed using the GROMACS 2021.1 simulation suite [7] patched with an in-house modified version of the PLUMED2 v2.8.0 enhanced sampling libraries [8]. This patch implements the PINES approach by integrating the newly developed module named PINES and the ANN module previously contributed by Chen and Ferguson[10]. We developed the PINES module in PLUMED2 using the PIV module previously contributed by Pipolo and Pietrucci[142, 141] as a template. In the ANN module, we added support for batch normalization. Biasing of the translationally, rotationally, and permutationally-invariant CVs discovered by PINES can be performed by any off-the-shelf CV biasing enhanced sampling technique. In this work, we use the PBMetaD PLUMED2 module contributed by Bonomi and Pfaendtner[147] in PLUMED2 for enhanced sampling calculations. The new PINES and

modified ANN modules along with a helper bash script for patching with PLUMED2 are available at <https://github.com/Ferg-Lab/pines.git>, and will be submitted for incorporation into a future release of PLUMED2[8]. All the data and PLUMED input files required to perform the molecular dynamics simulations of the three systems have been hosted on PLUMED-NEST, the public repository of the PLUMED consortium [168], under project ID plumID:23.024 (<https://www.plumed-nest.org/eggs/23/024/>).

3.3.7 13-particle Ar cluster

We consider a system of 13 Ar atoms to demonstrate the utility of PINES in driving assembly and disassembly of a simple Lennard-Jones cluster of identical particles[169]. The Ar atoms were modeled in vacuum using Lennard-Jones non-bonded interactions with $\sigma = 0.34$ nm and $\epsilon = 1$ kJ mol⁻¹. [170, 171]. Particles were randomly placed in a 2.5×2.5×2.5 nm³ cubic box implementing periodic boundary conditions. The system was subjected to steepest descent energy minimization to eliminate forces larger than 1 kJ mol⁻¹ nm⁻¹. After energy minimization, a short equilibration simulation was performed in a NVT ensemble for 100 ps. The classical mechanical equations of motion were integrated using the leapfrog algorithm[172] with a 2 fs time step. Particle velocities were initialized according to the Maxwell-Boltzmann distribution at 50 K. Simulations were conducted in the NVT ensemble in which temperature is maintained using a velocity-rescaling thermostat[173] employing a time constant of 0.1 ps. Lennard-Jones interactions were smoothly shifted to zero at a cutoff of 1.0 nm. The unbiased production run for the initial PINES round was conducted for 25 ns and frames were saved at a frequency of 1 ps for training the autoencoder in the subsequent PINES round. We conducted 25 ns of PBMetaD simulations in subsequent PINES iterations at a temperature of 50 K employing an initial Gaussian height of $W = 0.5$ kJ mol⁻¹, a Gaussian width of $\sigma = 0.1$ for each discovered CV, a bias factor of $\gamma = 10$, and deposition pace of 500 steps (1 ps). The PIV featurization contained one Ar-Ar interaction block with

78 elements corresponding to all possible pairwise distances. Simulations were performed on 5×2.40 GHz Intel Xeon Gold 6148 CPU cores and $1 \times$ NVIDIA TITAN V GPU, achieving execution speeds of ~ 120 ns per day.

3.3.8 *NaCl ion pair in water*

An NaCl ion pair in water was used as a classic test system to assess the capacity of PINES to discover solvent-inclusive CVs governing the association/dissociation of the ion pair and use these to efficiently accelerate sampling[174, 175, 176, 3, 177] We utilized the Amber99SB-ILDN forcefield[62] with Joung-Cheatham[178, 179] ion parameters. The *Na+* and *Cl-* ions were solvated to a density of 1 g cm^{-3} with 508 TIP3P[180] water molecules in a cubic simulation box of size $2.5 \times 2.5 \times 2.5 \text{ nm}^3$. A cutoff of 1 nm was applied for the Lennard-Jones interactions. Long-range electrostatics were treated by particle mesh Ewald summation[181, 70] employing a real-space cutoff of 1.0 and 0.16 nm Fourier grid spacing that were optimized during runtime. The neighbor list was generated using the Verlet method and was updated every 10 steps (20 fs). An energy minimization procedure with steepest-descent algorithm was applied to relax the system after solvation by adjusting the configuration to have forces between particles with less than $1 \text{ kJ mol}^{-1} \text{ nm}^{-1}$. After energy minimization, a short equilibration simulation was performed in a NVT ensemble for 100 ps. The leapfrog algorithm[172] with a timestep of 2 fs was applied for propagating the equations of motion in the simulations. The initial velocities for the equilibration run were generated using the Maxwell-Boltzmann distribution at 300 K. A velocity-rescale thermostat[173] was applied for controlling the temperature during the equilibration run. For the production run, the final configuration and velocities from the equilibration run were used to initiate the simulation. Frames at a frequency of 2 ps were saved from the production run for training the autoencoder in the subsequent PINES rounds. Our PIV vectors consider the distance between the ion pair and the ion distances to the O and H atoms of the 10 most proximate water molecules such that

the PIV featurization contained five interaction blocks with a total of 61 elements distributed across the blocks as 1 Na-Cl, 10 Na-O, 20 Na-H, 10 Cl-O, and 20 Cl-H elements.

The initial production unbiased simulation was performed for 50 ns. For PBMetaD in subsequent rounds, we set the initial gaussian height as $W = 1.2$ kJ/mol, a Gaussian width of $\sigma = 0.1$ for each of the discovered CVs, a bias factor of $\gamma = 20$, and a deposition pace of 500 timesteps (1 ps). Each simulation in the PINES rounds with enhanced sampling along discovered CVs using PBMetaD was initially performed for 15 ns in a NVT ensemble at 300 K but was extended until all the convergence metrics for PBMetaD were satisfied. Simulations were performed on 5×2.40 GHz Intel Xeon Gold 6148 CPU cores and $1 \times$ NVIDIA TITAN V GPU, achieving execution speeds of ~ 50 ns per day.

3.3.9 Aqueously solvated *n*-pentatetracontane ($C_{45}H_{92}$) polymer chain

An aqueously solvated *n*-pentatetracontane ($C_{45}H_{92}$) polymer chain was used to demonstrate the ability of PINES to learn solvent-inclusive CVs for hydrophobic collapse and use these CVs to enhance sampling of the elongation/collapse transition[1]. The initial structure of the *n*-pentatetracontane alkane chain was modeled using the TraPPE-UA[182, 183] potential that models the chain as a string of *CH3/CH2* united atoms. The system was solvated with 11,188 SPCE[184] water molecules in a cubic simulation box of size $7.0 \times 7.0 \times 7.0$ nm³. A cutoff of 1 nm was applied for the Lennard-Jones interactions. Long-range electrostatic interactions are handled by particle mesh Ewald method [181, 70] employing a real-space cutoff of 1.0 and 0.16 nm Fourier grid spacing that were optimized during runtime. The neighbor list was generated using the Verlet method[185, 186] and is updated every 20 steps (40 fs). An energy minimization procedure with steepest-descent algorithm was applied to relax the system after solvation by adjusting the configuration to have forces between particles with less than $1 \text{ kJ mol}^{-1} \text{ nm}^{-1}$. After energy minimization, a short equilibration simulation was performed in a NVT ensemble for 100 ps. The leapfrog algorithm[172] with

a timestep of 2 fs was applied for propagating the equations of motion. The initial velocities for the equilibration run were generated using the Maxwell-Boltzmann distribution at 300 K. A velocity-rescale thermostat[173] was applied for controlling the temperature during the equilibration run. For the production run, the final configuration and velocities from the equilibration run were used to initiate the simulation. The unbiased production run in the initial PINES round was conducted for 100 ns with a timestep of 2 fs. Frames at a frequency of 2 ps were saved from the production run for training the autoencoder in the subsequent PINES rounds. The PIV featurization contained 45 interaction blocks with a total of 108 elements distributed across the blocks. The blocks are characterized by the intramolecular interactions between nine evenly spaced *CH2* united atoms from position 3 to 43 in the chain, yielding 36 blocks each containing only a single element, as well as the intermolecular interactions with their respective eight closest water oxygens, yielding nine blocks each containing eight elements. We note that the *CH2-CH2* united atom blocks could be represented as 18 blocks of two elements to mod out the symmetry associated with the head-to-tail inversion symmetry of the alkane chain, however we choose not to do this here.

For the PBMetaD calculations, we set the initial gaussian height as $W = 1.2 \text{ kJ mol}^{-1}$, a Gaussian width of $\sigma = 0.1$ for each discovered CV, a bias factor of $\gamma = 20$, and a deposition pace of 500 timesteps (1 ps). Each simulation in the PINES rounds with enhanced sampling along discovered CVs using PBMetaD was initially performed for 30 ns in a NVT ensemble at 300 K but was extended until all the convergence metrics for PBMetaD were satisfied. Simulations were performed on $5 \times 2.40 \text{ GHz}$ Intel Xeon Gold 6148 CPU cores and $1 \times \text{NVIDIA TITAN V GPU}$, achieving execution speeds of $\sim 10 \text{ ns per day}$.

3.4 Results and Discussion

We demonstrate the general utility and applicability of PINES to discover translationally, rotationally, and permutationally invariant CVs for enhanced sampling in applications to

three molecular systems: (i) self-assembly of a 13-atom *Ar* cluster, (ii) assembly/disassembly of a *NaCl* ion pair in water, and (iii) hydrophobic collapse of a $C_{45}H_{92}$ *n*-pentatetracontane polymer chain.

3.4.1 *Self-assembly of a 13-particle Ar cluster*

We first demonstrate the application of PINES to the self-assembly of a 13-particle *Ar* cluster in vacuum. We hypothesize that PINES can learn permutationally-invariant multi-body collective variables associated with the assembly and disassembly of the identical *Ar* atoms between solid-like, liquid-like, and vapor-like phases. After some trial-and-improvement exploration, we chose to perform the simulations at a temperature of 50 K, which we observed to provide good unbiased sampling of the solid, liquid, and vapor-like phases and therefore access to an unbiased ground truth against which to compare the biased sampling calculations under PINES. As in our previous work with MESA, we determine the dimensionality of the latent space by training independent networks with 1-8 nodes in the bottleneck layer of autoencoder and utilizing the L-method by Salvador and Chan[154] to identify a knee in the fraction of variance explained (FVE) curve (Fig. S2). With this method, we identified a latent space of dimensionality $d=3$.

In Fig. 3.3a we show the round-by-round sampling as 2D projections along the three CVs ξ_0 , ξ_1 , and ξ_2 learned from the terminal PINES network. The initial unbiased simulation used to seed PINES (gray) provided a good initial coverage of the accessible configurational space and each subsequent biased round (different colors) extends the frontier into new unsampled regions of phase space around the periphery of the explored region and probes higher free energy configurations. We present in Fig. 3.3b the reweighted (i.e., unbiased) free energy surfaces (FES) associated with each sampling round that better expose the variations in sampling density over the three CVs. In particular, we observe that the biased sampling identifies a low- ξ_0 metastable minimum that was not resolved by the unbiased calculations

in Round 0. As we discuss below, this minimum at $(\xi_0 \approx -2, \xi_1 \approx 0, \xi_2 \approx 1)$ corresponds to a rarefied vapor-like state that is incompletely explored in the unbiased calculations.

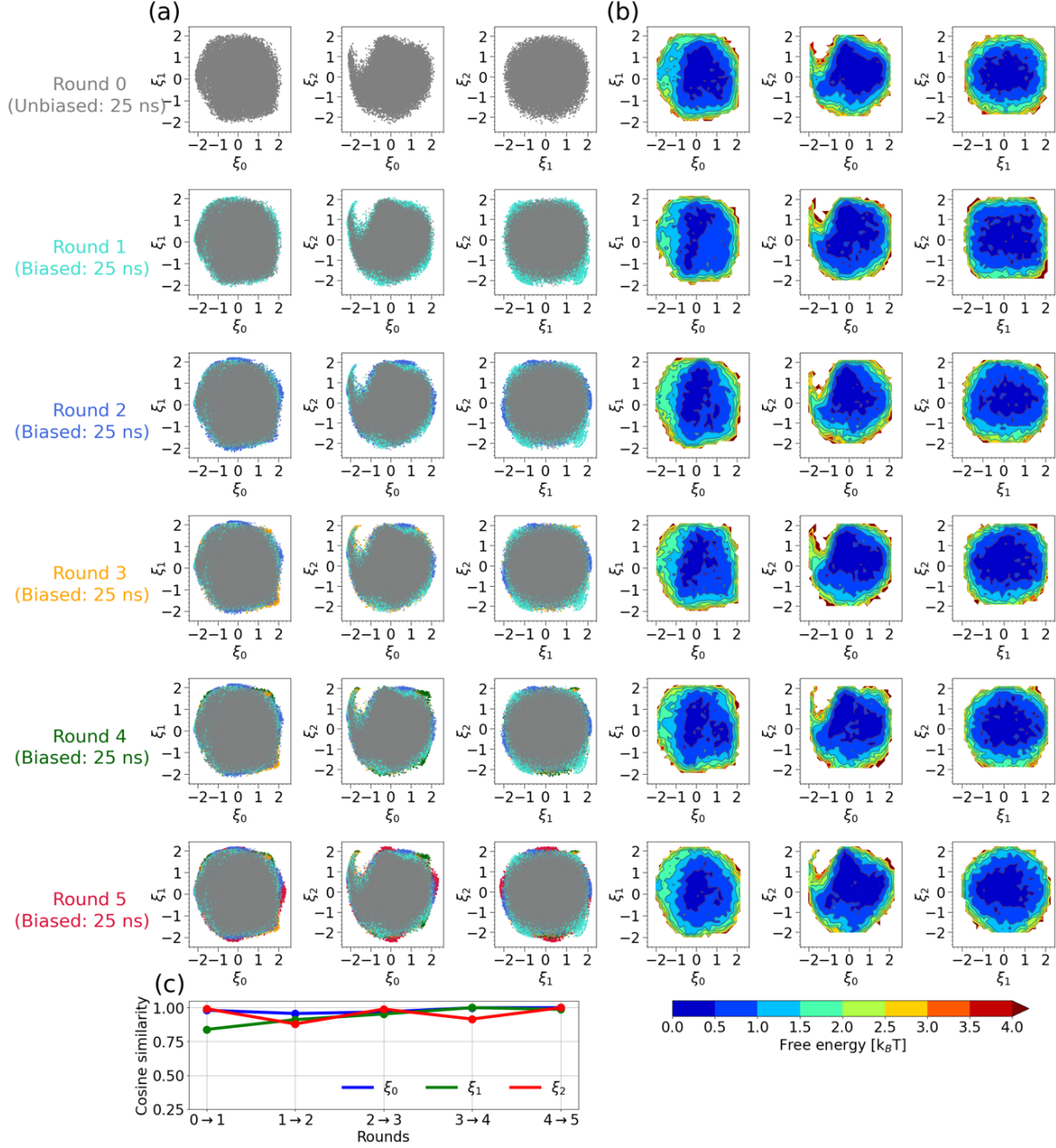


Figure 3.3: Application of PINES to a 13-particle Ar system in vacuum at 50 K. (a) Each column shows a 2D embedding of the round-by-round accrued sampling projected into the converged CVs ξ_0 , ξ_1 , and ξ_2 at the terminal Round 5. The unbiased simulation quite thoroughly samples the thermally accessible phase space, yet each biased round advances the frontier into unsampled regions on the periphery of the explored region. (b) Corresponding free energy surfaces (FES) computed by reweighting of the biased simulation data collected in the current round of sampling. (c) Cosine similarity between corresponding CVs in successive rounds. The high cosine similarity value indicates rapid convergence of the iterative CV discovery and enhanced sampling cycles within the five rounds of PINES.

We assess convergence of the iterative cycles of CV learning and biased sampling by computing the cosine similarity between corresponding CVs in successive rounds. As illustrated in Fig. 3.3c, we find each of the three learned CVs to quickly stabilize with round-to-round cosine similarity scores in excess of $\rho = 0.9$ for Round 1 and above. Together with the relatively modest expansion of the configurational phase space (Fig. 3.3a) and stabilization of the FES (Fig. 3.3b), we declare the PINES procedure to be converged within the five rounds.

It is informative to analyze the terminal learned CVs to provide insight about the high variance emergent collective modes within the system. By extracting the weights of the trained encoder, we perform layerwise matrix multiplication to compute a single scalar weight encapsulating the extent to which each PIV element contributes to each of the learned CVs integrated over all possible pathways through the network. Although this analysis does not account for correlations between the various PIV elements, it is useful in providing a gross assessment of how strongly each permutationally invariant pairwise distances influences each of the learned CVs. We present in Fig. 3.4a a graphical illustration of the contribution of each PIV element v_k to each of the the terminal learned CVs $\{\xi_0, \xi_1, \xi_2\}$, and also a cumulative sum over these weights indicating the net contribution and cooperative impact over multiple PIV elements at increasing pairwise distances. The absence of strong outliers in the bar charts indicates that none of the $\binom{13}{2} = 78$ PIV elements substantially dominate the values of the three learned CVs. A similar trend is manifested in the cumulative sums that indicate an approximately linear trend indicative of similar magnitude and similar signed contributions out to PIV element 35 or so. Recalling that the pairwise distances are passed through a switching function (Fig. 3.2) the larger pairwise distances appear earlier in the PIV, and for the shorter range interactions beyond element 35, we see a more interesting and challenging to interpret structure in the sign and magnitude of the PIV weights. The identical nature of all particles in this system and necessarily cooperative nature of assembly and disassembly is consistent with the absence of any single dominant contribution from the PIV.

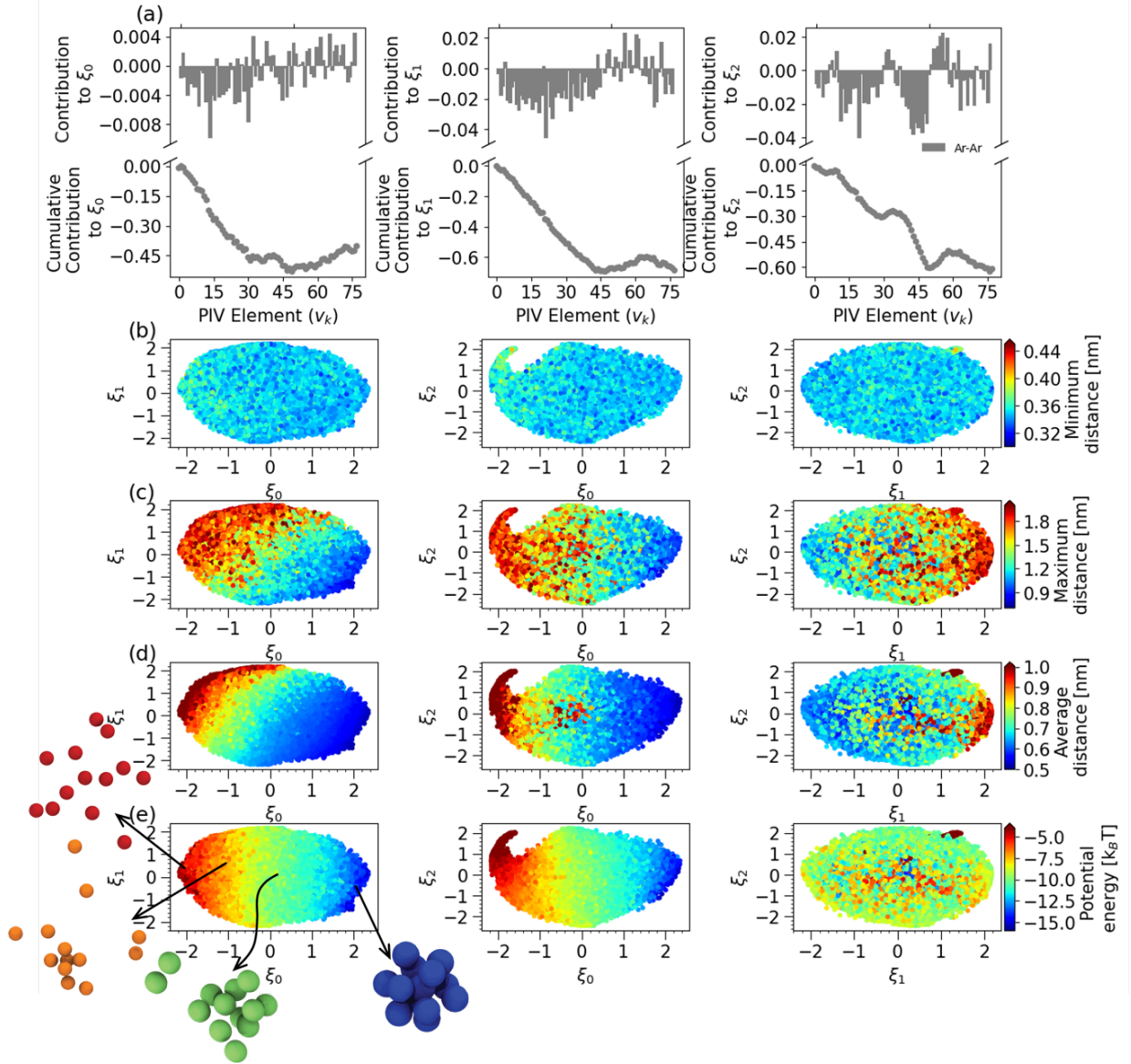


Figure 3.4: Bottom-up and top-down analysis of terminal learned CVs for the 13-particle Ar system. (a) Split plots show the signed contribution from each PIV element to each CV $\{\xi_0, \xi_1, \xi_2\}$ as a bar chart on top and as a cumulative sum of the signed contributions on the bottom. The magnitude of contributions on the bar chart emphasize the importance of individual PIV elements while the slopes in different regions of the scatter plot illuminate the concerted effect of atom groups on each CV. (b-e) Projections of the terminal Round 5 sampling data into the terminal CVs and colored by physically interpretable candidate variables. The potential energy, which is known a priori to be the putative CV for a Lennard-Jones fluid, correlates well with ξ_0 . Selected snapshots are presented to show the smooth transition across ξ_0 from the vapor-like phase (red, orange) to the condensed phase (green), and finally to the highly ordered solid-like phase (blue). The gradients in ξ_0 and ξ_1 with respect to the average and maximum pairwise distances indicates that PINES has learned these as important variables in driving assembly/disassembly.

We complement this bottom-up analysis of the learned CVs with a top-down interrogation in which we color projections of the Round 5 simulation data into the terminal CVs by candidate physical order parameters (Fig. 3.4b-e). In doing so, we find ξ_0 to be strongly correlated with the potential energy of the system (Fig. 3.4e). Since the potential energy is a sum over pairwise interactions, this makes ξ_0 a good correlate for the degree of association of the 13-particle cluster, with the compact, low-potential energy solid-like state residing at high ξ_0 , the dissociated, high-potential energy vapor-like state at low ξ_0 , and the liquid-vapor-like, partially associated / partially dissociated states at intermediate ξ_0 . The average pairwise distance (Fig. 3.4d) and maximum pairwise distance (Fig. 3.4c) are both well-correlated with a linear combination of ξ_0 and ξ_1 , with the “red hook” apparent in the ξ_0 - ξ_2 projections containing the most highly dissociated and rarefied vapor-like configurations. The minimum pairwise distance fails to show strong correlations with any of the three CVs (Fig. 3.4b). In sum, the three permutationally-invariant CVs learned by PINES as important collective modes to drive sampling of the configurational exploration of the system reflect the potential energy, compactness, and average/maximum pairwise interatomic distance.

Finally, in Fig. 3.5 we present our reweighted FES collected from our final round of enhanced sampling in the converged terminal learned CVs. We emphasize that these data were collected from enhanced sampling calculations conducted in the learned CVs (cf. Fig. 3.3), but that after sampling is complete, we are at liberty to reweight and project these data into arbitrary order parameters of our choosing for the purposes of expository illustration. The 1D FES constructed in the average pairwise distance clearly exposes the solid-like, filled icosahedron as the global free energy minimum of the system at an average distance of 0.62 nm (Fig. 3.5a). At higher average distances we enter into a liquid/vapor-like phase in which some fraction of the particles in the system are associated, and then terminate in a vapor-like phase with all particles dispersed lying within a weak free energy minimum at 1.17 nm approximately $4 k_B T$ higher in free energy than the solid-like phase. The 2D FES constructed

in the potential energy and maximum pairwise distance better illustrates the expansion of accessible phase space for the more disordered liquid-like and vapor-like states compared to the solid-like filled icosahdron (Fig. 3.5b).

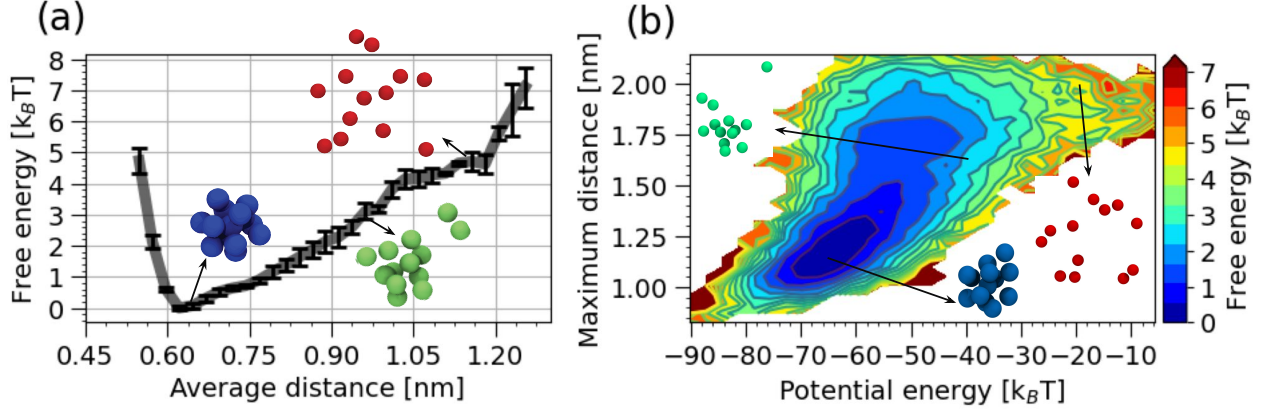


Figure 3.5: Reweighted free energy surfaces from the terminal round of enhanced sampling for the 13-particle Ar system projected into physically interpretable order parameters. (a) A 1D FES projected into the average pairwise interatomic distance. Representative snapshots show that the global free energy minimum to be occupied by a solid-like, filled icosahedron with an average pairwise distance of 0.62 nm. (b) A 2D FES projected into the potential energy and maximum pairwise interatomic distance. Representative snapshots illustrate the solid-like global free energy minimum and liquid-like and vapor-like phases.

The location and relative stability of the solid, liquid, and vapor-like phases is a strong function of the box size and temperature. In particular, the $2.5 \times 2.5 \times 2.5 \text{ nm}^3$ box places strict finite size effects on the maximum possible pairwise distances available to the particles in the vapor-like phase and limits the phase space volume accessible to it. The temperature also changes the relative balance of energetic and entropic contributions and modulating this away from the 50 K temperature at which our simulations were conducted can strongly modify the underlying free energy surface. To test the transferability of the converged CVs learned by PINES in a $2.5 \times 2.5 \times 2.5 \text{ nm}^3$ box at 50 K, we conducted enhanced sampling calculations in these CVs to recover the FES in an 8-fold larger $5 \times 5 \times 5 \text{ nm}^3$ simulation box (Fig. S3) and at temperatures of 25 K and 100 K (Fig. S4). Pleasingly, the CVs learned by PINES at $2.5 \times 2.5 \times 2.5 \text{ nm}^3$ and 50 K were capable of efficient exploration of the thermally-accessible configurational phase under these new conditions without any additional retraining. In particular, the enhanced sampling calculations had no difficulty in sampling the thermodynamically unfavorable solid-like phase in the large $5 \times 5 \times 5 \text{ nm}^3$ box (Fig. S4) or in traversing the $\sim 40 k_B T$ difference between the solid-like and vapor-like phases in the 25 K low temperature calculations (Fig. S4).

Taken together, this application of PINES to a simple 13-particle Ar system demonstrates that the method is capable of learning translationally, rotationally, and permutationally invariant CVs, using these CVs to efficiently drive PBMetaD enhanced sampling calculations, and furnishing transferable order parameters capable of driving sampling at other thermodynamic state points. We now proceed to test the method in applications to multi-component systems with more complex, rugged, and challenging-to-sample free energy surfaces.

3.4.2 Association/dissociation of a NaCl ion pair in water

We now apply PINES to discover CVs to drive the association/dissociation of a NaCl ion pair in water at 300 K. It is well known that solvent degrees of freedom play an important role

in mediating this process via many-body collective rearrangements of solvation shell water molecules [174, 175, 176, 3, 177]. We hypothesize that PINES should be able to identify permutationally-invariant ion and solvent-inclusive CVs associated with the indistinguishable solvent molecules and use these to drive sampling of the solvated ion pair dynamics.

We identified a $d=3$ dimensional latent space from an initial 50 ns unbiased simulation trajectory and confirm that this dimensionality remained unchanged in each of the five successive rounds of PINES (Fig. S2). We illustrate in Fig. 3.6a the expansion of the sampled phase space under each PINES iteration. The initial unbiased simulation provides good initial sampling of the central portion of the latent space (gray) but each successive biased round (different colors) provide additional sampling of the periphery. In particular, Rounds 1-3 (blue, yellow, green) show increased exploration of the CV space in half or less of the simulation time compared to the unbiased run and indicate that the CVs are accelerating sampling of novel states. Rounds 4-5 (red, black), despite being longer than the previous runs, offer diminishing returns in additional configurational sampling. (We recall that the run length in each round is dictated by convergence criteria of the PBMetaD simulations (cf. Sec. 3.3.4).)

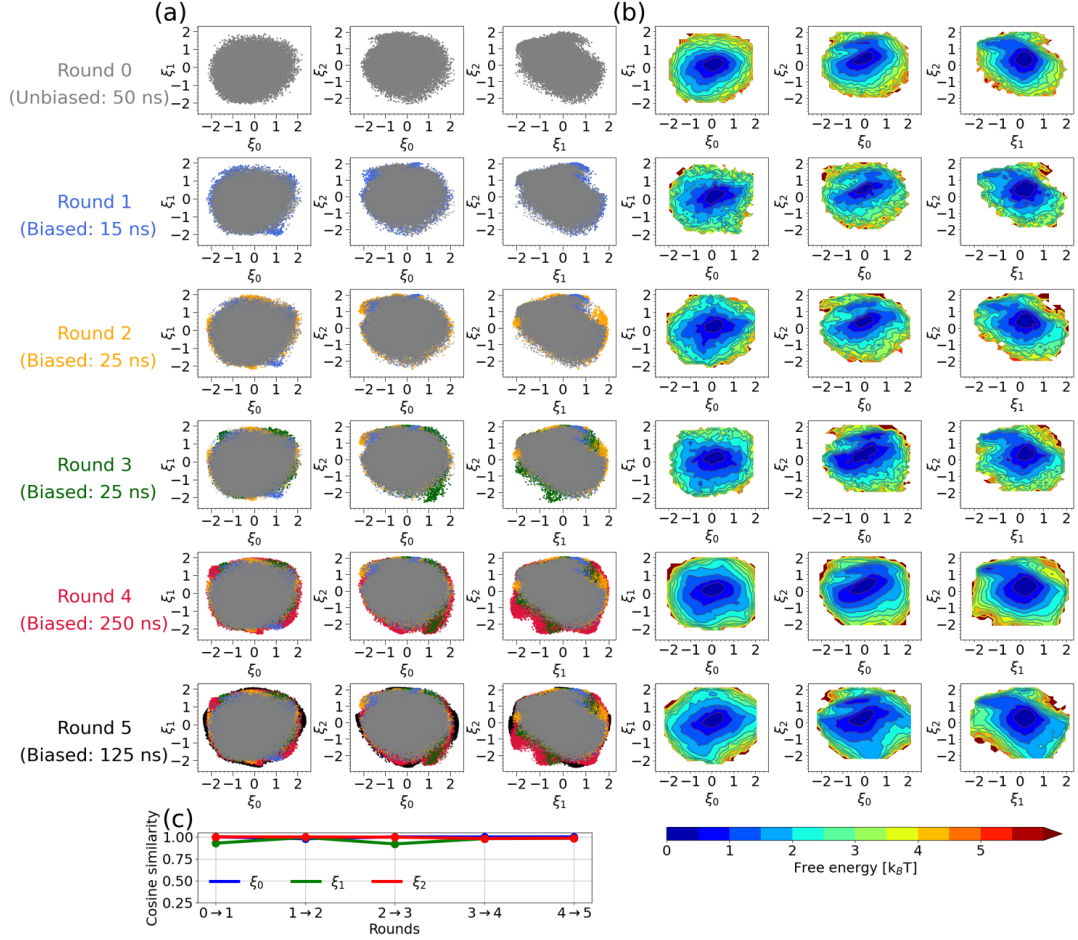


Figure 3.6: Application of PINES to a NaCl ion pair in water at 300 K. (a) Each column shows a 2D embedding of the round-by-round accrued sampling projected into the converged CVs ξ_0 , ξ_1 , and ξ_2 at the terminal Round 5. The unbiased simulation quite thoroughly samples the thermally accessible phase space, yet each biased round advances the frontier into unsampled regions on the periphery of the explored region. Rounds 1-3 show broader exploration of the CV space in half or less of the simulation time compared to the unbiased run and indicate that the CVs are accelerating sampling of novel states. Rounds 4-5, despite being much longer than the previous biased runs, provide diminishing returns in the exploration of new areas of CV space. (b) Corresponding free energy surfaces (FES) computed by reweighting of the biased simulation data collected in the current round of sampling. (c) Cosine similarity between corresponding CVs in successive rounds. The high cosine similarity values between all neighboring rounds indicate that the CVs discovered in the initial unbiased round are not substantially altered by learning over the biased data and the PINES protocol rapidly converges.

In Fig. 3.6b, we present the corresponding reweighted FES for each round. We observe visually apparent expansion of the FES into high-free energy peripheral regions and changes in the free energy distributions through Rounds 0-4. The FES between Rounds 4-5, however, are very similar in terms of the explored regions and free energy contours. As we discuss below, the core of the latent space that was readily sampled by the unbiased simulation corresponds to dissociated configurations with average levels of hydration. The peripheral states that require biased simulations for adequate exploration are associated with higher free energy states corresponding to metastable configurations such as the solvent-shared pair.

We present in Fig. 3.6c the cosine similarity between CVs in successive PINES rounds. In this case we find that the CVs almost immediately stabilize, with those learned from the unbiased simulation data changing relatively little over the course of the iterative CV discovery and enhanced sampling protocol and always remaining above $\rho = 0.9$. Taken together with the observed trends in the latent space and FES, we surmise that the initial CVs provide a good global parameterization of the system configurational space, but require approximately three rounds of enhanced sampling and minor CV refinement to attain convergence in sampling of the thermally accessible phase space.

We elected to use PIVs that considered the 10 closest water molecules to each ion (Section 2.2.2), with this number selected to be sufficiently large to track all water molecules within approximately the first hydration shell. To assess the dependence of the number of water molecules included within the PIV, we reran our analysis using the 15, 25, and 35 most proximate water molecules to explore the effect of including water molecules in higher hydration shells. We find that both the CVs discovered by the PINES protocol (Fig. S5) and that the exploration of phase space within the CVs (Fig. S6) are quite insensitive to the number of proximal water molecules included in the PIV. This suggests that the approximately 10 proximal water molecules necessary to capture the important solvation effects within the first hydration shell expose sufficient information for the discovery of CVs capable

of characterizing and driving good global sampling of the relevant configurational motions of the system. In general, the appropriate number of closest solvent molecules to consider is anticipated to be context and system dependent, and a parametric analysis such as this can be useful in identifying an appropriate threshold.

In Fig. 3.7a, we expose the integrated weights linking each of the 61 PIV elements to each of the three CVs. Inspection of the individual (upper) and cumulative (lower) weights indicates that particular interactions appear to contribute more significantly to each of the three CVs. In particular, PIV elements corresponding to Na-O and Cl-O distances make large magnitude contributions to ξ_0 , whereas Na-H and Cl-H distances seemingly contribute more strongly to ξ_1 , and Na-O makes the largest contribution to ξ_2 . Interestingly, the Na-Cl interaction is not a top contributor to any CVs, suggesting that the information about the Na-Cl state may be delocalized among the information on solvation shell water distances.

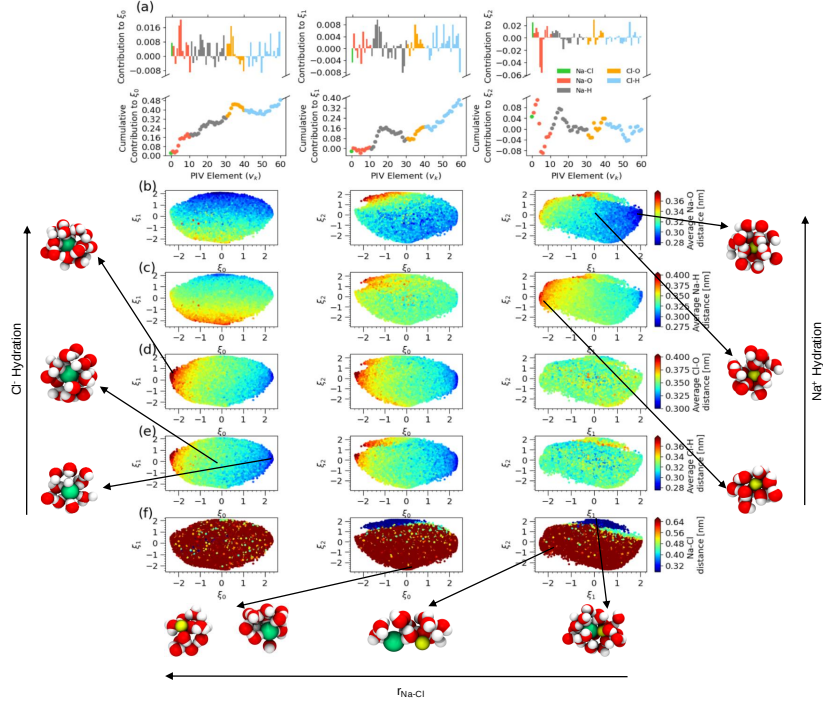


Figure 3.7: Bottom-up and top-down analysis of terminal learned CVs for the NaCl ion pair in water. (a) Split plots show the signed contribution from each PIV element to each CV $\{\xi_0, \xi_1, \xi_2\}$ as a bar chart on top and as a cumulative sum of the signed contributions on the bottom. The magnitude of contributions on the bar chart emphasize the importance of individual PIV elements while the slopes in different regions of the scatter plot illuminate the concerted effect of atom groups on each CV. (b-f) Projections of the terminal Round 5 sampling data into the terminal CVs and colored by physically interpretable candidate variables. ξ_0 is well-correlated with the average Cl-O and Cl-H distance, ξ_1 with the average Na-O and Na-H distance, and ξ_2 with the Na-Cl distance. Three distinct color bands are apparent in panel (f) corresponding to the contact pair (blue), the solvent-shared (green), and disassociated (red) states (cf. Fig. 3.8c).

We further explore these trends in Fig. 3.7b-f where we color the latent space by candidate physical variables. These plots are illuminating in showing that ξ_0 is well-correlated with the average Cl-O and Cl-H distance, ξ_1 with the average Na-O and Na-H distance, and ξ_2 with the Na-Cl distance. This analysis offers a straightforward interpretation of the three learned CVs as corresponding to Cl ion hydration (ξ_0), Na ion hydration (ξ_1), and Na-Cl proximity (ξ_2).

We present in Fig. 3.8 the reweighted FES collected from our final round of PINES enhanced sampling projected into physical order parameters and annotated with representative molecular configurations. A 1D FES constructed in the Na-Cl distance clearly shows the contact pair ($r_{\text{Na-Cl}} = 0.26$ nm), solvent-shared ($r_{\text{Na-Cl}} = 0.51$ nm), and disassociated ($r_{\text{Na-Cl}} \gtrsim 0.6$ nm) states (Fig. 3.8a). The free energy barrier between the contact pair and solvent-shared state lies at $r_{\text{Na-Cl}} = 0.36$ nm with a height relative to the contact pair of $(4.9 \pm 0.1) k_B T$. This stands in good agreement with reported literature values of 4.8-5.3 $k_B T$ and provides good support for the accurate sampling of the configurational space and FES generated by PINES [187, 188]. We present in Fig. 3.8b a 2D FES constructed in the Na-Cl distance and the solvent coordination number of the Na ion n_{Na} . This projection is illuminating in showing a splitting of the contact pair at $r_{\text{Na-Cl}} = 0.26$ nm into two basins with coordination numbers of $n_{\text{Na}} = 4$ and $n_{\text{Na}} = 5$. The FES suggests that the transition to the solvent-shared state at $r_{\text{Na-Cl}} = 0.51$ nm appears to proceed from the $n_{\text{Na}} = 5$ state and then drops into the lower free energy $n_{\text{Na}} = 6$ state after separation has completed, although dynamical analyses such as transition path sampling would be required to test this conjecture.

To assess the importance of the solvent features learned by PINES in sampling the thermally relevant phase space, we present in Fig. S7 a comparison of the 1D FES in the Na-Cl distance and the 2D FES in the Na-Cl distance and the solvent coordination number of the Na ion n_{Na} computed by PINES and by well-tempered metadynamics enhanced sampling

in the Na-Cl distance alone. This analysis shows that while metadynamics in the Na-Cl distance alone accurately recovers the 1D FES in the Na-Cl distance in good accord with PINES, it fails to expose low-free energy regions associated with highly solvated states of the Na ion that are only exposed by driving sampling in solvent-based collective variables learned by PINES.

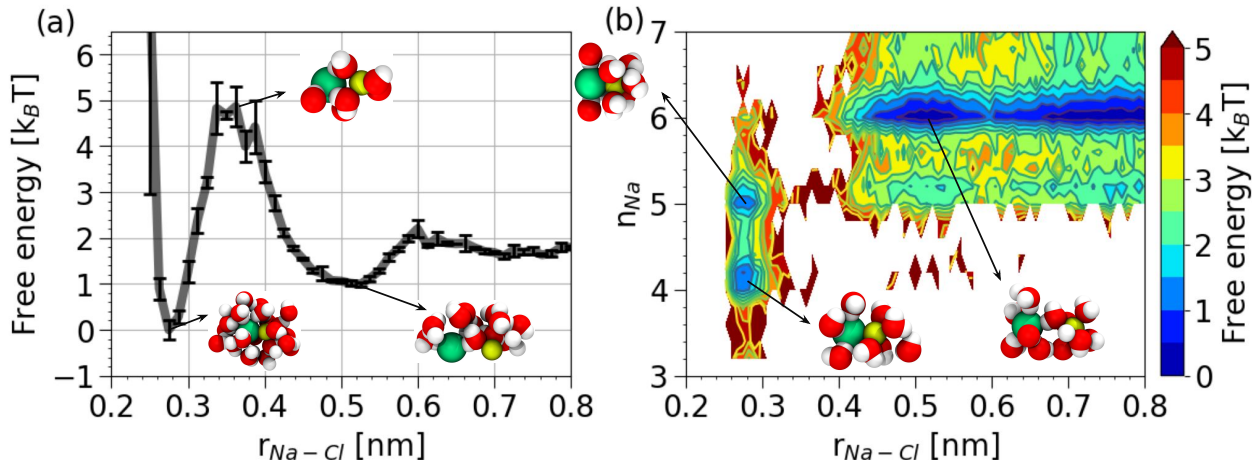


Figure 3.8: Reweighted free energy surfaces from the terminal round of enhanced sampling for the NaCl ion pair in water projected into physically interpretable order parameters. (a) A 1D FES projected into the Na-Cl distance. The landscape clearly illustrates the contact pair ($r_{Na-Cl} = 0.26$ nm), solvent-shared ($r_{Na-Cl} = 0.51$ nm), and disassociated ($r_{Na-Cl} \gtrsim 0.6$ nm) states. The free energy barrier lying between the contact pair and solvent-shared states is $(4.9 \pm 0.1) k_B T$ as measured from the contact pair, and involves a collective rearrangement of solvent shell water molecules to accommodate the separating ion pair [3, 4]. Uncertainties in the free energy correspond to standard errors estimated by five-fold block averaging. (b) A 2D FES projected into the Na-Cl distance and the solvent coordination number of the Na ion n_{Na} . The contact pair comprises multiple solvation states of the Na ion. Transitions to the solvent separated pair appear to proceed most readily through the $n_{Na} = 5$ contact pair state.

To assess the transferability of the converged CVs learned by PINES for the NaCl ion pair association/dissociation process, we conducted enhanced sampling calculations in these CVs for a hypothetical divalent ion system generated by doubling the charge on each ion (Fig. S8). Whereas an unbiased simulation of the divalent ion pair is completely localized near the initial state corresponding to the contact pair because of the strong electrostatic attractions (Fig. S8a,b), PBMetaD enhanced sampling simulations in the monovalent CVs are capable of surmounting a $\sim 45 k_B T$ free energy barrier to drive sampling of the dissociated state (Fig. S8c,d).

This analysis indicates the transferability of CVs learned over one system applied to another in which it is more challenging to drive sampling.

In sum, this application to a NaCl ion pair in water demonstrates the capacity of PINES to identify permutationally-invariant CVs incorporating information on the ion pair and local solvation environment and use these variables to efficiently drive enhanced sampling of ion pair association/dissociation. This illustrates the capacity of the method to learn permutationally-adapted CVs in a solvated system of chemically indistinguishable water molecules. The unbiased simulations used to seed the PINES campaign provided a relatively comprehensive sampling of the phase space and the learned CVs remained relatively unchanged over the course of the iterative search. In the next application, we demonstrate a case where the PINES enhanced sampling is instrumental in good sampling of phase space.

3.4.3 Hydrophobic collapse of an n -pentatetracontane ($C_{45}H_{92}$) polymer chain in water

In our final application, we consider hydrophobic collapse of an n -pentatetracontane ($C_{45}H_{92}$) polymer chain in water to demonstrate an application of PINES to a solvated polymer system. The dynamical collapse of hydrophobic polymers in water is known to be intimately related to drying transitions and the formation of water cavities [189, 190, 143, 1]. Solvent-centric

CVs are therefore expected to play an important role in the collapse/extension transitions of the $C_{45}H_{92}$ polymer chain and, for such a long chain, it can also be challenging to comprehensively explore the configurational phase space using unbiased simulations alone. This system is therefore anticipated to be a good test of the PINES methodology.

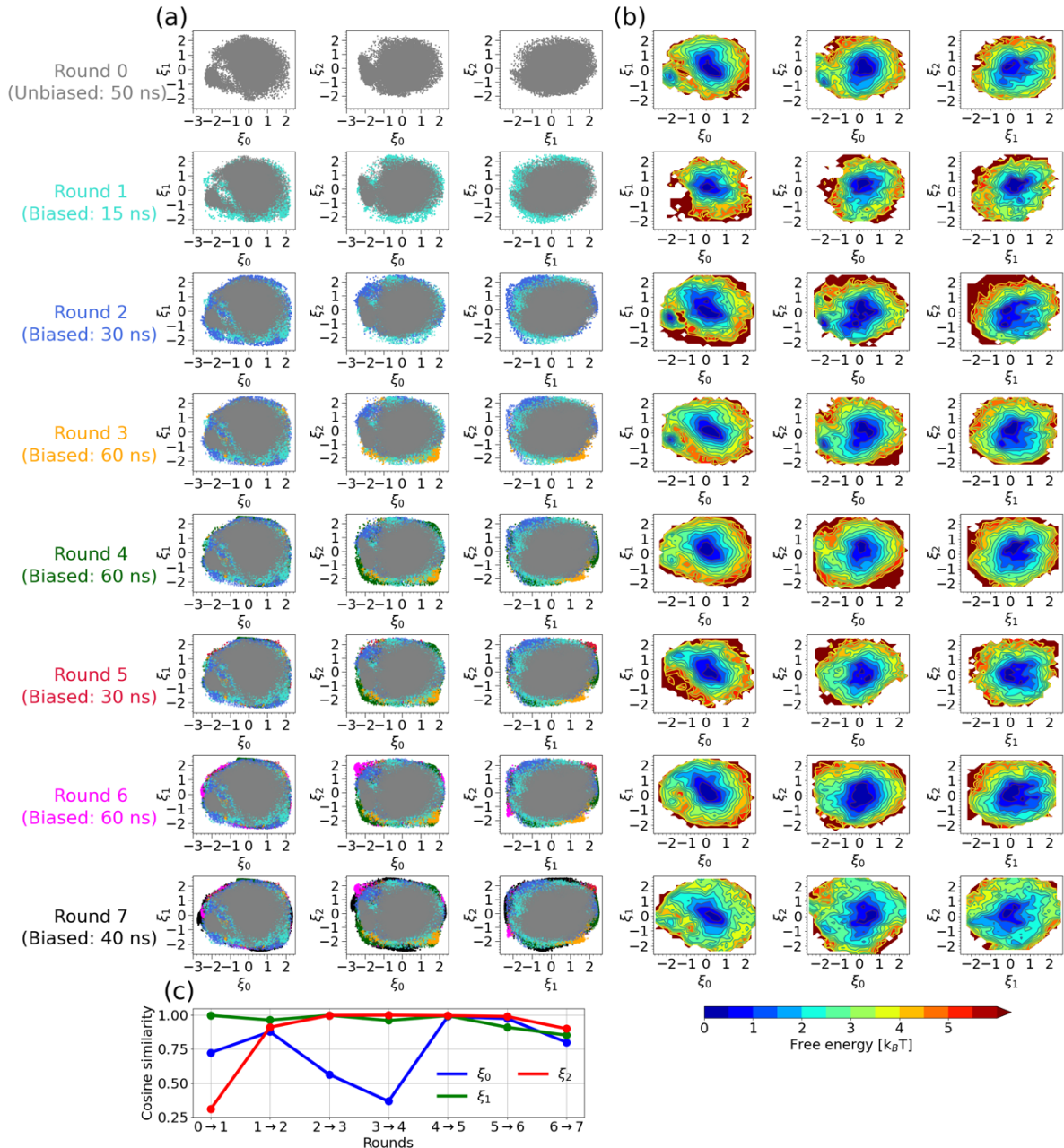


Figure 3.9: Application of PINES to an *n*-pentatetracontane ($C_{45}H_{92}$) chain in water at 300 K. (a) Each column shows a 2D embedding of the round-by-round accrued sampling projected into the converged CVs ξ_0 , ξ_1 , and ξ_2 at the terminal Round 7. Rounds 1-4 show significantly increased exploration of novel states compared to the unbiased run and indicate that the CVs are accelerating sampling of unexplored regions of phase space. In Rounds 5-6 the sampling of the CV space begins to saturate and suggests that the CVs have converged. (b) Corresponding free energy surfaces (FES) computed by reweighting of the biased simulation data collected in the current round of sampling. (c) Cosine similarity between corresponding CVs in successive rounds. The CVs change significantly round-to-round over Rounds 1-4 before converging in Rounds 5-7.

Once again, we identified a $d=3$ dimensional latent space from an initial 50 ns unbiased trajectory that remained unchanged over the course of the seven subsequent rounds of PINES (Fig. S2). We present in Fig. 3.9a the expansion of the sampled latent space over the PINES iterations. We see significant expansion into new, unexplored regions by the enhanced sampling calculations in Rounds 1-4, before sampling approximately converges and achieves only minor additional exploration in Rounds 5-7. We draw particular attention to the interior “hole” in the latent space apparent at $(\xi_0=-1, \xi_1=0)$ in Round 0 (gray) and which is filled in by the enhanced sampling trajectories in Round 1 (cyan) and Round 2 (blue). As we discuss below, this hole corresponds to a high-free energy transition region between collapsed and extended states of the chain that is only adequately sampled under PBMetaD biasing in the learned CVs. Fig. 3.9b presents analogous FES plots for each PINES round that expose the density distribution of points over the latent space and convergence of the landscapes over the PINES rounds. Fig. 3.9c presents the cosine similarity between CVs in successive PINES rounds. We observe significant changes in the learned CVs over Rounds 1-4 as they are exposed to additional configurational data supplied by the enhanced sampling calculations. The CVs stabilize in Rounds 5 and 6 with all CVs maintaining cosine similarity values of $\rho > 0.9$.

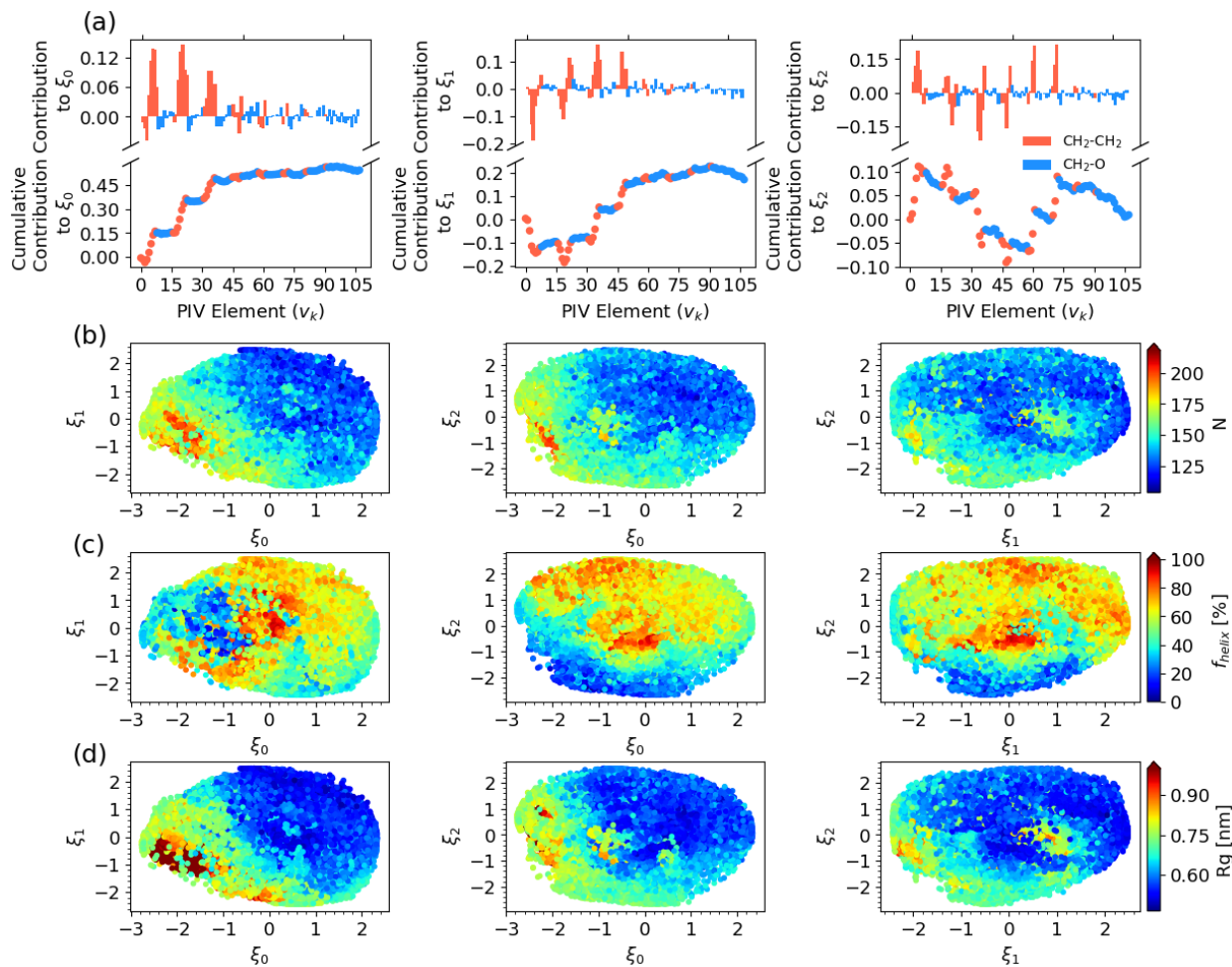


Figure 3.10: Bottom-up and top-down analysis of terminal learned CVs for the *n*-pentatetracontane chain in water. (a) Split plots show the signed contribution from each PIV element to each CV $\{\xi_0, \xi_1, \xi_2\}$ as a bar chart on top and as a cumulative sum of the signed contributions on the bottom. The magnitude of contributions on the bar chart emphasize the importance of individual PIV elements while the slopes in different regions of the scatter plot illuminate the concerted effect of atom groups on each CV. (b-d) Projections of the terminal Round 7 sampling data into the terminal CVs and colored by physically interpretable candidate variables. ξ_0 and ξ_1 are correlated with the chain radius of gyration, R_g , and the number of waters, N , in the first solvation shell of the chain. We highlight representative configurations on the ξ_0 - ξ_1 surface corresponding to the extended (red), collapsed (blue), and partially collapsed (green) structures. ξ_2 is correlated with the helical pitch f_{helix} . We highlight representative configurations on the ξ_1 - ξ_2 surface corresponding to planar rings (blue) and corkscrews (red).

In Fig. 3.10, we gain insight into the learned CVs by presenting the network weights linking each of the 108 PIV elements to the three CVs and coloring the latent space embeddings by candidate physical variables. Interpreting the CVs for more complex systems can be challenging, but we find ξ_0 and ξ_1 to possess significant correlation with both the chain radius of gyration R_g and the number of waters N in the first solvation shell of the chain, defined by counting the number of water molecule O atoms within 0.6 nm of any united atom *CH2* or *CH3* bead of the alkane chain. ξ_2 is correlated, albeit non-monotonically, with the fraction of helical pitch f_{helix} . This observable, defined as $f_{\text{helix}} = r_{CH3-CH43} - (r_{CH3-CH23} + r_{CH13-CH33} + r_{CH23-CH43})$ and min-max normalized, measures deviations from a common helical chain structure with a 20 united atom turn, where $r_{CH3-CH43}$ is the distance between the 3rd and 43rd united atoms in the chain and $(r_{CH3-CH23} + r_{CH13-CH33} + r_{CH23-CH43})$ are the distances between 20 united atom separations in the chain. Large values of f_{helix} correspond to an extended chain with regular turns of 20 united atoms while low values correspond to wider helices and ring-like structures. It also appears that solute-solvent PIV elements are most strongly represented within ξ_2 . Taken together, the CVs appear to contain signatures of the compactness, global hydration, and helicity of the polymer chain within the water solvent.

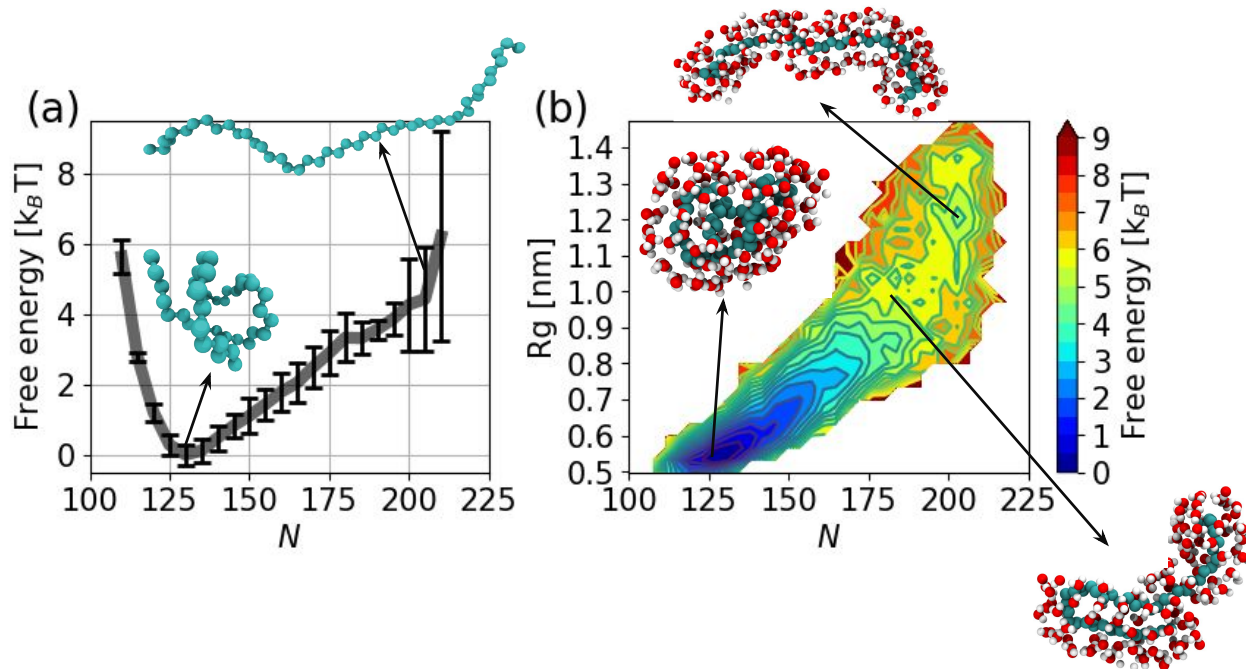


Figure 3.11: Reweighted free energy surfaces from the terminal round of enhanced sampling for the *n*-pentatetracontane chain in water. (a) A 1D FES projected into the number of waters N in the first solvation shell of the chain. A collapsed state with $N \approx 130$ is more stable than the fully extended state with $N \approx 200$ by approximately $4 k_B T$. Uncertainties in the free energy correspond to standard errors estimated by five-fold block averaging. (b) A 2D FES projected into the number of waters N in the first solvation shell of the chain and the radius of gyration R_g . The global free energy minimum at $(N \approx 130, R_g \approx 0.57 \text{ nm})$ contains the hydrophobically collapsed chain.

In Fig. 3.11 we show our converged and reweighted FES from the terminal round of PINES projected into physical order parameters. A 1D FES projected into the number of waters N in the first solvation shell of the chain exposes the collapsed state of the chain to be more stable than the extended by approximately $(4\pm1) k_B T$, in good agreement with the value of $3.3 k_B T$ reported by Dhabal *et al.* [1] (Fig. 3.11a). A 2D FES constructed in N and R_g clearly exposes the relationship between hydration and chain extent, with the collapsed state of the chain better shielding the hydrophobic beads from solvent and occupying the global free energy minimum at this thermodynamic state point (Fig. 3.11b).

3.5 Conclusions

In this work we have introduced Permutationally Invariant Network for Enhanced Sampling (PINES) as an approach for data-driven discovery and enhanced sampling in translationally, rotationally, and permutationally invariant collective variables for molecular systems. The approach essentially integrates Permutation Invariant Vector (PIV) featurizations with the Molecular Enhanced Sampling with Autoencoders (MESA) approach to perform interleaved rounds of symmetry-adapted collective variables and parallel bias metadynamics enhanced sampling to perform comprehensive sampling of the thermally accessible phase space and obtain converged thermodynamic averages. Convergence of the iterative sampling process is assessed by stabilization of the learned collective variables and saturation of the explored configurational phase space. PINES is well-suited to data-driven collective variable discovery and enhanced sampling in challenging-to-sample systems of indistinguishable particles frequently encountered in self-assembling and solvated systems. We have demonstrated PINES in applications to the assembly/disassembly of a 13-particle Ar cluster, association/dissociation of a NaCl ion pair in water, and hydrophobic collapse of a *n*-pentatetracontane ($C_{45}H_{92}$) chain in water. In all instances, PINES efficiently discovers symmetry-adapted multi-body collective variables that are used to efficiently drive sampling of configurational space. We

anticipate that PINES may be of value in applications to diverse molecular systems comprising indistinguishable particles and we have made the approach freely available as a new user-friendly module at <https://github.com/Ferg-Lab/pines.git> that can be patched with PLUMED2 enhanced sampling libraries and will be submitted for incorporation into a future release of PLUMED2 [8].

In future work, we would like to apply PINES to engage more complex processes such as protein-folding, DNA self-assembly, and ligand-binding. Additionally, we envision a number of future technical improvements and modifications to PINES. One of the key advantages of PINES is its flexibility through modularity. Any generic CV-based biasing method is immediately compatible with PINES, but with small modifications to the open-source code the featurization choice and the network architecture may also be adapted to employ alternative CV discovery techniques such as, for example, SRVs to discover slow, as opposed to high-variance, CVs. We also note that a PyTorch PLUMED2 module [191, 192] enables extension of the PINES approach to use arbitrarily complicated ANN architectures within the PINES pipeline. In addition, we are also implementing PINES in the flexible GPU accelerated enhanced sampling library PySAGES [193]. A principal use case of PINES is its application to solvated systems. The current implementation of PINES employs only solute-solvent distances within the PIV and does not consider the solvent-solvent interaction block. While this has the benefit of reducing the size of the PIV, the inclusion of such interactions would provide additional information on the structure of the near-field solvent that may be important for particular systems such as proteins with highly structured solvation shells. Lastly, we note that the current method makes no use of the decoder portion of the network other than its necessity in training the autoencoder network. The decoder, however, could possibly be utilized to generate novel conformations by sampling from sparse regions in the latent space to seed enhanced sampling calculations and accelerate convergence of the PINES protocol.

3.6 Acknowledgements

Work by N.S.M.H. was supported by the National Science Foundation under Grant No. DMR-1841807. Work by S.D. was supported by MICCoM (Midwest Center for Computational Materials), as part of the Computational Materials Science Program funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, through Argonne National Laboratory, under Contract No. DE-AC02-06CH11357. This work was completed in part with resources provided by the University of Chicago Research Computing Center. We gratefully acknowledge computing time on the University of Chicago high-performance GPU-based cyberinfrastructure supported by the National Science Foundation under Grant No. DMR-1828629.

3.7 Appendix: Supporting Information

Supplementary information comprising switching function parameterization process from pairwise RDFs, FVE plots used to determine the appropriate dimensionality of the latent space, analysis of convergence and converged FES for 13-atom Argon system in an entropically dominated $5 \times 5 \times 5 \text{ nm}^3$ box, analysis of transferability of learned CVs for the 13 atom Argon system to out-of-training temperature regimes, cosine similarity of discovered CVs for NaCl system with various numbers of proximal water considered, projection of the trajectories for various NaCl water systems into the CVs identified for the 35 water system, comparison of the free energy surfaces for the NaCl system obtained by PINES to enhanced sampling in the Na-Cl distance alone, and analysis of transferability of learned CVs from the NaCl system to a divalent ion pair by scaling the $Na+$ and $Cl-$ valency.

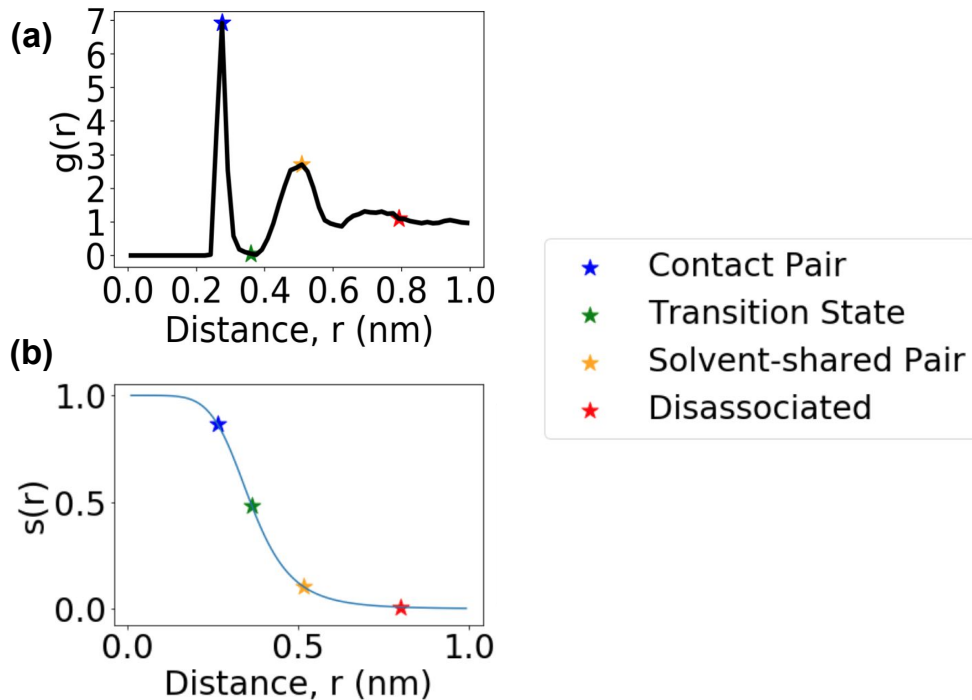


Figure 3.12: Switching function parameterization process from pairwise RDFs. (a) The Na-Cl pair correlation function is used to determine the r_0 parameter of the switching function, which is chosen to be the midpoint between the first two peaks of $g(r)$. This process naturally embeds transition regions close to r_0 on the switching function curve. (b) The second parameter, n , controls the steepness of the switching function curves with larger values producing steeper curves. n is chosen as the lowest positive integer which embeds the first $g(r)$ peak at a value greater than or equal to 0.9 and embeds the second $g(r)$ peak at a value less than or equal to 0.1. Four distinct system states are labeled along the Na-Cl pair correlation function and Na-Cl switching function, showing the conversion of pairwise distances to PIV values and the separation of states along the (0,1) range.

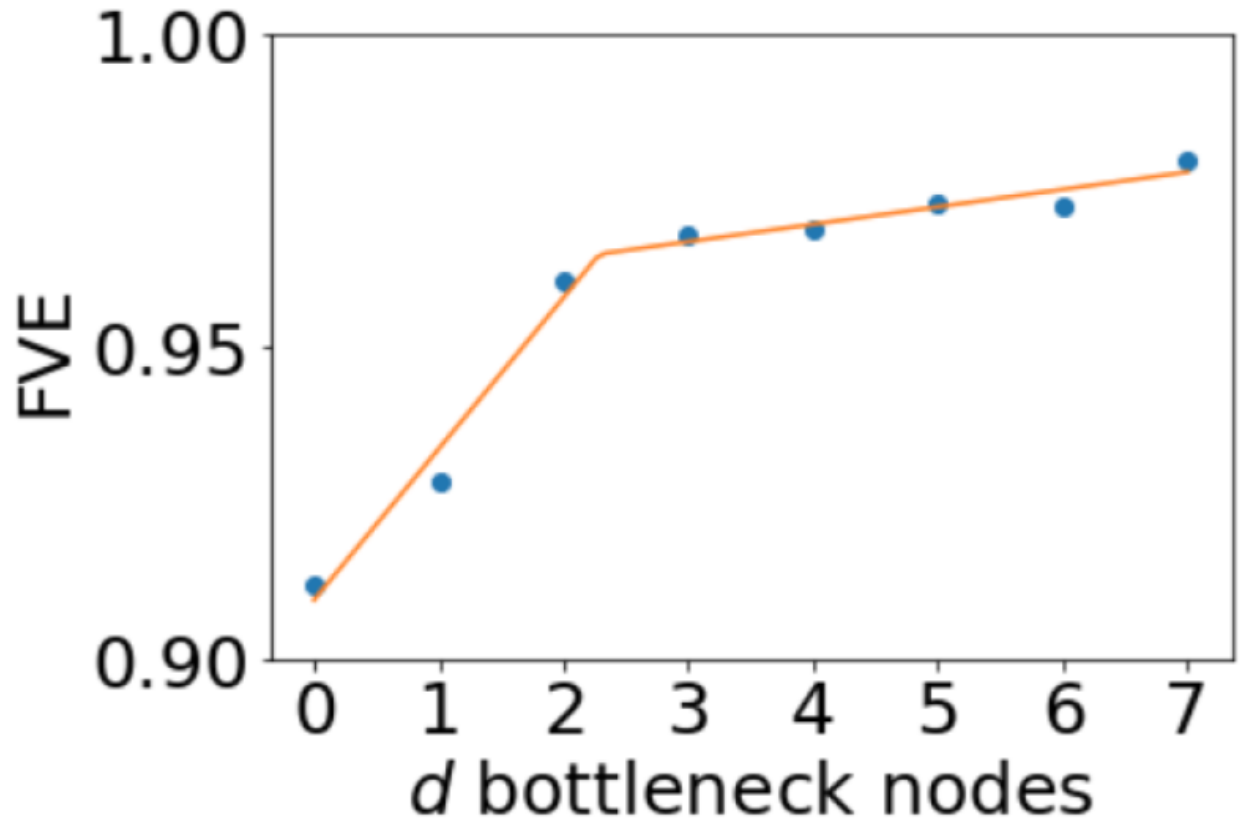


Figure 3.13: Example of the fraction of variance explained (FVE) plots used to determine the appropriate dimensionality of the latent space. In line with prior work, we identify a knee in the FVE plot that indicates the optimal number of bottleneck nodes. This example shows the intrinsic dimensionality identification for the network trained on the aggregated data set and is found to be $d = 3$.

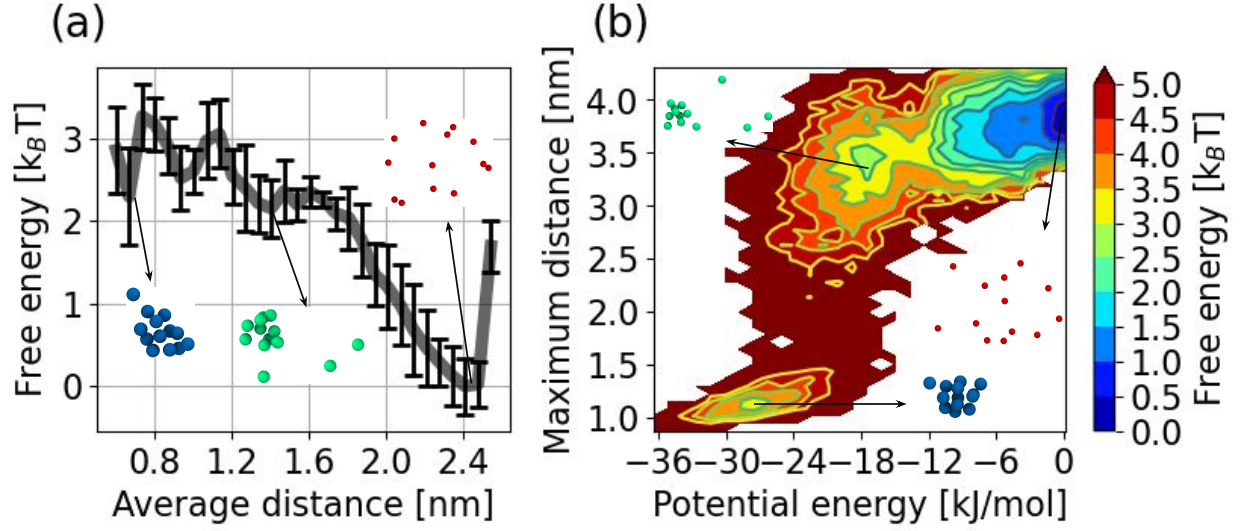


Figure 3.14: Analysis of convergence and converged FES for 13 Argon atoms in an entropically dominated $5 \times 5 \times 5 nm^3$ box. (a) The probability distribution, and (b) log probability distribution, for the $5 \times 5 \times 5 nm^3$ box show that the gas-like state is highly favored compared to the liquid-like and solid-like phases. Volume effects are particularly impactful on finite size systems and the increase in the box size represents a shift to a system dominated by entropy rather than enthalpy. (c) The deposited gaussian height along each CV as a function of time. It can be seen that all three CVs are sufficiently converged as evidenced by the severe drop in the deposition height by the end of the trajectory. (d,e) The 1D (d) and 2D (e) FES show a global energy minimum for the gas-like state and a metastable liquid-like state.

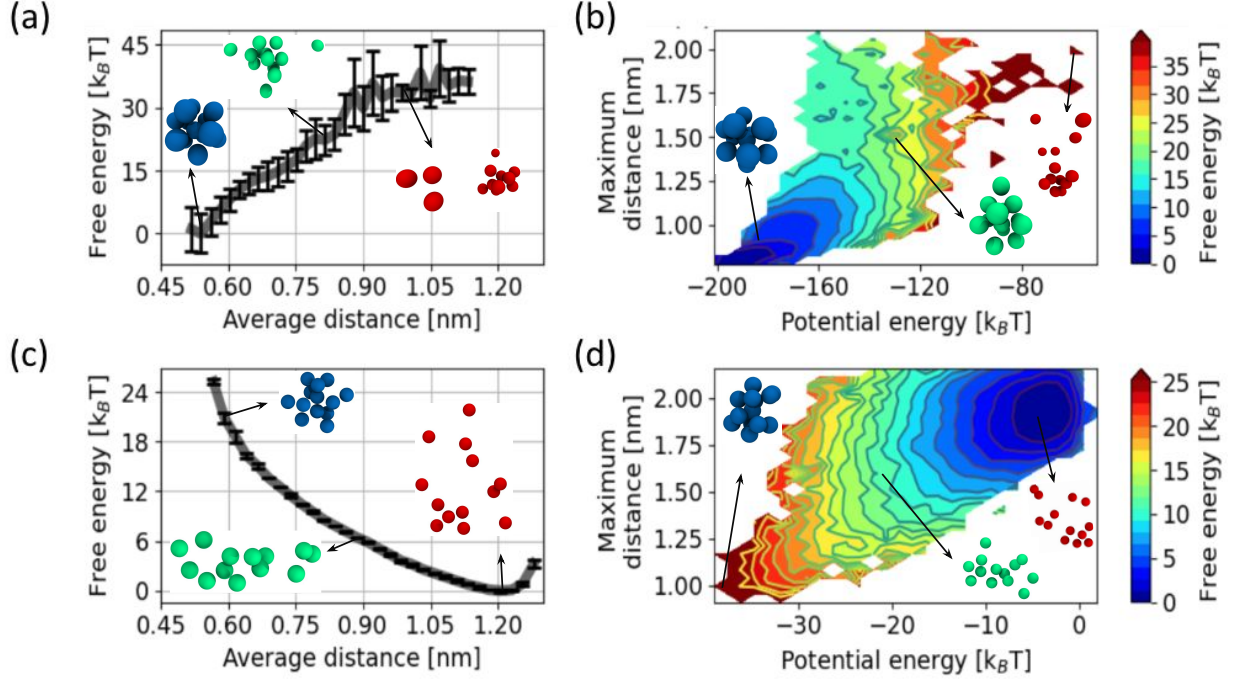


Figure 3.15: Analysis of transferability of learned CVs for the 13 atom Argon system to out-of-training temperature regimes. (a,b) The respective 1D and 2D FES for the system at 25K. The solid-like phase is strongly favored and the learned CVs continue to effectively facilitate psuedo-phase transtions and exploration of the liquid-like and gas-like phases. (c,d) The respective 1D and 2D FES for the system at 100K. The gas-like phase is strongly favored and the learned CVs continue to effectively facilitate psuedo-phase transtions and exploration of the liquid-like and solid-like phases.

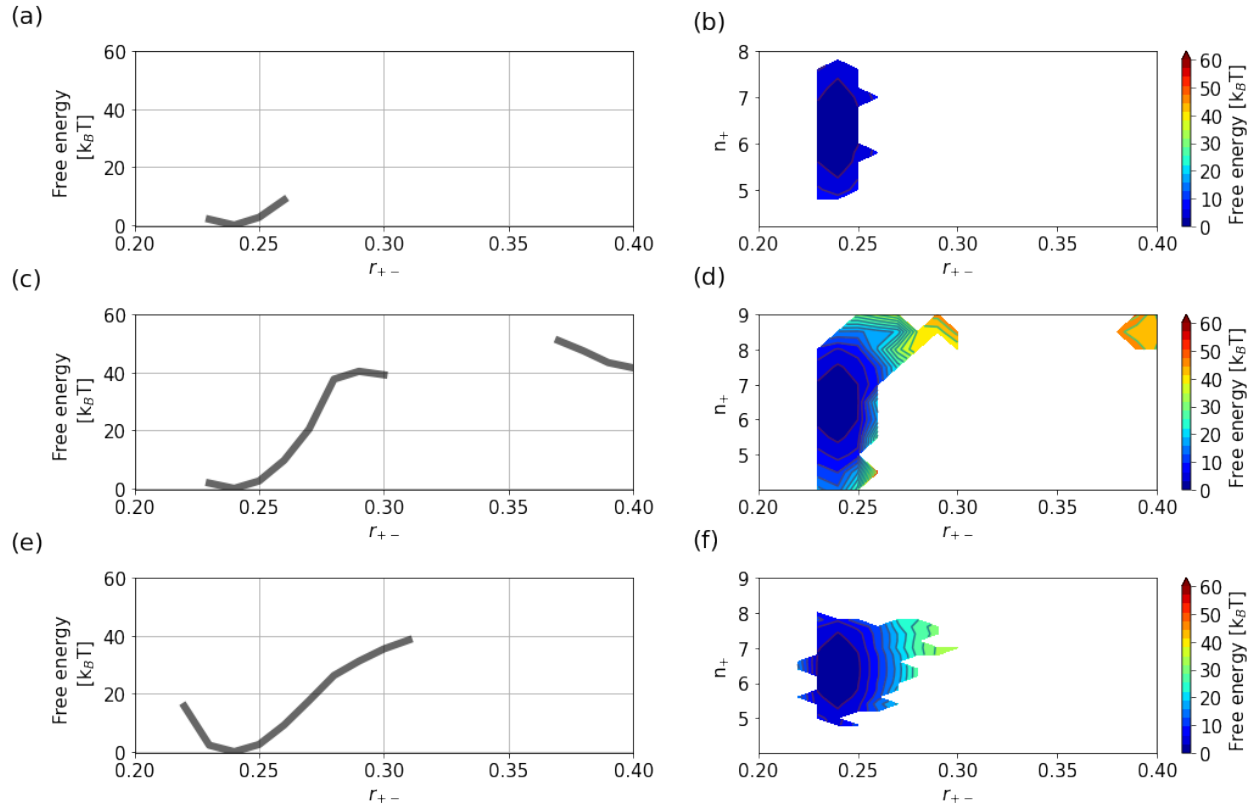


Figure 3.16: Analysis of transferability of learned CVs for the NaCl system to out-of-training charge regimes as divalent ions. (a,b) The respective 1D and 2D FES for a 25 ns unbiased simulation of the divalent system. The increased charges on the ions prohibit exploration of configurations outside of the contact pair. (c,d) The respective 1D and 2D FES for a 25 ns PBMetaD simulation of the divalent system along the PINES-identified CVs. The transition barrier separating the contact pair and solvent-shared pair states is $\approx 30\times$ larger compared to the monovalent system, however acceleration along the PINES CVs is sufficient to traverse the transition barrier. (e,f) The respective 1D and 2D FES for a 25 ns MetaD simulation along the ion distance of the divalent system. While an improvement on the exploration of the unbiased simulation, the ion distance is not a sufficient CV to efficiently traverse the large transition barrier.

CHAPTER 4

ACTIVE LEARNING FOR OPTIMAL THERMORESPONSIVE DNA-BASED SWITCHABLE SUPERLATTICES

The following chapter covers a manuscript that is still in progress, but which will constitute another first-author publication to be submitted by the end of Summer 2025.

N.S.M. Herringer, M. Gruschka, Q. Chen, D. McKeen, O. Gang, J. Mittal, and A.L. Ferguson “Active Learning for Optimal Thermoresponsive DNA-based Switchable Superlattices” in preparation, 2025

4.1 Abstract

DNA-functionalized nanoparticles (DFPs) are a powerful platform for engineering reconfigurable materials with programmable interactions. Recent innovations have introduced switchability in these systems via mechanisms such as strand displacement, but these often suffer from limitations including irreversibility, environmental insensitivity, and finite cyclability. In contrast, our work introduces a strategy for environmentally responsive and thermally reversible switching in DFP-based lattices, exploiting the tunable hybridization properties of DNA. To efficiently navigate the vast design space of these multivalent systems, we implement a high-throughput virtual screening framework coupled with Bayesian optimization in an active learning cycle. This approach allows us to iteratively refine predictive models and focus computational sampling on promising candidates. Using this framework, we discover DFP systems that assemble into binary superlattices exhibiting two thermally accessible structural states. These bistable configurations, enabled by tunable DNA-mediated interactions, demonstrate reversible switching behavior under temperature control. Our re-

sults reveal emergent design principles and provide a path toward programmable, stimulus-responsive metamaterials. Looking forward, this approach sets the stage for closed-loop experimental-computational design of hybrid systems. Our framework is readily generalizable to other forms of switchability, such as pH, light, or chemical stimuli, and will be of interest to researchers in soft matter, materials science, and nanotechnology aiming to rationally design dynamic, responsive materials.

4.2 Introduction

Metamaterials are designed materials built from unit building blocks, each composed of more fundamental molecular motifs, that exhibit emergent properties at larger length scales. One particularly powerful motif is DNA, which has enabled the development of DNA-functionalized particles (DFPs): hybrid nanosystems in which single-stranded DNA (ssDNA) is grafted to the surface of nanoparticles (NPs), typically gold, via gold-thiol bonds. DFPs exploit the specificity and programmability of Watson–Crick base pairing to tune interparticle interactions with exceptional precision, enabling the bottom-up assembly of complex nanostructures. These systems have shown promise in a wide array of fields, from programmable photonic crystals and nanoscale circuitry to biomedical applications such as targeted drug delivery, thanks to their biocompatibility and molecular recognition capabilities.[194, 195, 196]

Foundational studies in the mid-2000s by Mirkin, Gang, and Alivisatos demonstrated that DNA could be used to reliably drive the assembly of nanoparticles into ordered superlattices with a variety of symmetries and tunable lattice parameters.[197, 198, 199] These early works established the use of DNA as a programmable linker and provided the blueprint for controlling not just composition and stoichiometry but also symmetry and lattice spacing. Since then, extensive progress has been made in both experimental and computational realms. Researchers have elucidated how parameters such as strand length, grafting den-

sity, spacer design, and sequence identity can be tuned to control both the thermodynamic ground state and the kinetic pathways of self-assembly.[200, 201, 202]

In particular, the introduction of multiflavoring, in which particles are grafted with multiple DNA strands of varying complementarity, has emerged as a powerful technique for decoupling like-like and unlike-pair interactions, dramatically expanding the accessible design space of binary and ternary superlattices.[17] Recent studies have leveraged these interactions to generate diverse 2D and 3D lattice symmetries, including square, hexagonal, Kagome, CsCl, and CuAu structures.[17, 203, 204] These architectures can be stabilized or transformed by tuning enthalpic contributions (i.e., binding strength asymmetries) or introducing thermal and environmental perturbations.[194, 195, 196]

A particularly compelling direction has been the pursuit of thermoresponsive reconfiguration wherein a single DFP system transitions between stable lattice states as a function of temperature. This behavior, known as diffusionless or martensitic transformation, has been observed in both experiment and simulation. Recent work by Mao et al. demonstrated that DFP systems can exhibit reversible transformations between FCC and BCC lattice types, governed by crystallite size and temperature rather than nucleation kinetics.[194, 195, 196] This opens the door to truly reconfigurable metamaterials that can respond to external stimuli without requiring particle dissolution and reassembly. Developing this technology further has broad applications ranging from tuning photonic and mechanical properties through lattice modulation to acting as a thermally-controlled molecular sieve for either targeted drug delivery or membrane-like gating.

Despite these advances, existing models often rely on ultra-coarse-grained representations of entire DFPs as single beads, limiting insight into the microscopic features that govern hybridization and structural dynamics. Moreover, many prior reconfiguration strategies, such as those based on DNA strand displacement, face limitations in reversibility due to the accumulation of displaced strands within the lattice.[194, 195, 196, 197] These designs

tend to permit only one-time switching or require system flushing and resetting to restore functionality.

Our work aims to address these challenges by developing simulation models and active learning strategies to identify and optimize DFP systems capable of reversible, thermally controlled switching between structurally distinct states. This approach combines molecular insight with scalable optimization, setting the stage for rational design of responsive metamaterials that are both physically grounded and practically realizable.

Here, we propose an alternate designed system mechanism in lieu of strand displacement. This mechanism utilizes a complimentary A-B DFP scheme with two types of functionalized DNA (fDNA) of different lengths per nanoparticle. We test our mechanism with a computational active learning cycle designed to efficiently identify candidates, which not only exhibit the desired thermoresponsive behavior but are in fact the systems with the most optimal response, in a space that is intractably large to exhaustively simulate. We constrain our system parameters to include the lengths and grafting density ratio of the short and long fDNA as well as vary the hybridizing sequence of the short fDNA, limiting it to only A and T nucleobases with a length in the range [197, 204]. The constraints reduce the design space to roughly 1.65×10^6 unique states, which, while better, is still clearly an intractably large space to sample exhaustively. We leverage Bayesian optimization methods to find the most optimal systems while sampling less than 0.1% of the space. We synthesize and experimentally validate our identified top candidates while further improving our high-throughput computational model by supplementing it with an additional low-throughput/high-fidelity experimental stream of data.

This work primarily contributes to the efficient optimization of thermoresponsive DFP superlattices, which have direct applications in biomedicine and nanoelectronics. Additionally, the post-hoc analysis of the simulation data enables design principles to be uncovered. This work is, to the best of our knowledge, the first to extend a formal active learning

based approach to these DFP systems which hold such immense possibilities for fine-grained rational design of metamaterials.

The structure of the remainder of this chapter is as follows. In Section 2 we discuss our computational methods and in Section 3 we present the current results from our active learning cycle and expect convergence of later rounds in the following month. We then discuss the efficacy of the model to identify top candidates and perform post-hoc analysis to ascertain design principles. In Section 4 we share our conclusions and future directions for this work.

4.3 Methods

4.3.1 Thermoresponsive DFP Mechanism

The switching mechanism underlying the thermoresponsive behavior of DNA-functionalized nanoparticle (DFP) superlattices is conceptually straightforward yet chemically powerful. Each DFP is grafted with both short and long single-stranded DNA strands, and the superlattice is stabilized by intermolecular hybridization between complementary strands on neighboring particles (Fig. 4.1 a). The two different strand lengths can be thought of as concentric radii that define the bistable lattice constants (Fig. 4.1 b). Critically, the short and long strands are designed to differ in thermodynamic stability: the short strands are composed entirely of adenine–thymine (A–T) base pairs, which form weaker interactions, while the long strands incorporate only guanine–cytosine (G–C) base pairs to increase their melting temperature. At low temperatures, the enthalpic gain from maximizing hybridization leads to the formation of a tightly connected superlattice, with both short/short and long/long hybridizations contributing to particle-particle contacts. As the temperature increases, however, the less thermostable short/short interactions begin to dehybridize while the long/long connections, protected by their stronger G–C base pairing, remain hybridized

over this temperature range and preserve the structural integrity of the superlattice. Even more critical than the engineered enthalpic difference between the long and short strand hybridizations is the entropically-motivated imbalance in the relative grafting density of the short and long strands. By operating in a regime where the short strand grafting density is always greater than the long strand grafting density, the greater number of entropic contributions from all short strands relative to all long strands for any given DFP shifts the balance so that the hybridization state of the short strands is the dominant contributor to the system entropy. At higher temperatures the dehybridization of the short strands, which is entropically favorable by increasing the number of available conformational states for the now-unbound short strands, can be substantial enough to change the overall system free energy. In other words, in accordance with the Gibbs free energy equation (Eq. 4.1) when the temperature is low the free energy of the system is dominated by the enthalpic contributions of the short strand hybridizations, but at higher temperatures the entropic benefit from short strand dehybridization is the dominant contributor to the free energy of the transition.

$$\Delta G = \Delta H - T\Delta S \begin{cases} T_{Low} : \Delta G_{Expansion} > 0 \\ T_{High} : \Delta G_{Expansion} < 0 \end{cases} \quad (4.1)$$

The net result is a thermally-driven, reversible expansion in the DFP center-to-center distance (Fig. 4.1 c,d), as the system transitions from a configuration in which both strand types are hybridized to one in which only the long strands maintain contact. This shift redefines the lattice constant of the structure from being approximately the length of the short strand hybridizations to the length associated with the long strand hybridization and serves as the physical basis for the thermoresponsive behavior observed in simulation and experiment.

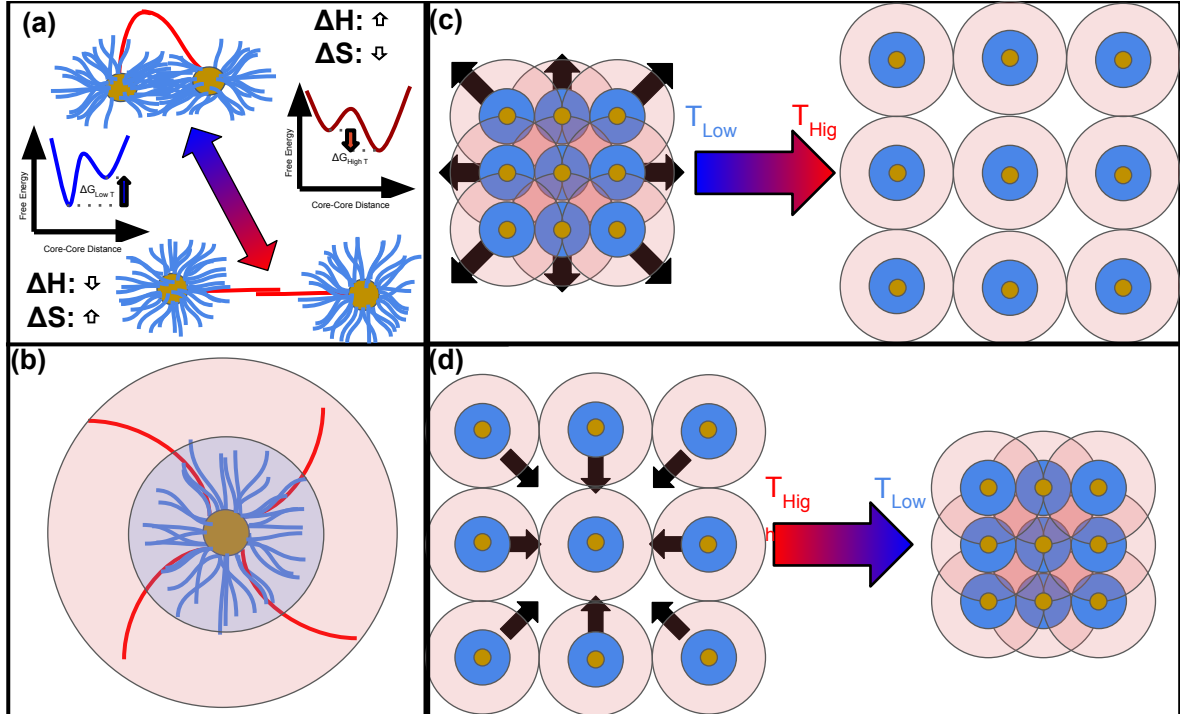


Figure 4.1: The proposed novel switching mechanism for DFP systems. In (a) we show a dimer representation of the DFPs and note the opposing entropic/enthalpic trends for transitioning between two states. In (b) we represent the coarsening of the DFPs with semi-opaque circles corresponding to the bi-length functionalized DNA, which we use to depict the process of (c) DFP lattice expansion and (d) compression.

4.3.2 CG-DNA Simulations

All molecular simulations were conducted at $T = 0.25 \approx 415K$ and utilized expanded-ensemble On-the-fly Probability Enhanced Sampling (OPES) [205] to accelerate convergence of a 1-dimensional, multi-thermal free energy surface (FES) along the collective variable (CV) described by the center-of-mass distance between nanoparticle centers. Simulations were performed using the LAMMPS June 23, 2022 MD engine [206] patched with the pair potentials from the CG-DNA forcefield developed by the Mittal group [207] and were conducted in LJ units. The characteristic energy, $\epsilon \approx 3.3$ kcal/mol, and lengthscale, $\sigma \approx 0.55$ nm, were established from empirical measurements of the pairing of CG nucleobases. Simulations were run on 128 AMD EPYC-7702 CPU cores to simulate 32 parallel walkers, each with 4 cores, achieving execution speeds of $\approx 1.25\mu s$ per day.

The PLUMED v2.8.0 software suite [8] was used to perform enhanced sampling, specifically expanded-ensemble OPES, in all production runs. The distance between the center points of each nanoparticle was used as the CV for umbrella-style acceleration while the internal energy of the system was used to sample a multicanonical distribution of temperatures ranging from $T=0.15$ to $T=0.35$ with $T=0.25$ as the thermostat temperature.

Assessing Convergence Several convergence metrics are utilized to assess each production OPES run. First, the center-of-mass distance CV distribution must show diffuse sampling across the CV range covered by OPES umbrella potentials. Secondly, the convergence of the RCT score, a measure unique to OPES but similar in concept to the convergence of well-tempered metadynamics simulations towards a maximum energy, implies the convergence of the CV space, i.e. no new CV space is being explored. Finally, the average bias of the 32 walkers is plotted as a function of time and convergence is indicated by a quasi-static bias.

4.3.3 *Evaluating System Fitness*

Hybridization Analysis Tool While the computationally efficient 2-bead DNA model [202] is necessary for high-throughput screening, the simplicity of the model does not account for the directional 5'-to-3' component of DNA hybridization and thus single strands can hybridize in both parallel (illegal) (Fig. 4.2 a) and anti-parallel (legal) interactions (Fig. 4.2 b). Rigorously addressing this at the level of the coarse-grained model is a nontrivial task so we implement a simple work around to account for this behavior. Noting that the unphysical DNA binding is most prominent at very close nanoparticle COM distances (Fig. 4.2 c), we have created an in-house tool, the Hybridization Analysis Tool (HAT), to analyze these DFP trajectories and determine a minimum cutoff distance where more than 80 percent of the bonds are physical. We then discard the near-field part of the FES containing the additional illegal hybridizations, which would otherwise serve to artificially stabilize the FES at small distances, and calculate $\Delta\Delta G$ values using the intermediate and long-range portions of the FES. These regions are effectively decoupled from the close-range interactions that suffer from spurious hybridizations and thus aren't significantly affected by the possibility of this behavior. In this way, we maintain the critical high-throughput nature of our initial screening while mitigating effects from using a simpler (and orders of magnitude faster) force field compared to more sophisticated options such as the 3SPN model [208]. Finally, we plan to subject our top identified candidates to additional, more expensive, simulations with a higher resolution force field.

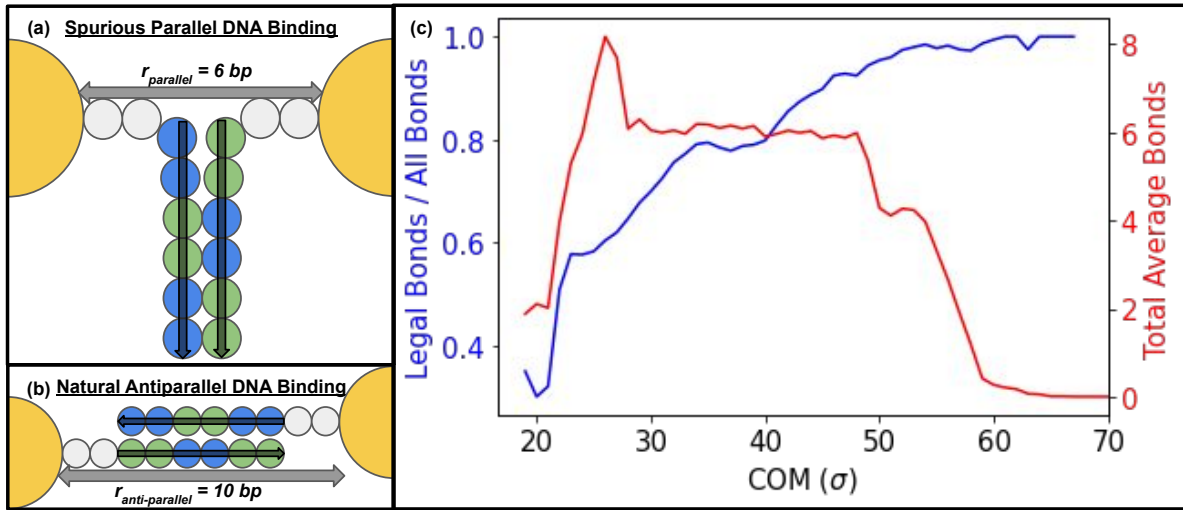


Figure 4.2: We show a shortcoming of the forcefield to account for directionality in DNA binding and the resulting parallel (a) and anti-parallel (b) hybridization modes. In (c) we show that unphysical parallel binding is only prevalent at short DFP separations and that we only use DFP separations with $>80\%$ allowed hybridization.

Calculating $\Delta\Delta G$ It is desirable to identify candidates that demonstrate a large energy difference between the short-bound and long-bound states at low temperature but a small (or even negative) difference between those states at higher temperatures (Fig. 4.3 a). To this end we evaluate the FES of the OPES simulations along the nanoparticle COM distance through $\Delta\Delta G$ by first computing the ΔG between the short-bound and long-bound states at a cold temperature ($T=0.16$) as well as at a hot temperature ($T=0.20$), and then $\Delta\Delta G = \Delta G_{cold} - \Delta G_{hot}$ where the best candidate systems are expected to have large $\Delta\Delta G$ values (Fig. 4.3 b) and the active learning loop is utilized to identify these top candidates in the vast design space.

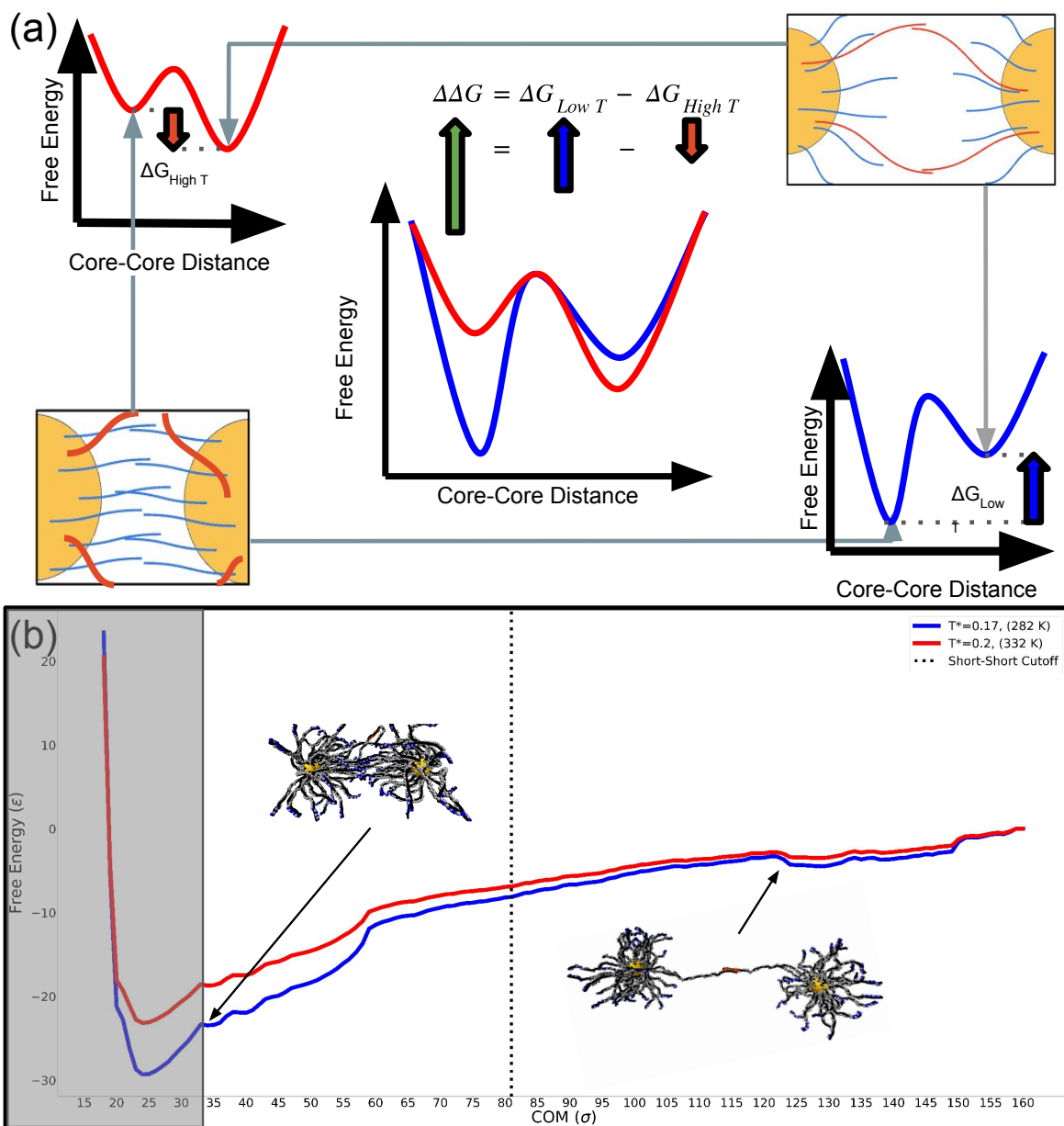


Figure 4.3: In (a) we demonstrate leveraging system thermodynamics to modulate the DFP FES as a function of temperature and our corresponding proxy evaluation of $\Delta\Delta G$ for switchability. In (b) we provide an example from our active learning campaign for this process.

4.3.4 Surrogate Modeling with Bespoke Molecular Similarity Kernel

We incorporate information about all parameters of the system through a 40-dimensional vector where the first four values correspond to the short spacer length, long spacer length, short spacer grafting density, and sequence length successively and are minmax normalized. The remaining 36 elements of the vector are used for a one-hot encoding of the 12 bp where $[0,1,0]$, $[1,0,0]$ represent A and T nucleobase beads, and $[0,0,1]$ is used for padding sequences that are less than 12 bp long.

Kernel PCA The combination of ordinal and nominal data in our dataset necessitates the use of a custom kernel function (Eq 4.2).

$$K(x_{\alpha\beta}, x'_{\alpha\beta}) = \exp(-\gamma T_{Distance} \times T_{Distance}(x_{\beta}, x'_{\beta}) + \sum_{i=1}^{i=5} -\gamma_i \|x_{\alpha} - x'_{\alpha}\|^2) \quad (4.2)$$

We choose to use standard RBF kernels for each of the four ordinal data values, each with their own optimized lengthscale, and a modified RBF kernel for the nominal sequence data which employs the Tanimoto distance (Eq 4.3)

$$T_{Distance} = 1 - T(A, B) \quad (4.3)$$

where the Tanimoto distance can be seen as a simple extension of the more common Tanimoto similarity (Eq 4.4) between sequences.

$$T(A, B) = \frac{A \cdot B}{\|A\| + \|B\| - A \cdot B} \quad (4.4)$$

These six kernels are multiplied together to produce a kernel space that contains information on all system parameters, which we use both as the kernel for our GP surrogate model and as the basis for Kernel PCA to help with visualization and interpretation. The leading eigenvalues for the KPCs are rank-ordered in magnitude and importance to the model and

we find the top 15 KPCs to be sufficient, accounting for more than 95% of the cumulative eigenvalue total. Analysis of the embedding space illustrates local order with regard to both system features and predicted $\Delta\Delta G$ values (Fig. 4.4)

4.3.5 *Bayesian Optimization*

We tested a variety of different acquisition functions such as the quasi-Monte Carlo Expected Improvement (qEI) and Upper Confidence Bound (qUCB) however the sparsely explored space as well as the presence of an outlying high performing system resulted in poor candidate diversity (exploration) which is particularly important in the early rounds of active learning. For these reasons we ultimately chose to employ Thompson Sampling as the basis for our Bayesian optimization because the exploratory emphasis is inherently baked-in to the stochastic process of sampling from the posterior of the model. We selected K new candidates each round for simulation, where the batch size of K was chosen for the maximum number of simulations that we could perform in parallel. The active learning cycle is considered converged when the performance difference and stabilizing prediction metrics indicate convergence. We also developed a Python library for MD-based active learning campaigns, **Active Learning Pipeline For Optimal Ranking Estimation** (ALPineFOREst), which serves as a framework for facilitating the interconnection of common active learning libraries (Fig 4.5).

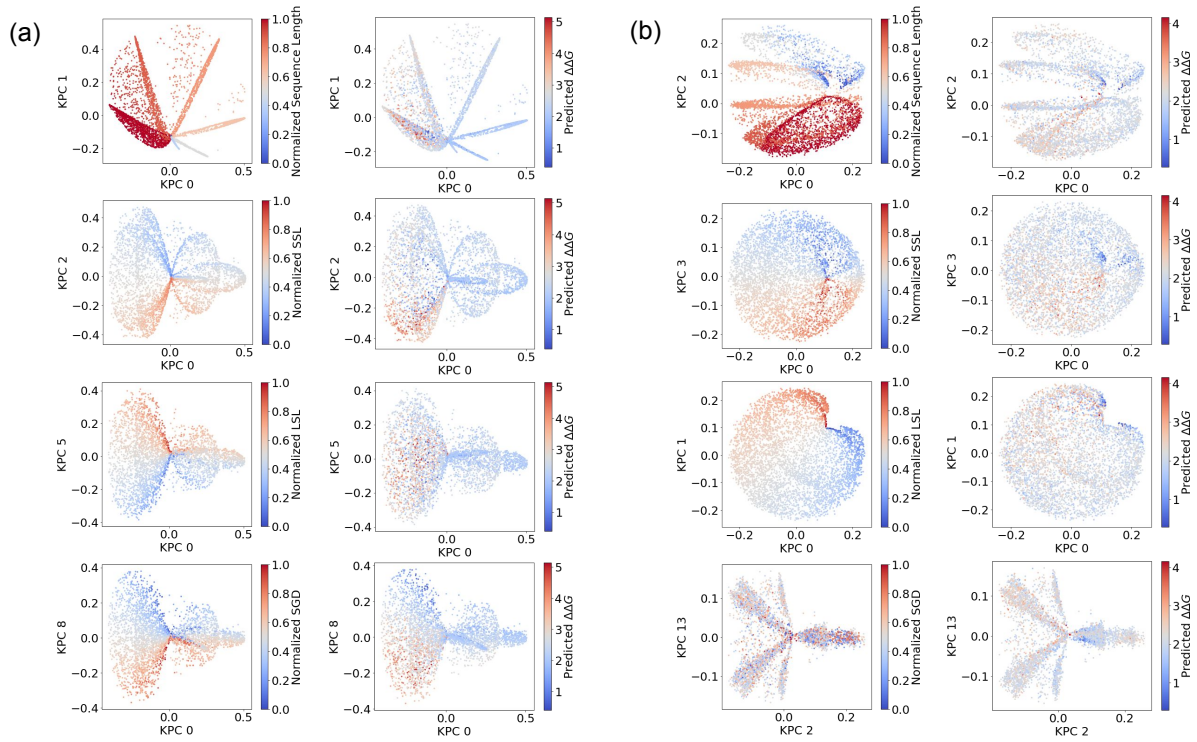


Figure 4.4: The KPC similarity space shows strong organization into meaningfully local regions from system features, and predicted $\Delta\Delta G$ values illuminate the impact of the features on the system's switchability for the initial dataset (a) and the first round of active learning (b).

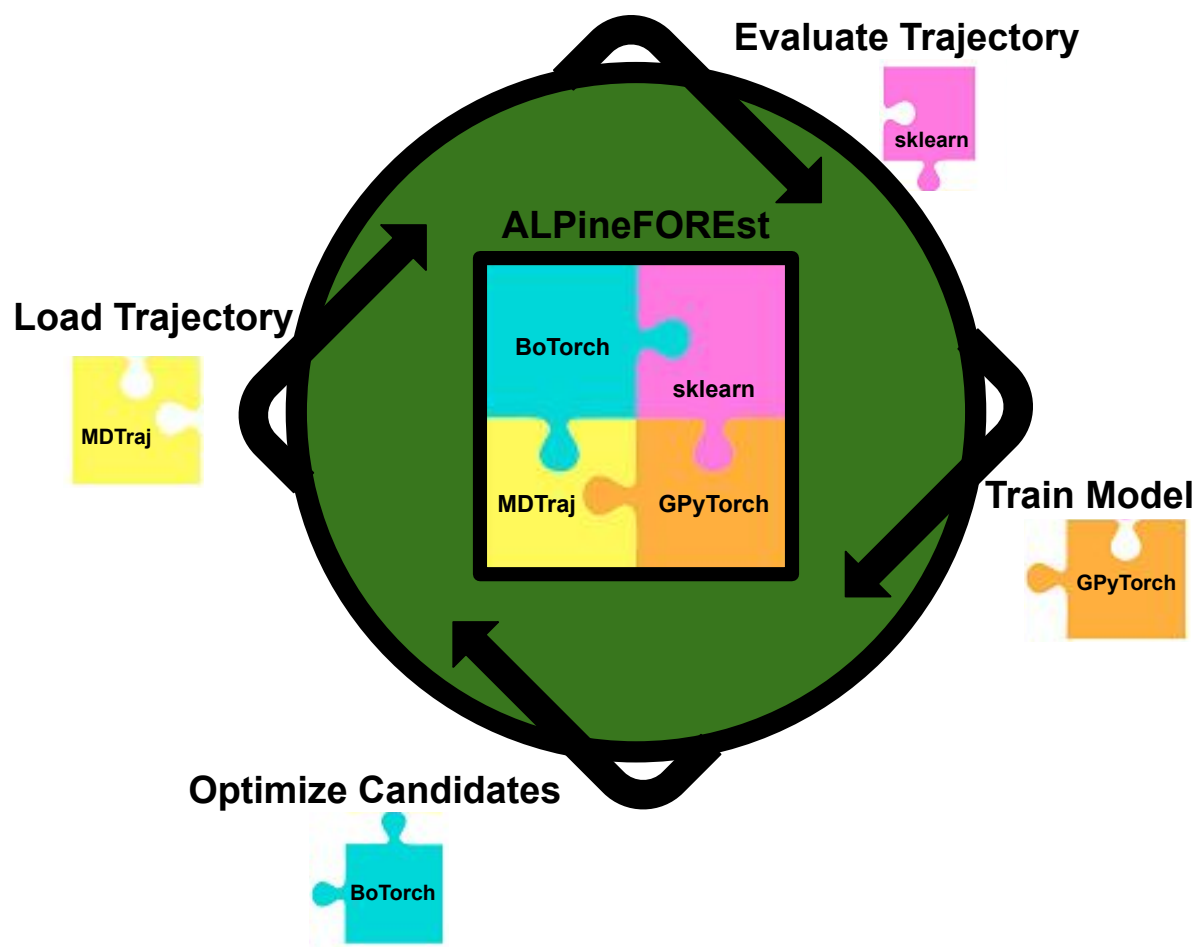


Figure 4.5: ALPineFOREst as a framework for connecting active learning python libraries for MD campaigns.

4.4 Results and Discussion

At the time of writing, the DNA-Functionalized Nanoparticle (DFP) active learning campaign is ongoing and submission for publication is expected in the month of August, 2025. The system has been successfully initialized and has completed two rounds of the ALPineFOREst pipeline: an initial exploration round and the first model-guided active learning round. During these stages, free energy proxies ($\Delta\Delta G$) were computed for a curated subset of candidate strand architectures, serving both as training data and as a performance baseline for the learning model.

Due to the computational cost of free energy estimation and the vast number of possible systems, only a small fraction of the full design space has been sampled thus far. Nevertheless, the pipeline has demonstrated full functionality across its modular components, including structure generation, biasing simulations, collective variable evaluation, Gaussian Process model training, and candidate selection using Thompson Sampling. These early stages validate the infrastructure and interoperability of the ALPineFOREst framework.

While the dataset is currently limited, preliminary analysis indicates a clear correlation between sequence length and the predicted $\Delta\Delta G$ values. Specifically, longer hybridization regions appear to support a higher ceiling for $\Delta\Delta G$, likely due to the increased number of possible base pairs contributing both to the stability of short strand hybridization at cold temperatures as well as the entropically-driven lattice expansion at warm temperatures.

As the campaign progresses, further rounds of active learning will enable more detailed mapping of the design landscape. These data will ultimately support a deeper understanding of the structural and thermodynamic determinants of temperature-responsive switching behavior in DFP superlattices.

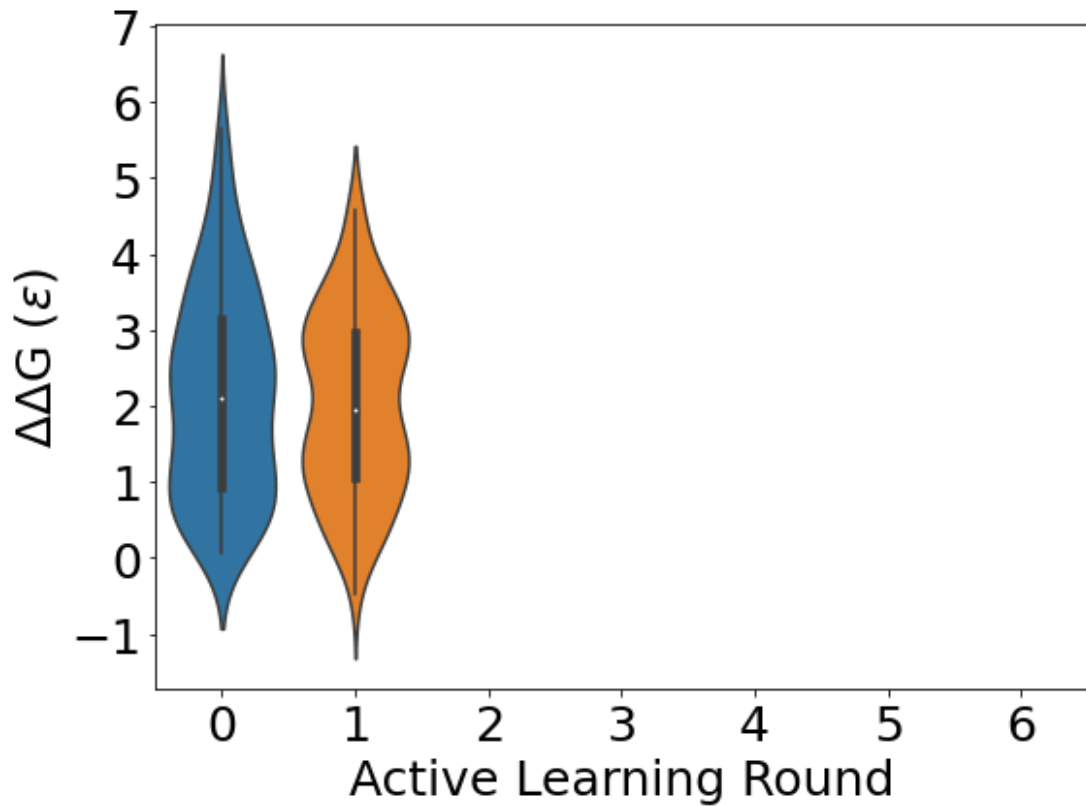


Figure 4.6: Violin plots by round of simulated candidates. The campaign is ongoing and will include further rounds, currently shown as whitespace without violin plots. Final results from this work are expected to be published in August 2025

4.5 Conclusions

In this work we led a computational active learning campaign in a vast chemical space of possible switchable, thermoresponsive DFP superlattices. Our approach centered on the idea of engineering competition between two classes of DNA-mediated interactions (short/short and long/long hybridizations) each with distinct thermodynamic properties. In particular, the hybridization sequence and the grafting density of the long and short strands were leveraged to tune the enthalpic and entropic contributions of the strands. This design creates a built-in switching mechanism: at low temperatures, both types of hybridizations are stable and contribute to a compact lattice but as temperature increases, short/short connections dehybridize, leading to a measurable change in lattice spacing while long/long interactions persist. We utilize Bayesian optimization to guide candidate selection within our on-going campaign and we show round by round violin plots for each batch of candidates (Fig. 4.6). Additionally, this work presented the development and application of a data-driven simulation workflow to investigate thermoresponsive behavior in DNA-functionalized nanoparticles (DFPs). By integrating enhanced sampling, molecular design heuristics, and active learning, we constructed a framework capable of identifying DFP superlattices that exhibit sharp, tunable switching behavior as a function of temperature. The results of this study will inform both the design principles for next-generation switchable nanomaterials and the methodological directions for applying machine learning frameworks to molecular simulation. Future work will explore incorporating experimental data in a multi-fidelity active learning cycle, implementing more sophisticated DNA force fields in the computational loop, and addressing other relevant environmental factors like ion concentration.

4.6 Acknowledgements

Work by N.S.M.H. was supported by the U.S. Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences (BES), as part of the Energy Frontier Research Centers program: CSSAS—The Center for the Science of Synthesis Across Scales—under Award No. DE-SC0019288. This work was completed in part with resources provided by the University of Chicago Research Computing Center. We gratefully acknowledge the University of Chicago for computing time on high-performance GPU-based cyberinfrastructure supported by the National Science Foundation under Grant No. DMR-1828629. We also gratefully acknowledge the resources awarded to us by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award BES-ERCAP0034608, as well as from the Explore allocation tier through the Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS) program, which is supported by U.S. National Science Foundation grants No. 2138259, 2138286, 2138307, 2137603, and 2138296.

CHAPTER 5

FROM PINES TO ALPINEFOREST: SEEING THE FOREST FOR THE TREES

This chapter is partially adapted from work published in the *Journal of Chemical Physics*.

*G.A. Tribello, M. Bonomi, G. Bussi, C. Camilloni, B.I. Armstrong, A. Arsiccio, S. Aureli, F. Ballabio, M. Bernetti, L. Bonati, S.G.H. Brookes, Z.F. Brotzakis, R. Capelli, M. Ceriotti, K.-T. Chan, P. Cossio, S. Dasetty, D. Donadio, B. Ensing, A.L. Ferguson, G. Fraux, J.D. Gale, F.L. Gervasio, T. Giorgino, N.S.M. Herringer, G.M. Hocky, S.E. Hoff, M. Invernizzi, O. Languin-Cattoën, V. Limongelli, O. Lopez-Acevedo, F. Marinelli, P.F. Martinez, M. Masetti, S. Mehdi, A. Michaelides, M.H. Murtada, M. Parrinello, P.M. Piaggi, A. Pietropaolo, F. Pietrucci, S. Pipolo, C. Pritchard, P. Raiteri, S. Raniolo, D. Rapetti, V. Rizzi, J. Rydzewski, M. Salvalaglio, C. Schran, A. Seal, A.S. Zadeh, T.F.D. Silva, V. Spiwok, G. Stirnemann, D. Sucerquia, P. Tiwary, O. Valsson, M. Vendruscolo, G.A. Voth, A.D. White, and J. Wu “PLUMED Tutorials: A collaborative, community-driven learning ecosystem” *J. Chem. Phys.* 162 092501 (2025) [<https://doi.org/10.1063/5.0251501>]*

In computational research, software is often treated as a means to an end, hastily written to generate results then abandoned or rewritten for the next project. Over time, this approach leads to a bloated and brittle ecosystem of scripts, notebooks, and ad hoc tools: barely functional when first written, and nearly impossible to understand or reuse months later. The proliferation of these one-off solutions is a form of scientific pollution and a by-product of working in haste and in isolation, without consideration for sustainability or reproducibility.

This chapter is about a different approach, one that strongly echoes the philosophy of F.A.I.R. data principles [209, 210, 211] applied to the development of computational methods, but which chronicles my specific journey from writing narrowly-scoped, fragile scripts to building general, reusable, and performant research software. The chapter centers on two projects: Permutationally Invariant Networks for Enhanced Sampling (PINES), a molecular descriptor-based module for custom pairwise interactions, and the Active Learning Pipeline For Optimal Ranking Estimation (ALPineFOREst), a cohesive framework for active learning in molecular simulations. PINES addresses the need for discovering symmetry-adapted CVs and shows immediate value in sampling solvated states thoroughly in addition to future applications simulating chemical processes involving indistinguishable particles and both global as well as local system descriptors. ALPineFOREst facilitates active learning campaigns in the functionalized nanoparticle space and opens the door to providing a general framework for molecular active learning projects. Together, these tools reflect my growing belief that thoughtful software design is not just an engineering preference but a scientific necessity.

In what follows, I reflect on the software practices that guided my refactors and rewrites, and how design choices made at the individual level scale up to impact group productivity and community reuse. I argue that software in science should be developed with a focus on researcher-friendly interfaces, designed modularity, and long-term utility to maximize the utility to the scientific research community. PINES is now in active use across multiple projects within our group and has also been submitted to the PLUMED enhanced sampling suite, has contributed to a recent tutorial article, and is now supported by new documentation and user guides available on the PLUMED website. ALPineFOREst is freely available to pip install from PyPi and is expected to become a critical tool for active learning projects in the group moving forward.

5.1 Reduce and Reuse, Don't Recycle

Recycling old code is often celebrated as a practical and efficient habit aimed at avoiding reinvention. But in research computing, recycling often takes the form of copying, pasting, and modifying fragile code without truly understanding or improving it. Over time, this leads to a proliferation of barely functional duplicated scripts with uninformative names, many of which contain hard-coded paths, magic constants, or outdated assumptions. In this environment, small changes become costly, collaboration becomes difficult, and progress slows.

Instead, I advocate for a different mantra: reduce and reuse. Rather than recycling dozens of redundant scripts, we should reduce the complexity of our codebase by identifying common functionality and packaging it in reusable modules. This practice not only improves clarity and maintainability, but also enables scaling and sharing, thereby making it easier to validate results, onboard collaborators, and apply methods across projects.

5.1.1 *Respect the Environment*

One of the most immediate barriers to reproducible research is the software environment. Differences in package versions, dependency trees, or compiler flags can derail a workflow that “worked perfectly” on another machine. This is especially daunting for early-career researchers, who may be encountering environment issues for the first time without institutional support.

To address this, I find that it is critical to develop robust environment management strategies such as `conda` environments, version-pinned `requirements.txt` and `environment.yml` files, and cluster-specific `module` files. While these do not eliminate the complexity of scientific software, they formalize it, enabling others (and future me) to reconstruct the computational context in which results were generated.

Environment issues are often treated as an afterthought in research code but in practice,

they are often the first obstacle to reuse. Respecting the environment is a prerequisite for reproducible, shareable science.

5.1.2 *Quick and Dirty Solution Pollution*

In many labs, rapid prototyping is both necessary and encouraged. Novel ideas are often tested by writing short scripts, copying from previous projects, and modifying as needed. While this approach can be productive in the short term, it creates long-term entropy in the codebase: poor documentation, inconsistent logic, and hidden assumptions that make it difficult for anyone else to understand or build on the work.

The problem is not that quick-and-dirty scripts exist, it's that they often become permanent infrastructure by default. Without time or incentives to refactor them into generalized tools, these scripts persist, spread, and become the basis for future projects. The result is a code ecosystem that is fragile, non-reproducible, and non-collaborative.

Good software design can prevent this. By treating tools as first-class research products worth documenting, testing, and sharing we reduce solution pollution and encourage a healthier research environment. As the later sections of this chapter show, even modest investments in generalization and modularity pay dividends when building reusable infrastructure for simulation and machine learning workflows.

5.2 PINES

The PINES module was initially developed as a proof-of-concept for enhanced sampling based on local atomic environments. Its early success came at the cost of long-term usability: assumptions baked into the code made it difficult to extend beyond its original application. The architecture coupled data structures directly with control logic, lacked modular components, and relied on hard-coded interaction rules, leading to what I later recognized as significant research tech debt.

As new scientific questions pushed PINES beyond its original scope, I undertook a comprehensive refactor guided by three core principles: modularity, generality, and performance. The refactored version treats PINES as a framework rather than a fixed tool, enabling a platform for building custom pairwise interactions driven by molecular descriptors and evaluated under user-defined tolerances. Two major architectural decisions exemplify this shift.

First, I replaced the brute-force pairwise interaction evaluations with a heap-based scheme for pair list generation. In this design, each atom maintains a max-heap of candidate partners ranked by interaction strength. This makes it possible to dynamically prune the pair list, retaining only the top k interactions or those exceeding a specified threshold. These heaps are updated incrementally across timesteps, allowing the system to flag and update only interactions that become stale, i.e. that have changed beyond a tolerance threshold. Each block of atoms maintains its own update counters, enabling efficient, parallel pairlist updates. This change led to significant performance gains in both large systems and high-frequency evaluation regimes.

Second, I introduced a flexible filtering system based on molecular descriptors. The original version of PINES relied solely on atom types to define interaction partners, limiting the ability to model complex or coarse-grained systems. The new framework allows users to define filters based on atom type, atom indices, and resids. These filters can be combined using logical operators and applied during pair list construction, enabling precise control over which atoms interact. This capability made it possible to model coarse-grained DNA systems with directional hybridization logic, exclude intramolecular interactions, and include solvent ions selectively without rewriting core logic.

Together, these changes mark a shift in my approach to scientific software development. Rather than writing bespoke scripts to solve specific problems, I now aim to build flexible, reusable infrastructure that enables new problems to be posed. PINES, once a narrowly-scoped tool, has become a general-purpose interaction engine designed to support diverse

applications in enhanced sampling. It has already enabled the interrogation of vacancy formation and mobility in silicon-carbide crystals as well as consideration of the structural ordering for water in regards to the Mpemba effect, and its contribution to PLUMED further gives other researchers easy access to a data-driven, symmetry-adapted CV discovery method.

5.3 ALPineFOREst

ALPineFOREst (**A**ctive **L**earning **P**ipeline **F**or **O**ptimal **R**anking **E**stimation) was designed to support simulation-driven active learning at scale. Its core premise is that scientific pipelines should be engineered with the same care and modularity as production software. Too often, active learning workflows are glued together from scattered scripts written for a single experiment, leading to brittle systems that are hard to maintain, reuse, or scale. ALPineFOREst aims to break this cycle by offering a reusable framework structured around three core principles: modularity, scalability, and reusability.

Each of the four stages of the pipeline, simulation, evaluation, modeling, and optimization, is implemented as a self-contained module with a clean interface. These components can be independently replaced, configured, or extended without disrupting the surrounding logic. This modularity allows the pipeline to adapt to evolving project needs while maintaining internal consistency and compatibility.

To address the computational demands of modern simulation campaigns, ALPineFOREst integrates scalable methods such as batched kernel computation, precomputed Gram matrices, and Nyström approximations. These enable the use of Gaussian Process models on datasets that would otherwise exceed memory limits. Additionally, common utilities, such as free energy reweighting, kernel evaluation, and data filtering, are implemented as reusable helpers, allowing rapid prototyping without redundancy.

A key aspect of ALPineFOREst’s design is what I refer to as either active or passive

development. Active development focuses on building the core infrastructure: designing clean APIs, managing dependencies, and formalizing interfaces. Passive development happens when solving project-specific challenges, such as implementing a hybrid Euclidean/sequence kernel, results in reusable modules added back into the library. By building these features within the existing architecture rather than around it, I ensured that ALPineFOREst grew organically into a library rather than a pile of scripts.

This design philosophy stems directly from lessons learned in my earlier work on PINES, where a narrow, expedient design led to difficult refactors later. With ALPineFOREst, I made an upfront investment in modularity to reduce future tech debt, enable team contributions, and support long-term evolution.

Crucially, this modularity doesn't just support maintainability, it enables automation. When each module communicates through predictable formats and standardized interfaces, it's straightforward to chain them together into automated pipelines. This makes the software suitable for both interactive experimentation and high-throughput workflows.

In our active learning campaigns, ALPineFOREst operated across more than five supercomputing clusters using Globus for data transfer and SLURM for job scheduling. The pipeline launched hundreds of simulations in parallel, analyzed them automatically, and fed the results back into model training and candidate selection. These workflows would have been prohibitively complex to manage without a well-defined software backbone. Instead, ALPineFOREst facilitated rational exploration of DFP system parameters and offers seamless extensibility to ELP-functionalized NPs, a future direction of interest for the CSSAS EFRC.

5.4 Conclusions

The work described in this chapter reflects not only a shift in my perspective on scientific software development but also the creation of two tools that I believe contribute meaningfully

to the broader scientific enterprise.

PINES began as a highly system-specific implementation, but its growing generality and eventual integration into PLUMED marked a turning point. Through rigorous refactoring and modularization, PINES evolved into a flexible, production-quality module that now enables force reconstruction of arbitrarily precise pairwise interactions and implements a novel neighbor list type that efficiently maintains a tolerance-guaranteed ranked list of pairwise interactions for any molecular system and group definition, functionality that was not previously accessible in PLUMED. Its incorporation into the core codebase makes it available to a global user base and opens the door for new force-learning, enhanced sampling, and post hoc analysis methods.

In contrast, ALPineFOREst was designed from the ground up with modularity and extensibility as core principles. Rather than treating active learning as a fixed pipeline, I developed ALPineFOREst as a framework for integrating machine learning into molecular simulation workflows. It has already supported large-scale campaigns across multiple supercomputing clusters and is now being adopted by other researchers in our lab to enable system-specific discovery, including drug-like molecule design and nucleic acid engineering. Its flexible design and clean interfaces have lowered the barrier for others to apply active learning to new domains.

Ultimately, these tools are more than technical artifacts, they are products of scientific value. Good software practices are not merely engineering concerns; they are scientific enablers. Well-designed code amplifies reproducibility, encourages collaboration, and expands what is computationally feasible. By investing in sustainable software at the level of the individual, the research group, and the broader community, we shape the science we are capable of doing. PINES and ALPineFOREst each represent steps in that direction as tools that both reflect and advance the evolving relationship between scientific inquiry and software development.

CHAPTER 6

CONCLUSION

6.1 Contributions to the field

This dissertation presents a cohesive research program that advances the application of machine learning in molecular simulation, not merely through the development of new software tools, but by enabling novel discoveries and methodologies in physical chemistry. Each chapter builds toward a central goal: to accelerate and deepen our ability to explore complex molecular systems by integrating learning-based methods with simulation-driven science. The core contributions lie not only in the tools we’ve developed, but in what they made possible both scientifically and methodologically.

In Chapter 2, we learned design rules for self-assembling peptide-conjugated perylene diimide (PDI) molecules by elucidating the rich relationship between peptide sequence, supramolecular structure, and the resulting optoelectronic properties. Through all-atom molecular dynamics simulations of the self-assembly of π -conjugated peptide dimers, we quantified the effects of small changes in side-chain composition on stacking geometry, dimerization free energies, and chiral twist distributions. These insights validated experimental findings and offered a mechanistic rationale for the observed differences in photophysical response. This work demonstrates how computational modeling can provide predictive, atomistic understanding of sequence-structure-function relationships in π -conjugated peptide materials.

In Chapter 3, we introduced PINES, a new approach to discovering CVs that respects permutational symmetry. This longstanding challenge in enhanced sampling has historically limited the transferability of CVs between chemically similar systems. By combining local atomic environments with a symmetry-aware pair list sorting scheme, PINES provides a robust solution to this problem. Using PINES, we uncovered transferrable, low-dimensional

CVs for three chemically distinct systems: a solvated NaCl pair, a simple LJ-fluid, and a solvated alkane chain. In each case, PINES learned symmetry-adapted CVs strongly correlated with known high-quality handcrafted system CVs and recovered FES in agreement with prior work in the literature. Furthermore, it furnished more comprehensive sampling of solvation structure compared to traditional CV heuristics and discovered an additional subtle alkane CV correlated with the chain’s fraction of helical pitch. These results suggest that PINES could provide a unique approach towards elucidating computationally challenging processes, such as nucleation, by identifying arbitrarily complex symmetry-adapted CVs generated from simulation data. Traditional methods often have great difficulty with rare-event sampling and the necessity of handling both global and local system descriptors simultaneously, so PINES stands poised to have a substantial impact in extending the capabilities of the field.

In Chapter 4, we applied these ideas to the design of thermoresponsive DNA-based nanoparticle superlattices. Here, the goal was to identify sequences and strand architectures that enable a central property to many programmable material applications: reversible assembly and disassembly in response to temperature. The primary challenge was combinatorial in that the space of candidate systems is vast, and conventional simulation methods are too slow to evaluate more than a small fraction. To overcome this, we integrated coarse-grained molecular dynamics with Bayesian optimization in an active learning loop. Our strategy successfully identified the top-performing designs after simulating less than 0.1% of a one-million-element design space. This work introduced a principled, data-driven framework for materials design, one that combines physical modeling with efficient sampling of compositional variables. The discoveries made here, such as identifying the optimal hybridization asymmetry and strand composition for switching, would not have been feasible with Edisonian screening or intuition-guided search alone.

In Chapter 5, we developed ALPineFOREst, a modular pipeline that formalizes this active learning cycle and extends it into a general framework for simulation-guided molecular

design. We demonstrated its application to an active learning campaign of DFP systems, which are of general interest to my collaborators at CSSAS for their immense promise as reconfigurable materials, and emphasized that the framework is readily adaptable for extension to ELP-functionalized NPs. While ALPineFOREst includes novel software infrastructure, its importance lies in how it enables reproducible, scalable, and domain-specific workflows for learning in the loop with simulation. By designing the pipeline with a robust structure from the start, we created a platform that supports both flexible experimentation and high-throughput campaigns across distributed computing environments. This work advances physical chemistry both by contributing a bespoke tool for investigating and evaluating DFP supramolecular assembly as well as by providing a general framework to lower the barrier for researchers to integrating machine learning into simulation campaigns. Furthermore, we discussed several significant improvements that we’ve made to the PINES code since its initial publication and the impact of this tool on the scientific community as a contribution to PLUMED.

Taken together, the contributions in this dissertation demonstrate a shift in how we approach simulation-based research in molecular science. Where traditional methods often relied on static CVs or exhaustive sampling, we introduced adaptive, learning-driven techniques that respond to the structure of the system and the needs of the problem. Where software was once treated as incidental, we showed that principled software engineering grounded in chemical intuition can expand what is scientifically possible. The resulting tools and methodologies enable researchers to answer questions that were previously out of reach about configurational mechanisms, optimal material designs, and the effective use of computation in chemical discovery.

In a broader sense, this dissertation contributes to the emerging paradigm of data-driven physical chemistry, one in which simulation, machine learning, and molecular design are not siloed disciplines, but parts of a single, unified workflow. By combining methodological

innovation with applied discovery, we offer both new insights into specific systems and a general approach that others can adopt and build upon. The field is moving rapidly, but with tools like PINES and ALPineFOREst, we are better equipped to keep pace and to shape the questions we ask in the first place.

6.2 Directions for the Future

6.2.1 *Applications for PINES and Further Developments*

This section outlines three key directions for future development and impact of the PINES framework: (i) innovations already implemented since its initial publication, (ii) proposals for future technical improvements that would enhance its usability and performance, and (iii) anticipated scientific applications where PINES can help address open questions in molecular design and engineering.

(i) Recent innovations. Since the initial release of PINES (v0.1.0), a number of infrastructural and algorithmic advances have significantly broadened the method’s applicability. Chief among these is the introduction of flexible group definitions for each PIV block. In earlier versions, PINES operated on homogeneous atomic groups, restricting its use to systems like Lennard-Jones fluids or neat water. With recent updates, users can now define chemically or spatially meaningful groupings, such as individual DNA strands, residues, ions, or solvent shells, tailored to the molecular features of interest. This enables the construction of localized, chemically interpretable CVs and allows PINES to operate effectively in heterogeneous, multicomponent systems.

This expanded group handling opens the door to a wider class of physical systems, including mixtures of indistinguishable species such as electrolyte solutions and biomolecular environments. For example, PINES can now be applied to ionic fluids where solvent molecules and ions are both present, enabling the identification of ion-specific collective behaviors and

solvent-mediated effects.

(ii) Future technical innovations. While the core framework of PINES is now general and modular, several targeted improvements could substantially enhance its accessibility and impact. One such extension is the automatic selection of switching function parameters. Currently, users must manually tune distance cutoffs and tolerance thresholds, often relying on heuristics. A procedure to infer these parameters from a short unbiased trajectory, based on characteristics such as interatomic distances or radial distribution features, would lower the barrier to adoption and improve robustness across use cases. While PINES will soon be available as part of the upcoming PLUMED2 release, another valuable future development would be porting the method to additional packages such as SSAGES, to help reach new user communities in coarse-grained modeling, materials science, and biomolecular dynamics. These suggested extensions are part of a broader direction focused on improving accessibility for the scientific community overall. This includes creating generally safe default parameters for novice users, maintaining clear and comprehensive documentation, and ensuring compatibility with external software workflows.

(iii) Future applications in molecular engineering. PINES is particularly well-suited for emerging applications in molecular design where symmetry, heterogeneity, and long-range correlations play central roles. One promising direction is the study of DNA hybridization and dehybridization [212]. By assigning individual strands or domains to separate PIV blocks, PINES can isolate the collective variables that govern strand pairing, mispairing, and unbinding, enabling fine-grained monitoring of hybridization dynamics that are crucial for DNA nanotechnology and strand displacement systems.

In the realm of biomineralization, PINES offers a compelling platform for modeling nucleation events around charged moieties in peptides and proteins. Its ability to resolve ion-solvent-protein coordination environments makes it ideal for identifying emergent CVs that describe the onset of mineral aggregation, a challenge that has long eluded standard

descriptors due to the high degree of symmetry and local disorder in the early stages of crystallization [213, 214].

PINES also presents an opportunity to advance understanding of homogeneous nucleation, where no obvious structural template exists but spontaneous symmetry-breaking leads to emergent order [215, 216, 217]. By scanning large spatial volumes with permutation-invariant filters, PINES could help identify low-dimensional representations of pre-nucleation clusters and early-stage crystal formation.

Finally, PINES may offer enhanced representations in protein folding by incorporating solvent directly into the CVs, enabling the tracking of water-mediated transitions and stabilizing interactions. While many folding studies use internal coordinates alone, the symmetry-adapted framework provided by PINES offers a route to uncover solvent-coupled folding pathways with greater fidelity to the true free energy landscape [218, 219, 220].

Going forward, the most impactful applications of PINES may lie at the intersection of emergent phenomena and molecular design for systems in which many-body interactions, solvent mediation, or spatial heterogeneity defy simple coordinate-based descriptors. For instance, PINES could be applied to the design of strand-displacement logic in synthetic DNA circuits [221], the elucidation of early-stage calcium phosphate cluster formation during bone growth [222, 223], the mapping of heterogeneous nucleation events in lipid bilayers and surfactant systems [224], or the enhanced sampling of folding and unfolding pathways in intrinsically disordered proteins [225]. By continuing to develop the infrastructure and documentation needed for broad accessibility, and by focusing on application areas where permutation symmetry poses a fundamental challenge to existing methods, PINES can evolve from a powerful method into a widely adopted tool for addressing long-standing problems in physical chemistry and molecular engineering.

6.2.2 *Cultivating the Forest*

ALPineFOREst was built with extensibility at its core, and its modular design leaves ample room for future growth. Looking ahead, we can imagine its development proceeding along three complementary paths: infrastructure improvements, incremental passive extensions, and more ambitious active expansions that introduce fundamentally new capabilities.

On the infrastructure front, one of the most impactful opportunities lies in deeper integration with distributed computing environments. In particular, tighter coupling with Globus [226], a platform for secure and high-throughput data transfer, would enable native support for scalable, automated active learning pipelines. A standalone Globus framework already exists within the group and has proven effective at coordinating simulation, featurization, and modeling across nodes. Incorporating this functionality directly into ALPineFOREst via Globus-aware job orchestration, stage-aware data triggers, or batch-level dispatching could transform it into a turnkey solution for high-throughput virtual screening campaigns across heterogeneous computing systems.

By contrast, passive development encompasses a set of lower-effort but still valuable enhancements that can be added organically as new needs arise. These typically take the form of modular alternatives to existing components: a new file loader for an additional simulation format, a different acquisition function for candidate selection, or a drop-in model built on tools from GPyTorch [227], BoTorch [228], or scikit-learn [229]. Because these additions reuse well-defined interfaces and extend existing functionality, they require relatively little overhead to implement. Importantly, the growth of this layer is self-regulating: modules are only created when there is a practical need, and they persist only if they prove useful across multiple applications. In this way, the codebase evolves under a kind of natural selection, prioritizing generality and relevance without accumulating unused or overly specialized components.

Finally, there are directions for active development that would involve deeper changes

to ALPineFOREst’s architecture and capabilities. These efforts aim to open up entirely new functionality for the library and, therefore, new options for active learning-based scientific research. One particularly promising direction is the introduction of multi-objective and multi-fidelity surrogate models [230]. Such models would enable hybrid campaigns that combine low- and high-fidelity simulations or blend computational and experimental measurements within a unified acquisition loop [231].

More broadly, ALPineFOREst could be extended to support experimental active learning just as naturally as computational active learning. In fact, because Globus software is increasingly deployed alongside scientific instruments for data management and synchronization, the infrastructure already exists to automate experimental workflows as easily as simulation ones. By leveraging these same automation hooks, ALPineFOREst could be adapted to guide real-world experimental sampling by triggering experiments, ingesting results, and updating models in a closed-loop fashion. This not only doubles the potential user base of the software, but also lowers the barrier to hybrid campaigns that draw on the strengths of both *in silico* and *in vitro* discovery.

Another major axis of active development involves leveraging large, publicly available molecular databases to improve the efficiency and generalizability of early-stage models. By implementing functionality to pretrain surrogate models on broad datasets, such as QM9 [232], MoleculeNet [233], or curated property-specific libraries, and then fine-tune those models within a Bayesian optimization loop, ALPineFOREst could dramatically reduce the number of expensive simulations required to initiate productive exploration. This would not only accelerate campaign startup but also make it easier to bootstrap active learning workflows for complex systems where simulation or experimental costs are prohibitive. Adding this capability would align ALPineFOREst with broader trends in machine learning toward transfer learning and foundation model architectures.

Looking forward, ALPineFOREst is well-positioned to support active learning campaigns

across a broad range of molecular design spaces beyond DNA-functionalized particles. One especially rich and well-studied target is π -conjugated oligopeptides, whose tunable photoelectronic properties and sequence diversity offer an ideal testbed for structure–property exploration [214, 59, 39]. Our lab’s prior success with active learning in this space [231] affirms its amenability to this approach. Other compelling targets include ligand–receptor binding landscapes, where predictive surrogate models could prioritize compounds for simulation or assay with minimal labeling [234]; metal–organic frameworks (MOFs), whose structural modularity and large chemical design space present a natural match for model-guided discovery [235, 236]; and coarse-grained polymer systems, where ALPineFOREst could assist in optimizing block sequences, functionalization patterns, or cross-linking strategies for targeted physical behaviors such as elasticity or self-assembly.

Beyond these materials and small-molecule systems, ALPineFOREst could also be adapted to support biologically focused discovery campaigns. In antibody design, active learning can guide sequence modifications to improve antigen binding affinity [237, 238], thermal stability, or developability, particularly when fine-tuning pretrained models on new target systems. Likewise, vaccine and adjuvant formulation presents an ideal case for Bayesian optimization over expensive, multi-dimensional biological endpoints, such as cytokine expression or immune activation levels [239]. Finally, in protein engineering, ALPineFOREst could be used to iteratively explore mutation landscapes to enhance enzymatic function [240], solubility, or expression yields. These biological domains often rely on costly wet-lab experiments, making them strong candidates for model-guided closed-loop design, particularly when integrated with automation platforms like Globus.

REFERENCES

- [1] Debdas Dhabal, Zhitong Jiang, Akash Pallath, and Amish J Patel. Characterizing the interplay between polymer solvation and conformation. *J. Phys. Chem. B*, 125(20): 5434–5442, 2021.
- [2] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: Visual molecular dynamics. *J. Mol. Graphics*, 14(1):33–38, 1996.
- [3] Phillip L Geissler, Christoph Dellago, and David Chandler. Kinetic pathways of ion pair dissociation in water. *J. Phys. Chem. B*, 103(18):3706–3710, 1999.
- [4] Andrew J Ballard and Christoph Dellago. Toward the mechanism of ionic dissociation in water. *J. Phys. Chem. B*, 116(45):13490–13497, 2012.
- [5] G. L. Eakins, J. K. Gallaher, R. A. Keyzers, A. Falber, J. E. A. Webb, A. Laos, Y. Tidhar, H. Weissman, B. Rybtchinski, P. Thordarson, and J. M. Hodgkiss. Thermodynamic factors impacting the peptide-driven self-assembly of perylene diimide nanofibers. *J. Phys. Chem. B*, 118:8642, 2014.
- [6] G. L. Eakins, R. Pandey, J. P. Wojciechowski, H. Y. Zheng, J. E. A. Webb, C. Valery, P. Thordarson, N. O. V. Plank, J. A. Gerrard, and J. M. Hodgkiss. Functional organic semiconductors assembled via natural aggregating peptides. *Adv. Funct. Mater.*, 25: 5640, 2015.
- [7] 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.
- [8] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2):604–613, 2014.
- [9] Wei Chen, Aik Rui Tan, and Andrew L Ferguson. Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *J. Chem. Phys.*, 149(7):072312, 2018.
- [10] 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.
- [11] 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.

- [12] Zineb Belkacemi, Paraskevi Gkeka, Tony Lelièvre, and Gabriel Stoltz. Chasing Collective Variables Using Autoencoders and Biased Trajectories. *Journal of Chemical Theory and Computation*, 18(1):59–78, January 2022. ISSN 1549-9618. doi:[10.1021/acs.jctc.1c00415](https://doi.org/10.1021/acs.jctc.1c00415). URL <https://doi.org/10.1021/acs.jctc.1c00415>. Publisher: American Chemical Society.
- [13] C. A. Mirkin. The polyvalent gold nanoparticle conjugate—materials synthesis, bio-diagnostics, and intracellular gene regulation. *MRS Bulletin*, 35(7):532–539, 2010. doi:[10.1557/mrs2010.602](https://doi.org/10.1557/mrs2010.602). URL <https://doi.org/10.1557/mrs2010.602>.
- [14] Nicolò Alvisi and Renko de Vries. Biomedical applications of solid-binding peptides and proteins. *Materials Today Bio*, 19:100580, April 2023. ISSN 2590-0064. doi:[10.1016/j.mtbio.2023.100580](https://doi.org/10.1016/j.mtbio.2023.100580). URL <https://www.sciencedirect.com/science/article/pii/S2590006423000406>.
- [15] L. Lin, Y. Liu, L. Tang, and J. Li. Electrochemical dna sensor by the assembly of graphene and dna-conjugated gold nanoparticles with silver enhancement strategy. *Analyst*, 136(22):4732–4737, 2011. doi:[10.1039/C1AN15610A](https://doi.org/10.1039/C1AN15610A). URL <https://doi.org/10.1039/C1AN15610A>.
- [16] R. Mao, E. Pretti, and J. Mittal. Temperature-controlled reconfigurable nanoparticle binary superlattices. *ACS Nano*, 15(5):8466–8473, 2021. doi:[10.1021/acs.nano.0c10874](https://doi.org/10.1021/acs.nano.0c10874).
- [17] M. Song, Y. Ding, H. Zerbe, M. A. Snyder, and J. Mittal. Binary superlattice design by controlling dna-mediated interactions. *Langmuir*, 34(3):991–998, 2018. doi:[10.1021/acs.langmuir.7b02835](https://doi.org/10.1021/acs.langmuir.7b02835). URL <https://doi.org/10.1021/acs.langmuir.7b02835>.
- [18] Runfang Mao, Brian Minevich, Daniel McKeen, Qizhan Chen, Fang Lu, Oleg Gang, and Jeetain Mittal. Regulating phase behavior of nanoparticle assemblies through engineering of DNA-mediated isotropic interactions. *Proceedings of the National Academy of Sciences*, 120(52):e2302037120, December 2023. doi:[10.1073/pnas.2302037120](https://doi.org/10.1073/pnas.2302037120). URL <https://www.pnas.org/doi/10.1073/pnas.2302037120>. Publisher: Proceedings of the National Academy of Sciences.
- [19] Runfang Mao and Jeetain Mittal. Self-Assembly of DNA-Functionalized Nanoparticles Guided by Binding Kinetics. *The Journal of Physical Chemistry B*, 124(51):11593–11599, December 2020. ISSN 1520-6106. doi:[10.1021/acs.jpcc.0c08225](https://doi.org/10.1021/acs.jpcc.0c08225). URL <https://doi.org/10.1021/acs.jpcc.0c08225>. Publisher: American Chemical Society.
- [20] F. J. M. Hoebe, P. Jonkheijm, E. W. Meijer, and A. P. H. J. Schenning. About supramolecular assemblies of -conjugated systems. *Chem. Rev.*, 105:1491, 2005.
- [21] M. Hasegawa and M. Iyoda. Conducting supramolecular nanofibers and nanorods. *Chem. Soc. Rev.*, 39:2420, 2010.

- [22] T. Aida, E. W. Meijer, and S. I. Stupp. Functional supramolecular polymers. *Science*, 335:813, 2012.
- [23] E. Moulin, J. J. Cid, and N. Giuseppone. Advances in supramolecular electronics - from randomly self-assembled nanostructures to addressable self-organized interconnects. *Adv. Mater.*, 25:477, 2013.
- [24] A. P. H. J. Schenning, P. Jonkheijm, E. Peeters, and E. W. Meijer. Hierarchical order in supramolecular assemblies of hydrogen-bonded oligo(p-phenylene vinylene)s. *J. Am. Chem. Soc.*, 123:409, 2001.
- [25] S. S. Panda, H. E. Katz, and J. D. Tovar. Solid-state electrical applications of protein and peptide based nanomaterials. *Chem. Soc. Rev.*, 47:3640, 2018.
- [26] R. Marty, R. Nigon, D. Leite, and H. Frauenrath. Two-fold odd-even effect in self-assembled nanowires from oligopeptide-polymer-substituted perylene bisimides. *J. Am. Chem. Soc.*, 136:3919, 2014.
- [27] A. Chworos, I. Severcan, A. Y. Koyfman, P. Weinkam, E. Oroudjev, H. G. Hansma, and L. Jaeger. Building programmable jigsaw puzzles with rna. *Science*, 306:2068, 2004.
- [28] D. R. Lovley. Electrically conductive pili: Biological function and potential applications in electronics. *Curr. Opin. Electrochem.*, 4:190, 2017.
- [29] M. Amit, S. Yuran, E. Gazit, M. Reches, and N. Ashkenasy. Tailor-made functional peptide self-assembling nanostructures. *Adv. Mater.*, 30:1707083, 2018.
- [30] J. D. Tovar. Supramolecular construction of optoelectronic biomaterials. *Acc. Chem. Res.*, 46:1527, 2013.
- [31] M. Kumar, N. L. Ing, V. Narang, N. Wijerathne, A. I. Hochbaum, and R. V. Ulijn. Amino acid-encoded biocatalytic self-assembly enables the formation of transient conducting nanostructures. *Nat. Chem.*, 10:696, 2018.
- [32] H. Shao, J. Seifert, N. C. Romano, M. Gao, J. J. Helmus, C. P. Jaroniec, D. A. Modarelli, and J. R. Parquette. Amphiphilic self-assembly of an n-type nanotube. *Angew. Chem., Int. Ed.*, 49:7688, 2010.
- [33] W.-W. Tsai, I. D. Tevis, A. S. Tayi, H. Cui, and S. I. Stupp. Semiconducting nanowires from hairpin-shaped self-assembling sexithiophenes. *J. Phys. Chem. B*, 114:14778, 2010.
- [34] M. A. Khalily, G. Bakan, B. Kucukoz, A. E. Topal, A. Karatay, H. G. Yaglioglu, A. Dana, and M. O. Guler. Fabrication of supramolecular n/p-nanowires via coassembly of oppositely charged peptide-chromophore systems in aqueous media. *ACS Nano*, 11:6881, 2017.

- [35] R. J. Kumar, J. M. MacDonald, T. B. Singh, L. J. Waddington, and A. B. Holmes. Hierarchical self-assembly of semiconductor functionalized peptide alpha-helices and optoelectronic properties. *J. Am. Chem. Soc.*, 133:8564, 2011.
- [36] T. Lee, S. S. Panda, J. D. Tovar, and H. E. Katz. Unusually conductive organic-inorganic hybrid nanostructures derived from bio-inspired mineralization of peptide/pi-electron assemblies. *ACS Nano*, 14:1846, 2020.
- [37] J. Cornil, D. A. dos Santos, X. Crispin, R. Silbey, and J. L. Brédas. Influence of interchain interactions on the absorption and luminescence of conjugated oligomers and polymers: a quantum chemical characterization. *J. Am. Chem. Soc.*, 120:1289, 1998.
- [38] S. Siddiqui and F. C. Spano. H- and j-aggregates of conjugated polymers and oligomers: a theoretical investigation. *Chem. Phys. Lett.*, 308:99, 1999.
- [39] B. D. Wall, A. E. Zacca, A. M. Sanders, W. L. Wilson, A. L. Ferguson, and J. D. Tovar. Supramolecular polymorphism: Tunable electronic interactions within -conjugated peptide nanostructures dictated by primary amino acid sequence. *Langmuir*, 30:5946, 2014.
- [40] S. S. Panda, K. Shmilovich, A. L. Ferguson, and J. D. Tovar. Computationally guided tuning of amino acid configuration influences the chiroptical properties of supramolecular peptide-peptide nanostructures. *Langmuir*, 36:6782, 2020.
- [41] M. Wang, P. Zhou, J. Wang, Y. Zhao, H. Ma, J. R. Lu, and H. Xu. Left or right: How does amino acid chirality affect the handedness of nanostructures self-assembled from short amphiphilic peptides? *J. Am. Chem. Soc.*, 139:4185, 2017.
- [42] S. Lin, Y. Li, B. Li, and Y. Yang. Control of the handedness of self-assemblies of dipeptides by the chirality of phenylalanine and steric hindrance of phenylglycine. *Langmuir*, 32:7420, 2016.
- [43] E. T. Pashuck, H. Cui, and S. I. Stupp. Stupp, s. i. tuning supramolecular rigidity of peptide fibers through molecular structure. *J. Am. Chem. Soc.*, 132:6041, 2010.
- [44] F. Würthner and M. Stolte. Naphthalene and perylene diimides for organic transistors. *Chem. Commun.*, 47:5109, 2011.
- [45] S. R. Greenfield, W. A. Svec, D. Gosztola, and M. R. Wasielewski. Multistep photochemical charge separation in rod-like molecules based on aromatic imides and diimides. *J. Am. Chem. Soc.*, 118:6767, 1996.
- [46] E. R. Draper, B. J. Greeves, M. Barrow, R. Schweins, M. A. Zwijnenburg, and D. J. Adams. ph-directed aggregation to control photoconductivity in self-assembled perylene bisimide films. *Chem.*, 2:716, 2017.

- [47] C. Huang, S. Barlow, and S. R. Marder. Perylene-3,4,9,10-tetracarboxylic acid diimides: Synthesis, physical properties, and use in organic electronics. *J. Org. Chem.*, 76:2386, 2011.
- [48] G. Horowitz, F. Kouki, P. Spearman, D. Fichou, C. Nogues, X. Pan, and F. Garnier. Evidence for n-type conduction in a perylene tetracarboxylic diimide derivative. *Adv. Mater.*, 8:242, 1996.
- [49] S. Rajaram, P. B. Armstrong, B. J. Kim, and J. M. J. Frechet. Effect of addition of a diblock copolymer on blend morphology and performance of poly(3-hexylthiophene):perylene diimide solar cells. *Chem. Mater.*, 21:1775, 2009.
- [50] S. V. Dayneko, M. Rahmati, M. Pahlevani, and G. C. Welch. Solution processed red organic light-emitting-diodes using an n-annulated perylene diimide fluorophore. *J. Mater. Chem. C*, 8:2314, 2020.
- [51] G. Klebe, F. Graser, E. Hadicke, and J. Berndt. Crystallochromy as a solid-state effect: correlation of molecular conformation, crystal packing and colour in perylene-3,4:9,10-bis(dicarboximide) pigments. *Acta Crystallogr., Sect. B: Struct. Sci.*, B45:69, 1989.
- [52] S. Roy, D. Kumar Maiti, S. Panigrahi, D. Basak, and A. Banerjee. A new hydrogel from an amino acid-based perylene bisimide and its semiconducting, photo-switching behaviour. *RSC Adv.*, 2:11053, 2012.
- [53] E. R. Draper, J. J. Walsh, T. O. McDonald, M. A. Zwiijnenburg, P. J. Cameron, A. J. Cowan, and D. J. Adams. Air-stable photoconductive films formed from perylene bisimide gelators. *J. Mater. Chem. C*, 2:5570, 2014.
- [54] E. R. Draper, L. J. Archibald, M. C. Nolan, R. Schweins, M. A. Zwiijnenburg, S. Sproules, and D. J. Adams. Controlling photoconductivity in pbi films by supramolecular assembly. *Chem. - Eur. J.*, 24:4006, 2018.
- [55] A. Datar, K. Balakrishnan, and L. Zang. One-dimensional self-assembly of a water soluble perylene diimide molecule by ph triggered hydrogelation. *Chem. Commun.*, 49: 6894, 2013.
- [56] S. Bai, S. Debnath, N. Javid, P. W. J. M. Frederix, S. Fleming, C. Pappas, and R. V. Ulijn. Differential self-assembly and tunable emission of aromatic peptide bola-amphiphiles containing perylene bisimide in polar solvents including water. *Langmuir*, 30:7576, 2014.
- [57] R. Marty, R. Szilluweit, A. Sánchez-Ferrer, S. Bolisetty, J. Adamcik, R. Mezzenga, E.-C. Spitzner, M. Feifer, S. N. Steinmann, C. Corminboeuf, and H. Frauenrath. Hierarchically structured microfibers of “single stack” perylene bisimide and quaterthiophene nanowires. *ACS Nano*, 7:8498, 2013.

- [58] J. K. Gallaher, E. J. Aitken, R. A. Keyzers, and J. M. Hodgkiss. Controlled aggregation of peptide-substituted perylene-bisimides. *Chem. Commun.*, 48:7961, 2012.
- [59] B. A. Thurston and A. L. Ferguson. Machine learning and molecular design of self-assembling -conjugated oligopeptides. *Mol. Simul.*, 44:930, 2018.
- [60] G. S. Vadehra, B. D. Wall, S. D. Diegelmann, and J. D. Tovar. On-resin dimerization incorporates a diverse array of -conjugated functionality within aqueous self-assembling peptide backbones. *Chem. Commun.*, 46:3947, 2010.
- [61] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi. Plumed 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185:604, 2014.
- [62] 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: Struct., Funct., Bioinf.*, 78(8):1950–1958, 2010.
- [63] A. W. Sousa da Silva and W. F. Vranken. Acpype - antechamber python parser interface. *BMC Res. Notes*, 5:367, 2012.
- [64] C. A. Reynolds, J. W. Essex, and W. G. Richards. Atomic charges for variable molecular conformations. *J. Am. Chem. Soc.*, 114:9075, 1992.
- [65] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *J. Phys. Chem.*, 97:10269, 1993.
- [66] 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, 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*. Gaussian, Inc., Wallingford, CT, 2016.
- [67] P. Mark and L. Nilsson. Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k. *J. Phys. Chem. A*, 105:9954, 2001.
- [68] R. W. Hockney. *Computer Simulation Using Particles*. CRC Press, Boca Raton, FL, 1988.

- [69] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18:1463, 1997.
- [70] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103(19):8577–8593, 1995.
- [71] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182, 1981.
- [72] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126:014101, 2007.
- [73] A. Barducci, G. Bussi, and M. Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100:020603, 2008.
- [74] S. S. Panda, K. Shmilovich, A. L. Ferguson, and J. D. Tovar. Controlling supramolecular chirality in peptide-peptide networks by variation of the alkyl spacer length. *Langmuir*, 35:14060, 2019.
- [75] J. S. Hub, B. L. de Groot, and D. van der Spoel. g_wham: free weighted histogram analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.*, 6:3713, 2010.
- [76] S. Pall, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl. Tackling exascale software challenges in molecular dynamics simulations with gromacs. In S. Markidis and E. Laure, editors, *Solving Software Challenges for Exascale. EASC 2014*, volume 8759 of *Lecture Notes in Computer Science*, pages 3–27. Springer, Cham, 2015.
- [77] L. R. Valverde, B. A. Thurston, A. L. Ferguson, and W. L. Wilson. Evidence for prenucleated fibrillogenesis of acid-mediated self-assembling oligopeptides via molecular simulation and fluorescence correlation spectroscopy. *Langmuir*, 34:7346, 2018.
- [78] A. K. Mishra, H. Weissman, E. Krieg, K. A. Votaw, M. McCullagh, B. Rybtchinski, and F. D. Lewis. Self-assembly of perylenediimide-single-strand-dna conjugates: Employing hydrophobic interactions and dna base-pairing to create a diverse structural space. *Chem. - Eur. J.*, 23:10328, 2017.
- [79] A. Ashcraft, K. Liu, A. Mukhopadhyay, V. Paulino, C. Liu, B. Bernard, D. Husainy, T. Phan, and J.-H. Olivier. A molecular strategy to lock-in the conformation of a perylene bisimide-derived supramolecular polymer. *Angew. Chem., Int. Ed.*, 59:7487, 2020.
- [80] D. Gorl, X. Zhang, and F. Wurthner. Molecular assemblies of perylene bisimide dyes in water. *Angew. Chem., Int. Ed.*, 51:6328, 2012.
- [81] R. A. Cormier and B. A. Gregg. Synthesis and characterization of liquid crystalline perylene diimides. *Chem. Mater.*, 10:1309, 1998.

- [82] Y. Xu, S. Leng, C. Xue, R. Sun, J. Pan, J. Ford, and S. Jin. A room-temperature liquid-crystalline phase with crystalline stacks. *Angew. Chem., Int. Ed.*, 46:3896, 2007.
- [83] H. Wu, L. Xue, Y. Shi, Y. Chen, and X. Li. Organogels based on j- and h-type aggregates of amphiphilic perylenetetracarboxylic diimides. *Langmuir*, 27:3074, 2011.
- [84] W. Wagner, M. Wehner, V. Stepanenko, S. Ogi, and F. Würthner. Living supramolecular polymerization of a perylene bisimide dye into fluorescent j-aggregates. *Angew. Chem.*, 129:16224, 2017.
- [85] N. J. Hestand and F. C. Spano. Expanded theory of h- and j-molecular aggregates: The effects of vibronic coupling and intermolecular charge transfer. *Chem. Rev.*, 118:7069, 2018.
- [86] N. J. Hestand, R. V. Kazantsev, A. S. Weingarten, L. C. Palmer, S. I. Stupp, and F. C. Spano. Extended-charge-transfer excitons in crystalline supramolecular photocatalytic scaffolds. *J. Am. Chem. Soc.*, 138:11762, 2016.
- [87] A. Oleson, T. Zhu, I. S. Dunn, D. Bialas, Y. Bai, W. Zhang, M. Dai, D. R. Reichman, R. Tempelaar, L. Huang, and F. C. Spano. Perylene diimide-based hj- and hj-aggregates: The prospect of exciton band shape engineering in organic materials. *J. Phys. Chem. C*, 123:20567, 2019.
- [88] Michael P Allen and Dominic J Tildesley. *Computer Simulation of Liquids*. Oxford University Press, New York, NY, 2017.
- [89] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Elsevier, Amsterdam, Netherlands, 2001.
- [90] Ken A Dill and Hue Sun Chan. From levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4(1):10–19, 1997.
- [91] Jérôme Hénin, Tony Lelièvre, Michael Shirts, Omar Valsson, and Lucie Delemotte. Enhanced sampling methods for molecular dynamics simulations [article v1. 0]. *Living J. Comput. Mol. Sci.*, 4(1):1583–1583, 2022.
- [92] Yi Isaac Yang, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Qin Gao. Enhanced sampling in molecular dynamics. *J. Chem. Phys.*, 151(7):070902, 2019.
- [93] Cameron Abrams and Giovanni Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2013.
- [94] 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.

- [95] 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.
- [96] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *J. Chem. Phys.*, 115(20):9169–9183, 2001.
- [97] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.*, 99(20):12562–12566, 2002.
- [98] Robert Zwanzig. *Nonequilibrium Statistical Mechanics*. Oxford University Press, New York, NY, 2001.
- [99] J Wang and AL Ferguson. Nonlinear machine learning in simulations of soft and biological materials. *Mol. Simul.*, 44(13-14):1090–1107, 2018.
- [100] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019.
- [101] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2022.
- [102] Zineb Belkacemi, Paraskevi Gkeka, Tony Lelièvre, and Gabriel Stoltz. Chasing collective variables using autoencoders and biased trajectories. *J. Chem. Theory Comput.*, 18(1):59–78, 2021.
- [103] Ferry Hooft, Alberto Pérez de Alba Ortíz, and Bernd Ensing. Discovering collective variables of molecular transitions via genetic algorithms and neural networks. *J. Chem. Theory Comput.*, 17(4):2294–2306, 2021.
- [104] Ao Ma and Aaron R Dinner. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B*, 109(14):6769–6779, 2005.
- [105] Rangsiman Ketkaew, Fabrizio Creazzo, and Sandra Luber. Machine learning-assisted discovery of hidden states in expanded free energy space. *J. Phys. Chem. Lett.*, 13(7):1797–1805, 2022.
- [106] Rangsiman Ketkaew and Sandra Luber. Deepcv: A deep learning framework for blind search of collective variables in expanded configurational space. *J. Chem. Inf. Model.*, 62(24):6352–6364, 2022.
- [107] Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.*, 148(24):241703, 2018.
- [108] Carlos X Hernández, Hannah K Wayment-Steele, Mohammad M Sultan, Brooke E Husic, and Vijay S Pande. Variational encoding of complex dynamics. *Phys. Rev. E*, 97(6):062412, 2018.

- [109] Mohammad M Sultan, Hannah K Wayment-Steele, and Vijay S Pande. Transferable neural networks for enhanced sampling of protein dynamics. *J. Chem. Theory Comput.*, 14(4):1887–1894, 2018.
- [110] Hannah K Wayment-Steele and Vijay S Pande. Note: Variational encoding of protein dynamics benefits from maximizing latent autocorrelation. *J. Chem. Phys.*, 149(21):216101, 2018.
- [111] Dedi Wang and Pratyush Tiwary. State predictive information bottleneck. *J. Chem. Phys.*, 154(13):134111, 2021.
- [112] 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.
- [113] Wei Chen, Hythem Sidky, and Andrew L Ferguson. Capabilities and limitations of time-lagged autoencoders for slow mode discovery in dynamical systems. *J. Chem. Phys.*, 151(6):064123, 2019.
- [114] 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. *J. Phys. Chem. A*, 127(15):3497–3517, 2023.
- [115] 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.
- [116] M Frassek, A Arjun, and PG Bolhuis. An extended autoencoder model for reaction coordinate discovery in rare event molecular dynamics datasets. *J. Chem. Phys.*, 155(6):064103, 2021.
- [117] Felix Musil, Andrea Grisafi, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. Physics-inspired structural representations for molecules and materials. *Chem. Rev.*, 121(16):9759–9815, 2021.
- [118] 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. *Comput. Phys. Commun.*, 247:106949, 2020.
- [119] Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. *Adv. Neural Inf. Process. Syst.*, 32:14537–14546, 2019.
- [120] Tess E Smidt. Euclidean symmetry and equivariance in machine learning. *Trends Chem.*, 3(2):82–85, 2021.

- [121] 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):2453, 2022.
- [122] Florian Sittel, Abhinav Jain, and Gerhard Stock. Principal component analysis of molecular dynamics: On the use of cartesian vs. internal coordinates. *J. Chem. Phys.*, 141(1):07B605_1, 2014.
- [123] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.*, 120(14):143001, 2018.
- [124] Connor Shorten and Taghi M Khoshgohar. A survey on image data augmentation for deep learning. *J. Big Data*, 6(1):1–48, 2019.
- [125] Robin Winter, Frank Noé, and Djork-Arné Clevert. Permutation-invariant variational autoencoder for graph-level representation learning. *Adv. Neural. Inf. Process. Syst.*, 34:9559–9573, 2021.
- [126] Bojun Liu, Mingyi Xue, Yunrui Qiu, Kirill Konovalov, Michael O’Connor, and Xuhui Huang. Graphvampnets for uncovering slow collective variables of self-assembly dynamics. *ChemRxiv preprint 10.26434/chemrxiv-2023-v0dww*, 2023.
- [127] Frederico V Prudente, Paulo H Acioli, and JJ Soares Neto. The fitting of potential energy surfaces using neural networks: Application to the study of vibrational levels of h 3+. *J. Chem. Phys.*, 109(20):8801–8808, 1998.
- [128] Hieu TT Nguyen and Hung M Le. Modified feed-forward neural network structures and combined-function-derivative approximations incorporating exchange symmetry for potential energy surface fitting. *J. Phys. Chem. A*, 116(18):4629–4638, 2012.
- [129] Bastiaan J Braams and Joel M Bowman. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.*, 28(4):577–606, 2009.
- [130] Zhen Xie and Joel M Bowman. Permutationally invariant polynomial basis for molecular energy surface fitting via monomial symmetrization. *J. Chem. Theory Comput.*, 6(1):26–34, 2010.
- [131] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98(14):146401, 2007.
- [132] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.*, 134(7):074106, 2011.
- [133] Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning. *Mach. Learn.: Sci. Technol.*, page 045017, 2022.

- [134] Chen Gu, Huang-Wei Chang, Lutz Maibaum, Vijay S Pande, Gunnar E Carlsson, and Leonidas J Guibas. Building markov state models with solvent dynamics. *BMC Bioinf.*, 14:1–9, 2013.
- [135] Matthew P Harrigan, Diwakar Shukla, and Vijay S Pande. Conserve water: A method for the analysis of solvent in molecular dynamics. *J. Chem. Theory Comput.*, 11(3): 1094–1101, 2015.
- [136] Robert Raddi and Vincent Voelz. A markov state model of solvent features reveals water dynamics in protein-peptide binding. *ChemRxiv preprint 10.26434/chemrxiv-2023-r6njk*, 2023.
- [137] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87(18):184115, 2013.
- [138] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O Anatole Von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and non-locality in chemical space. *J. Phys. Chem. Lett.*, 6(12):2326–2331, 2015.
- [139] Fabio Pietrucci and Wanda Andreoni. Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale. *Phys. Rev. Lett.*, 107(8): 085504, 2011.
- [140] Fabio Pietrucci. Novel enhanced sampling strategies for transitions between ordered and disordered structures. In Wanda Andreoni and Sidney Yip, editors, *Handbook of Materials Modeling: Methods: Theory and Modeling*, chapter 28: Novel enhanced sampling strategies for transitions between ordered and disordered structures, pages 597–619. Springer Nature, Switzerland AG, 2020. ISBN 978-3-319-44677-6. doi:10.1007/978-3-319-44677-6_51. URL https://doi.org/10.1007/978-3-319-44677-6_51.
- [141] 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.
- [142] Grégoire A Gallet and Fabio Pietrucci. Structural cluster analysis of chemical reactions in solution. *J. Chem. Phys.*, 139(7):074101, 2013.
- [143] Andrew L. Ferguson, Athanassios Z. Panagiotopoulos, Pablo G. Debenedetti, and Ioannis G. Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.*, 107(31):13597–13602, 2010.
- [144] Amish J Patel, Patrick Varilly, David Chandler, and Shekhar Garde. Quantifying density fluctuations in volumes of all shapes and sizes using indirect umbrella sampling. *J. Stat. Phys.*, 145(2):265–275, 2011.

- [145] Ziyue Zou, Eric R Beyerle, Sun-Ting Tsai, and Pratyush Tiwary. Driving and characterizing nucleation of urea and glycine polymorphs in water. *Proc. Natl. Acad. Sci. U. S. A.*, 120(7):e2216099120, 2023.
- [146] Valerio Rizzi, Luigi Bonati, Narjes Ansari, and Michele Parrinello. The role of water in host-guest interaction. *Nat. Commun.*, 12(1):1–7, 2021.
- [147] 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.
- [148] Bin Jiang and Hua Guo. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. *J. Chem. Phys.*, 139(5):054112, 2013.
- [149] 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.
- [150] Amit Singer. A remark on global positioning from local distances. *Proc. Natl. Acad. Sci. U. S. A.*, 105(28):9507–9511, 2008.
- [151] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Process. Mag.*, 32(6):12–30, 2015.
- [152] PLUMED2. Switching functions. <https://www.plumed.org/doc-v2.9/user-doc/html/switchingfunction.html>. (Accessed October 2023).
- [153] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2018.
- [154] Stan Salvador and Philip Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE International Conference on Tools with Artificial Intelligence*, pages 576–584. IEEE, 2004.
- [155] 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.
- [156] Oliver F Lange and Helmut Grubmüller. Can principal components yield a dimension reduced description of protein dynamics on long time scales? *J. Phys. Chem. B*, 110(45):22842–22852, 2006.
- [157] Angel E. García. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, 68:2696–2699, Apr 1992. doi:10.1103/PhysRevLett.68.2696. URL <https://link.aps.org/doi/10.1103/PhysRevLett.68.2696>.
- [158] Andrea Amadei, Antonius BM Linssen, and Herman JC Berendsen. Essential dynamics of proteins. *Proteins: Struct., Funct., Bioinf.*, 17(4):412–425, 1993.

- [159] 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.
- [160] 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.
- [161] Arushi Prakash, Christopher D Fu, Massimiliano Bonomi, and Jim Pfandtner. Biasing smarter, not harder, by partitioning collective variables into families in parallel bias metadynamics. *J. Chem. Theory Comput.*, 14(10):4985–4990, 2018.
- [162] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, editors, *Adv. Neural Inf. Process. Syst.*, volume 32, chapter PyTorch: An Imperative Style, High-Performance Deep Learning Library. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.
- [163] Falcon, William. Pytorch lightning. <https://github.com/PyTorchLightning/pytorch-lightning>. (Accessed April 2022).
- [164] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.
- [165] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.
- [166] 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.
- [167] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nat. Rev. Phys.*, 2(4):200–212, 2020.
- [168] Massimiliano Bonomi, Giovanni Bussi, Carlo Camilloni, Gareth A. Tribello, Pavel Banáš, Alessandro Barducci, Mattia Bernetti, Peter G. Bolhuis, Sandro Bottaro, Davide Branduardi, Riccardo Capelli, Paolo Carloni, Michele Ceriotti, Andrea Cesari, Haochuan Chen, Wei Chen, Francesco Colizzi, Sandip De, Marco De La Pierre, Davide Donadio, Viktor Drobot, Bernd Ensing, Andrew L. Ferguson, Marta Filizola, James S. Fraser, Haohao Fu, Piero Gasparotto, Francesco Luigi Gervasio, Federico Giberti, Alejandro Gil-Ley, Toni Giorgino, Gabriella T. Heller, Glen M. Hocky, Marcella Iannuzzi, Michele Invernizzi, Kim E. Jelfs, Alexander Jussupow, Evgeny Kirilin,

- Alessandro Laio, Vittorio Limongelli, Kresten Lindorff-Larsen, Thomas Löhr, Fabrizio Marinelli, Layla Martin-Samos, Matteo Masetti, Ralf Meyer, Angelos Michaelides, Carla Molteni, Tetsuya Morishita, Marco Nava, Cristina Paissoni, Elena Papaleo, Michele Parrinello, Jim Pfaendtner, Pablo Piaggi, GiovanniMaria Piccini, Adriana Pietropaolo, Fabio Pietrucci, Silvio Pipolo, Davide Provasi, David Quigley, Paolo Raiteri, Stefano Raniolo, Jakub Rydzewski, Matteo Salvalaglio, Gabriele Cesare Sosso, Vojtěch Spiwok, Jiří Šponer, David W. H. Swenson, Pratyush Tiwary, Omar Valsion, Michele Vendruscolo, Gregory A. Voth, and Andrew White. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods*, 16(8):670–673, Aug 2019. ISSN 1548-7105. doi:[10.1038/s41592-019-0506-8](https://doi.org/10.1038/s41592-019-0506-8). URL <https://doi.org/10.1038/s41592-019-0506-8>.
- [169] Doye, Jonathan. Lennard-jones clusters. <http://doye.chem.ox.ac.uk/research/forecast/LJ.html>. (Accessed September 2022).
- [170] A. Dawid and Z. Gburski. Interaction-induced light scattering in lennard-jones argon clusters: Computer simulations. *Phys. Rev. A*, 56:3294–3296, Oct 1997. doi:[10.1103/PhysRevA.56.3294](https://doi.org/10.1103/PhysRevA.56.3294). URL <https://link.aps.org/doi/10.1103/PhysRevA.56.3294>.
- [171] CL Briant and JJ Burton. Molecular dynamics study of the structure and thermodynamic properties of argon microclusters. *J. Chem. Phys.*, 63(5):2045–2058, 1975.
- [172] R.W Hockney, S.P Goel, and J.W Eastwood. Quiet high-resolution computer models of a plasma. *J. Comput. Phys.*, 14(2):148–158, 1974. ISSN 0021-9991. doi:[https://doi.org/10.1016/0021-9991\(74\)90010-2](https://doi.org/10.1016/0021-9991(74)90010-2). URL <https://www.sciencedirect.com/science/article/pii/0021999174900102>.
- [173] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007. doi:[10.1063/1.2408420](https://doi.org/10.1063/1.2408420). URL <https://doi.org/10.1063/1.2408420>.
- [174] Hendrik Jung, Roberto Covino, and Gerhard Hummer. Artificial intelligence assists discovery of reaction coordinates and mechanisms from molecular dynamics simulations. *arXiv preprint arXiv:1901.04595*, 2019.
- [175] Ryan Gotchy Mullen, Joan-Emma Shea, and Baron Peters. Transmission coefficients, committors, and solvent coordinates in ion-pair dissociation. *J. Chem. Theory Comput.*, 10(2):659–667, 2014.
- [176] Matthias Kellermeier, Paolo Raiteri, John K Berg, Andreas Kempter, Julian D Gale, and Denis Gebauer. Entropy drives calcium carbonate ion association. *ChemPhysChem*, 17(21):3535–3541, 2016.
- [177] Dedi Wang, Renjie Zhao, John D Weeks, and Pratyush Tiwary. Influence of long-range forces on the transition states and dynamics of nacl ion-pair dissociation in water. *J. Phys. Chem. B*, 126(2):545–551, 2022.

- [178] In Suk Joung and Thomas E Cheatham III. Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J. Phys. Chem. B*, 113(40):13279–13290, 2009.
- [179] In Suk Joung and Thomas E Cheatham III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, 112(30):9020–9041, 2008.
- [180] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.
- [181] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An $n \log(n)$ method for ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993. doi:[10.1063/1.464397](https://doi.org/10.1063/1.464397). URL <https://doi.org/10.1063/1.464397>.
- [182] Marcus G Martin and J Ilja Siepmann. Transferable potentials for phase equilibria. 1. united-atom description of n-alkanes. *J. Phys. Chem. B*, 102(14):2569–2577, 1998.
- [183] Becky L Eggimann, Amara J Sunnarborg, Hudson D Stern, Andrew P Bliss, and J Ilja Siepmann. An online parameter and property database for the trappe force field. *Mol. Simul.*, 40(1-3):101–105, 2014.
- [184] HJC Berendsen, JR Grigera, and TP Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91(24):6269–6271, 1987.
- [185] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159:98–103, Jul 1967. doi:[10.1103/PhysRev.159.98](https://link.aps.org/doi/10.1103/PhysRev.159.98). URL <https://link.aps.org/doi/10.1103/PhysRev.159.98>.
- [186] Szilárd Páll and Berk Hess. A flexible algorithm for calculating pair interactions on simd architectures. *Comput. Phys. Commun.*, 184(12):2641–2650, 2013.
- [187] Hythem Sidky, Yamil J. Colón, Julian Helfferich, Benjamin J. Sikora, Cody Bezik, Weiwei Chu, Federico Giberti, Ashley Z. Guo, Xikai Jiang, Joshua Lequieu, Jiyuan Li, Joshua Moller, Michael J. Quevillon, Mohammad Rahimi, Hadi Ramezani-Dakhel, Vikramjit S. Rathee, Daniel R. Reid, Emre Sevgen, Vikram Thapar, Michael A. Webb, Jonathan K. Whitmer, and Juan J. de Pablo. SSAGES: Software Suite for Advanced General Ensemble Simulations. *J. Chem. Phys.*, 148(4):044104, 01 2018. ISSN 0021-9606. doi:[10.1063/1.5008853](https://doi.org/10.1063/1.5008853). URL <https://doi.org/10.1063/1.5008853>.
- [188] Jeff Timko, Denis Bucher, and Serdar Kuyucak. Dissociation of nacl in water from ab initio molecular dynamics simulations. *J. Chem. Phys.*, 132(11):114510, 2010.
- [189] Pieter Rein ten Wolde and David Chandler. Drying-induced hydrophobic polymer collapse. *Proc. Natl. Acad. Sci. U.S.A.*, 99(10):6539–6543, 2002.

- [190] Manoj V Athawale, Gaurav Goel, Tuhin Ghosh, Thomas M Truskett, and Shekhar Garde. Effects of lengthscales and attractions on the collapse of hydrophobic polymers in water. *Proc. Natl. Acad. Sci. U.S.A.*, 104(3):733–738, 2007.
- [191] Luigi Bonati, Valerio Rizzi, and Michele Parrinello. Data-driven collective variables for enhanced sampling. *J. Phys. Chem. Lett.*, 11(8):2998–3004, 2020.
- [192] Bonati, Luigi. mlcvs: Machine learning collective variables. <https://github.com/luigibonati/mlcvs>. (Accessed April 2023).
- [193] Pablo F Zubieta Rico, Ludwig Schneider, Gustavo Perez-Lemus, Riccardo Alessandri, Siva Dasetty, Cintia A Menéndez, Yiheng Wu, Yezhi Jin, Trung Nguyen, John Parker, Andrew L. Ferguson, Jonathan K. Whitmer, and Juan J. de Pablo. Pysages: flexible, advanced sampling methods accelerated with gpus. *arXiv preprint arXiv:2301.04835*, 2023.
- [194] D. Nykypanchuk, M. M. Maye, D. van der Lelie, and O. Gang. Dna-guided crystallization of colloidal nanoparticles. *Nature*, 451(7178):549–552, 2008. doi:[10.1038/nature06560](https://doi.org/10.1038/nature06560).
- [195] S. Y. Park, A. K. R. Lytton-Jean, B. Lee, S. Weigand, G. C. Schatz, and C. A. Mirkin. Dna-programmable nanoparticle crystallization. *Nature*, 451(7178):553–556, 2008. doi:[10.1038/nature06508](https://doi.org/10.1038/nature06508).
- [196] A. P. Alivisatos, K. P. Johnsson, X. Peng, T. E. Wilson, C. J. Loweth, M. P. Bruchez Jr, and P. G. Schultz. Organization of 'nanocrystal molecules' using dna. *Nature*, 382(6592):609–611, 1996. doi:[10.1038/382609a0](https://doi.org/10.1038/382609a0).
- [197] R. J. Macfarlane, B. Lee, H. D. Hill, A. J. Senesi, and C. A. Mirkin. Assembly and organization processes in dna-directed colloidal crystallization. *Science*, 334(6053):204–208, 2011. doi:[10.1126/science.1208794](https://doi.org/10.1126/science.1208794).
- [198] H. Xiong, D. van der Lelie, and O. Gang. Dna linkers for the self-assembly of nanoarchitectures. *Angewandte Chemie International Edition*, 52(24):6580–6584, 2013. doi:[10.1002/anie.201302019](https://doi.org/10.1002/anie.201302019).
- [199] C. A. Mirkin, R. L. Letsinger, R. C. Mucic, and J. J. Storhoff. A dna-based method for rationally assembling nanoparticles into macroscopic materials. *Nature*, 382(6592):607–609, 1996. doi:[10.1038/382607a0](https://doi.org/10.1038/382607a0).
- [200] H. D. Hill and C. A. Mirkin. The bio-barcode assay for the detection of protein and nucleic acid targets using dtt-induced ligand exchange and nanoparticle aggregation. *Nature Materials*, 5(6):412–416, 2006. doi:[10.1038/nmat1641](https://doi.org/10.1038/nmat1641).
- [201] J. S. Lee, P. A. Ulmann, M. S. Han, and C. A. Mirkin. Using dna to link gold nanoparticles, polymers, and small molecules: a route to new materials. *Nature Nanotechnology*, 5(10):673–676, 2010. doi:[10.1038/nnano.2010.177](https://doi.org/10.1038/nnano.2010.177).

- [202] Y. Ding and J. Mittal. Insights into dna-mediated interparticle interactions from coarse-grained molecular dynamics simulations. *The Journal of Chemical Physics*, 141(18):184901, 2014. doi:[10.1063/1.4900881](https://doi.org/10.1063/1.4900881).
- [203] M. M. Maye, M. T. Kumara, D. Nykypanchuk, W. B. Sherman, and O. Gang. Switching binary states of nanoparticle superlattices and dimer clusters by dna strands. *Nature Nanotechnology*, 5(2):116–120, 2010. doi:[10.1038/nnano.2009.378](https://doi.org/10.1038/nnano.2009.378). URL <https://doi.org/10.1038/nnano.2009.378>.
- [204] E. Pretti, R. Fernandez, R. Mao, and J. Mittal. Assembly of three-dimensional binary superlattices using dna-coated nanoparticles. *The Journal of Physical Chemistry Letters*, 9(11):3133–3138, 2018. doi:[10.1021/acs.jpclett.8b00977](https://doi.org/10.1021/acs.jpclett.8b00977).
- [205] Michele Invernizzi. OPES: On-the-fly Probability Enhanced Sampling Method. *Il Nuovo Cimento C*, 44(405):1–4, September 2021. ISSN 03905551, 03905551. doi:[10.1393/ncc/i2021-21112-8](https://doi.org/10.1393/ncc/i2021-21112-8). URL <http://arxiv.org/abs/2101.06991>. arXiv:2101.06991 [physics].
- [206] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.*, 271:108171, 2022. doi:[10.1016/j.cpc.2021.108171](https://doi.org/10.1016/j.cpc.2021.108171).
- [207] Yajun Ding and Jeetain Mittal. Insights into DNA-mediated interparticle interactions from a coarse-grained model. *The Journal of Chemical Physics*, 141(18):184901, November 2014. ISSN 0021-9606, 1089-7690. doi:[10.1063/1.4900891](https://doi.org/10.1063/1.4900891). URL <https://pubs.aip.org/aip/jcp/article/675568>.
- [208] IV Knotts, Thomas A., Nitin Rathore, David C. Schwartz, and Juan J. de Pablo. A coarse grain model for dna. *The Journal of Chemical Physics*, 126(8):084901, 2007. doi:[10.1063/1.2431804](https://doi.org/10.1063/1.2431804).
- [209] Albert I. Ugochukwu and Peter W. B. Phillips. Open data ownership and sharing: Challenges and opportunities for application of FAIR principles and a checklist for data managers. *Journal of Agriculture and Food Research*, 16:101157, June 2024. ISSN 2666-1543. doi:[10.1016/j.jafr.2024.101157](https://doi.org/10.1016/j.jafr.2024.101157). URL <https://www.sciencedirect.com/science/article/pii/S2666154324001947>.
- [210] R. D. Kush, D. Warzel, M. A. Kush, A. Sherman, E. A. Navarro, R. Fitzmartin, F. Pétavy, J. Galvez, L. B. Becnel, F. L. Zhou, N. Harmon, B. Jauregui, T. Jackson, and L. Hudson. FAIR data sharing: The roles of common data elements and harmonization. *Journal of Biomedical Informatics*, 107:103421, July 2020. ISSN 1532-0464. doi:[10.1016/j.jbi.2020.103421](https://doi.org/10.1016/j.jbi.2020.103421). URL <https://www.sciencedirect.com/science/article/pii/S1532046420300496>.

- [211] Barend Mons, Erik Schultes, Fenghong Liu, and Annika Jacobsen. The FAIR Principles: First Generation Implementation Choices and Challenges. *Data Intelligence*, 2(1-2):1–9, January 2020. ISSN 2641-435X. doi:[10.1162/dint_e_00023](https://doi.org/10.1162/dint_e_00023). URL https://doi.org/10.1162/dint_e_00023.
- [212] 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. *Journal of the American Chemical Society*, 143(42):17395–17408, 2021. doi:[10.1021/jacs.1c05219](https://doi.org/10.1021/jacs.1c05219).
- [213] Sophia Iafolla, Maung Maung, Stefano Malissa, and Fabio Nudelman. The biomineralization proteome: Protein complexity for a universal skeleton. *Proteomics*, 19(24), 2019. doi:[10.1002/pmic.201900036](https://doi.org/10.1002/pmic.201900036).
- [214] Taein Lee, Sayak S. Panda, John D. Tovar, and Howard E. Katz. Unusually conductive organic–inorganic hybrid nanostructures derived from bio-inspired mineralization of peptide/-electron assemblies. *ACS Nano*, 14(2):1846–1855, 2020. doi:[10.1021/acsnano.9b07911](https://doi.org/10.1021/acsnano.9b07911).
- [215] Tianshu Li, Davide Donadio, Giovanna Russo, and Giulia Galli. Homogeneous ice nucleation from supercooled water. *Phys. Chem. Chem. Phys.*, 13(44):19807–19813, 2011.
- [216] Pablo M. Piaggi et al. Homogeneous ice nucleation in an ab initio machine-learning model. *Proc. Natl. Acad. Sci. U.S.A.*, 2022.
- [217] Qiang Sun, Shuai Cui, and Meixi Zhang. Homogeneous nucleation mechanism of nacl in aqueous solutions. *Crystals*, 10(2):107, 2020.
- [218] Ruhong Zhou and B. J. Berne. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins*, 53:148–161, 2003.
- [219] Margaret S. Cheung, Ángel E. García, and Jose N. Onuchic. Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. U.S.A.*, 99(2):685–690, 2002.
- [220] P. L. San Biagio, D. Bulone, V. Martorana, and M. B. Palma-Vittorelli. Solvent-induced free energy landscape and solute–solvent dynamic coupling in a multielement solute. *Biophys. J.*, 77:2477–2490, 1999.
- [221] Lulu Qian, Erik Winfree, and Jehoshua Bruck. Neural network computation with dna strand displacement cascades. *Nature*, 475:368–372, 2011. doi:[10.1038/nature10262](https://doi.org/10.1038/nature10262).
- [222] Sai Xue et al. Formation and evolution of nanoscale calcium phosphate clusters in simulated body fluids. *Anal. Chem.*, 93:..., 2021. doi:[10.1021/acs.analchem.1c01561](https://doi.org/10.1021/acs.analchem.1c01561).

- [223] et al. Zhang. Redefined ion association thermodynamics and cahpo4 pre-nucleation cluster formation. *Nat Commun*, 15:... , 2024. doi:[10.1038/s41467-024-47721-7](https://doi.org/10.1038/s41467-024-47721-7).
- [224] John M. Johnston et al. Physical mechanisms of amyloid nucleation on fluid membranes. *Proc. Natl. Acad. Sci. U.S.A.*, 117:... , 2020. doi:[10.1073/pnas.2007694117](https://doi.org/10.1073/pnas.2007694117).
- [225] Meng Han, Jun Xu, Yu Ren, et al. Sampling conformational space of intrinsically disordered proteins in explicit solvent: Comparison between well-tempered ensemble approach and solute tempering method. *J. Mol. Graph. Model.*, 72:136–147, 2017. doi:[10.1016/j.jmgm.2016.12.014](https://doi.org/10.1016/j.jmgm.2016.12.014).
- [226] Kyle Chard, Jim Pruyne, Kurt McKee, Josh Bryan, Brigitte Raumann, Rachana Ananthakrishnan, and Ian Foster. Globus automation services: Research process automation across the space–time continuum. *arXiv preprint arXiv:2208.09513*, 2022. Demonstrates event-driven automated workflows spanning distributed data, computation, and storage systems.
- [227] Jacob R. Gardner, Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7576–7586, 2018.
- [228] Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21524–21538, 2020.
- [229] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [230] 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*, 13:25–44, 2022. doi:[10.1146/annurev-chembioeng-092120-020803](https://doi.org/10.1146/annurev-chembioeng-092120-020803).
- [231] Kirill Shmilovich, Sayak Subhra Panda, Anna Stouffer, John D. Tovar, and Andrew L. Ferguson. Hybrid computational–experimental data-driven design of self-assembling -conjugated peptides. *Digital Discovery*, 1(4):448–462, 2022. doi:[10.1039/D1DD00047K](https://doi.org/10.1039/D1DD00047K).
- [232] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014. doi:[10.1038/sdata.2014.22](https://doi.org/10.1038/sdata.2014.22).

- [233] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9:513–530, 2018. doi:[10.1039/C7SC02664A](https://doi.org/10.1039/C7SC02664A).
- [234] Austin Clyde, Thomas Brettin, Alexander Partin, Hyunseung Yoo, Yadu Babuji, Ben Blaiszik, Andre Merzky, Matteo Turilli, Shantenu Jha, Arvind Ramanathan, and Rick Stevens. Protein–ligand docking surrogate models: A sars-cov-2 benchmark for deep learning accelerated virtual screening. *arXiv preprint arXiv:2106.07036*, 2021. Surrogate docking models achieve up to 10^6 x throughput compared to standard docking.
- [235] Yun Gy Chung et al. Recent trends of machine learning in metal–organic frameworks. *ACS Materials Letters*, 2024. Overview of ML methods for accelerating MOF design and screening.
- [236] Jianming Mao, Ningxin Jiang, Andrea Darù, Alexander S. Filatov, Jessica E. Burch, Jan Hofmann, Simon M. Vornholt, Karena W. Chapman, John S. Anderson, and Andrew L. Ferguson. Structure and synthesizability of iron–sulfur metal–organic frameworks. *Journal of the American Chemical Society*, 147(21):17651–17667, 2025. doi:[10.1021/jacs.4c16341](https://doi.org/10.1021/jacs.4c16341).
- [237] Alexandra Gessner, Sebastian W. Ober, Owen N. Vickery, Dino Oglic, and Talip Usar. Active learning for affinity prediction of antibodies. *arXiv preprint arXiv:2406.07263*, 2024. Demonstrates active learning loop combining RBFE simulations and BO for antibody optimization.
- [238] Asif Khan, Alexander I. Cowen-Rivers, Antoine Grosnit, Derrick-Goh-Xin Deik, Philippe A. Robert, Victor Greiff, Eva Smorodina, Puneet Rawat, Kamil Dreckowski, Rahmad Akbar, Rasul Tutunov, Dany Bou-Ammar, Jun Wang, Amos Storkey, and Haitham Bou-Ammar. Antbo: Towards real-world automated antibody design with combinatorial bayesian optimisation. *arXiv preprint arXiv:2201.12570*, 2022. Combinatorial BO framework for CDRH3 sequence optimization.
- [239] Tanya Pavlova et al. Bayesian optimization in bioprocess engineering—where do we go from here? *Biotechnology and Bioengineering*, 2024. doi:[10.1002/bit.28960](https://doi.org/10.1002/bit.28960). Discusses BO strategies in complex, multi-endpoint biological systems including immune response.
- [240] Jiahui Zhou and Meilan Huang. Navigating the landscape of enzyme design: from molecular simulations to machine learning. *Chemical Society Reviews*, 53(14):4275–4300, 2024. doi:[10.1039/d4cs00196f](https://doi.org/10.1039/d4cs00196f).
- [241] 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.

- [242] Alan Grossfield, Paul N Patrone, Daniel R Roe, Andrew J Schultz, Daniel W Siderius, and Daniel M Zuckerman. Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1. 0]. *Living J. Comput. Mol. Sci.*, 1(1), 2018.
- [243] Amish J Patel, Patrick Varilly, and David Chandler. Fluctuations of water near extended hydrophobic and hydrophilic surfaces. *J. Phys. Chem. B*, 114(4):1632–1637, 2010.
- [244] Joel M Bowman, Gabor Czako, and Bina Fu. High-dimensional ab initio potential energy surfaces for reaction dynamics calculations. *Phys. Chem. Chem. Phys.*, 13(18):8094–8111, 2011.
- [245] Jeffrey Comer, James C Gumbart, Jérôme Hénin, Tony Lelièvre, Andrew Pohorille, and Christophe Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B*, 119(3):1129–1151, 2015.
- [246] Christoph Dellago, Peter G Bolhuis, Félix S Csajka, and David Chandler. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.*, 108(5):1964–1977, 1998.
- [247] Titus S Van Erp, Daniele Moroni, and Peter G Bolhuis. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.*, 118(17):7762–7774, 2003.
- [248] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141–151, 1999.
- [249] E Weinan, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66(5):052301, 2002.
- [250] Rosalind J Allen, Patrick B Warren, and Pieter Rein Ten Wolde. Sampling rare switching events in biochemical networks. *Phys. Rev. Lett.*, 94(1):018104, 2005.
- [251] Bernd A Berg and Thomas Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.*, 68(1):9, 1992.
- [252] 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.
- [253] Arthur F Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.*, 78(20):3908, 1997.
- [254] Helmut Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E*, 52(3):2893, 1995.
- [255] Pratyush Tiwary and Axel van de Walle. A review of enhanced sampling approaches for accelerated molecular dynamics. *Multiscale materials modeling for nanomechanics*, 245:195–221, 2016.

- [256] Baron Peters. Reaction coordinates and mechanistic hypothesis tests. *Annu. Rev. Phys. Chem.*, 67:669–690, 2016.
- [257] K. L. Young, M. B. Ross, M. G. Blaber, M. Rycenga, M. R. Jones, C. Zhang, A. J. Senesi, B. Lee, G. C. Schatz, and C. A. Mirkin. Using dna to design plasmonic metamaterials with tunable optical properties. *Advanced Materials*, 26(4):653–659, 2014. doi:[10.1002/adma.201302938](https://doi.org/10.1002/adma.201302938). URL <https://doi.org/10.1002/adma.201302938>.
- [258] J. G. Heddle. Gold nanoparticle-biological molecule interactions and catalysis. *Catalysts*, 3(3):683–708, 2013. doi:[10.3390/catal3030683](https://doi.org/10.3390/catal3030683). URL <https://doi.org/10.3390/catal3030683>.
- [259] C. Zhang, R. J. Macfarlane, K. L. Young, C. H. J. Choi, L. Hao, E. Auyeung, G. Liu, X. Zhou, and C. A. Mirkin. A general approach to dna-programmable atom equivalents. *Nature*, 495(7440):348–352, 2013. doi:[10.1038/nature11975](https://doi.org/10.1038/nature11975).
- [260] Florian Gisperm, Robert Klausser, Mohamed ElShazly, Julian Kopp, Eva Brichová, and Oliver Spadiut. Bayesian optimization in bioprocess engineering—where do we stand today? *Biotechnology and Bioengineering*, 122(6):1313–1325, 2025. doi:[10.1002/bit.28960](https://doi.org/10.1002/bit.28960).
- [261] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. *Proteins: Struct., Funct., Genet.*, 78:1950, 2010.