

THE UNIVERSITY OF CHICAGO

RECONSTRUCTING DYNAMICAL SYSTEMS FROM TIME SERIES USING STATISTICAL
LEARNING AND TAKENS' DELAY EMBEDDING THEOREM

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

DEPARTMENT OF PHYSICS

BY
MAXIMILIAN TOMÁS TOPEL

CHICAGO, ILLINOIS

JUNE 2025

Copyright 2025 by Maximilian Tomás Topel
All Rights Reserved

A mis padres,
por todo lo que me enseñaron.

Yo sé los nombres extraños
De las yerbas y las flores,
Y de mortales engaños,
Y de sublimes dolores.

- José Martí

ABSTRACT

The field of dynamical systems studies how systems evolve in time according to deterministic, probabilistic, or hybrid rule sets. Embedding theorems can be used to describe how observations of these systems can be used to reconstruct their dynamics. However, these theorems do not provide a backmapping to the full-dimensional descriptions of these dynamics. The recovery of these full-dimensional descriptions from time series observations, such as those furnished by experiments, is central to the study of dynamical systems. In this work we propose and validate a pipeline that combines embedding theorems and statistical learning to represent dynamics from time series in a mathematically rigorous way and learn backmappings to full dimensional descriptions with universal function approximators in the context of molecular dynamics. We demonstrate that this framework is robust under noise and extend it for use with arbitrary dynamical systems. We introduce the TAKens Reconstruction (TAR) software package to perform univariate, multivariate and multi-temporal Takens' reconstructions of arbitrary time series and apply the work beyond physics, in the context of economic systems. Lastly, we investigate optimization of observable choice to maximize reconstruction quality. This work provides a theoretical and computational framework with which we can reconstruct and analyze the dynamics of arbitrary dynamical systems. It constitutes a contribution to a broader class of techniques bridging the gap between experimental observation and theoretical understanding of dynamics.

TABLE OF CONTENTS

LIST OF FIGURES	ix
LIST OF TABLES	xviii
1 INTRODUCTION	1
1.1 On the Reconstruction of Dynamical Systems	1
1.2 Takens' Reconstruction of Single Molecules	3
1.3 Takens' Reconstruction of Dynamics from Arbitrary Time Series	6
1.4 Molecular Probe Design For Per- and polyfluoroalkyl Substance Removal	8
1.5 Publications	10
2 RECONSTRUCTION OF PROTEIN STRUCTURES FROM SINGLE-MOLECULE TIME SERIES	12
2.1 Abstract	12
2.2 Introduction	13
2.3 Methods	15
2.3.1 Principles of STAR	15
2.3.2 Molecular Dynamics Simulations	26
2.4 Results and Discussion	28
2.4.1 C ₂₄ H ₅₀	29
2.4.2 Chignolin	33
2.5 Conclusions	36
3 LEARNED RECONSTRUCTION OF PROTEIN FOLDING TRAJECTORIES FROM NOISY SINGLE-MOLECULE TIME SERIES	41
3.1 Abstract	41
3.2 Introduction	42
3.3 Methods	45
3.3.1 Principles of STAR	45
3.3.2 Molecular Dynamics Simulations	53
3.4 Results and Discussion	54
3.4.1 Chignolin	54
3.4.2 Villin	60
3.5 Conclusions	66
3.6 Appendix	69
3.6.1 Synthetic smFRET noise model	69
3.7 Supplementary Methods	73
3.8 Supplementary Figures	80
4 OPTIMIZING DATA STRUCTURE TO LEARN DYNAMICS FROM TIME SERIES DATA	83

4.1	Abstract	83
4.2	Introduction	84
4.3	Dynamics from Time Series	87
4.3.1	Embedding theorems	87
4.3.2	Dimensionality Reduction and Manifold Learning	90
4.4	Construction of Delay Vectors	91
4.5	Computational Application: Architecture of Takens Reconstruction (TAR)	94
4.5.1	Data Structuring	94
4.5.2	Manifold Learning	94
4.5.3	Reconstruction	96
4.5.4	Synthesis	97
4.6	Dynamical systems metrics	97
4.7	Introducing Takens Informed Reconstruction: the Lotka Volterra Model	101
4.8	Experimental Observable Optimization: Molecular Dynamics of a Straight Chain Hydrocarbon	104
4.9	Multi-Temporal Takens' Theorem: Protein Dynamics and Folding	109
4.10	Phase Space Reconstruction in the absence of training data: Vanguard S&P 500 ETF	115
4.11	Conclusions and Future Work	118
4.12	Supporting Information	120
4.12.1	Information Theoretic Measurements for Determining Time Delay of Em- beddings	120
4.12.2	2D Lotka Volterra Predator Prey Dynamics	121
4.12.3	C ₂₄ H ₅₀ molecular dynamics and synthetic smFRET	122
4.12.4	Villin molecular dynamics and synthetic smFRET	125
5	KOLMOGOROV-SINAI ENTROPIES IDENTIFY OPTIMAL OBSERVABLES FOR PRE- DICTION AND DYNAMICS RECONSTRUCTION IN CHAOTIC SYSTEMS	128
5.1	Abstract	128
5.2	Introduction	129
5.3	Ergodic Dynamics and Attractors	130
5.4	Takens' Delay Embedding Theorem	131
5.5	Distortions to Embedded Manifolds	132
5.6	Oseledec's Multiplicative Ergodic Theorem	134
5.7	Lyapunov Exponents and the Kolmogorov Sinai Entropy	136
5.8	Theorem and Proof	137
5.8.1	Limitations	144
5.9	Conclusion	145
6	DATA-DRIVEN DISCOVERY OF LINEAR MOLECULAR PROBES WITH OPTIMAL SELECTIVE AFFINITY FOR PFAS IN WATER	147
6.1	Abstract	147
6.2	Introduction	148
6.3	Methods	153
6.3.1	Estimation of sensitivity ($-\Delta G_{PFOS}$) and selectivity ($-\Delta\Delta G_{PFOS-SDS}$) . .	153

6.3.2	Encoding of probe molecules using pre-trained variational autoencoder with SELFIES representation	156
6.3.3	Construction of Gaussian process regression (GPR) surrogate models . . .	158
6.3.4	Multi-objective Bayesian optimization (BO)	159
6.3.5	Training of LASSO regression models to infer important chemical substructures of probes	161
6.4	Results and Discussion	163
6.4.1	Discovery of optimal linear molecular probes using computational active learning	163
6.4.2	Identification of chemical substructures that are the principal determinants of sensitivity and selectivity	167
6.4.3	Correlation analysis of important identified chemical substructures	170
6.5	Conclusions	172
7	ESTIMATING BINDING FREE ENERGIES OF PERFLUOROOCTANESULFONIC ACID (PFOS) WITH CYCLODEXTRIN PROBES	176
7.1	Abstract	176
7.2	Introduction	177
7.3	Methods	178
7.3.1	All-atom molecular dynamics simulations of PFOS, SDS and TCAA interactions with cyclodextrin probes	179
7.3.2	Enhanced sampling using funnel well-tempered parallel bias metadynamics	181
7.3.3	Estimation of equilibrium binding free energies ΔG and binding constants K_b	185
7.3.4	Evaluation of binding energies	190
7.4	Conclusions	197
7.5	Acknowledgments	198
8	CONCLUSIONS	199
8.1	Summary of Contributions	199
8.2	A review of ideas that did not work and future directions	200
8.2.1	Generative Techniques in Latent Space and Latent Space Liouville Propagators	200
8.2.2	Botany Meets Dynamics	201
8.2.3	Econometrics	203
8.2.4	Topology, Topography and Holonomic Constraints	205
8.3	Contradictory Advice	206
	REFERENCES	208

LIST OF FIGURES

- 2.1 Cartoon schematic of Single-molecule TAKens Reconstruction (STAR). The trained STAR model predicts molecular configurations of a protein $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ from univariate time series in an experimentally-measurable observable of the system $v(t) \in \mathbb{R}^1$ through the four-step pathway $b \rightarrow d \rightarrow c \rightarrow e \rightarrow f$. Each panel corresponds to a different representation of the molecular system and arrows between them to mathematical operations. Red arrows represent unsupervised nonlinear manifold learning problems (i.e., nonlinear dimensionality reduction), blue arrows represent supervised learning problems (i.e., function approximation), and grey arrows represent deterministic operations. (a,b) Molecular dynamics simulations sufficiently long to sample the thermally-relevant configurational space provide training trajectories of molecular configurations $\mathbf{r}(t) \in \mathbb{R}^{3N}$ (panel a) and a scalar time series in an experimentally-measurable system observable $v(t) \in \mathbb{R}^1$ (panel b). In this work, we set $v(t)$ to be the head-to-tail ($h2t$) intramolecular distance that is, in principle, measurable by sm-FRET. (c) Interactions between atoms constrain the molecular trajectory $\mathbf{r}(t) \in \mathbb{R}^{3N}$ to a k -dimensional manifold $\mathcal{M} \subset \mathbb{R}^k \subset \mathbb{R}^{3N}$ with effective dimensionality $k \ll 3N$. The collective variables $\{\psi_1, \psi_2, \dots, \psi_k\}$ parameterizing the manifold are extracted from the simulation trajectory using diffusion maps. The manifold $\mathcal{M}$ supports the smFES $\beta F(\psi_1, \psi_2, \dots, \psi_k)$. (d) An image of $\mathcal{M}$ and the smFES is obtained from the time series data $v(t) \in \mathbb{R}^1$ by forming a d -dimensional delay embedding $\mathbf{y}(t) = [v(t), v(t - \tau), v(t - 2\tau), \dots, v(t - (d - 1)\tau)] \in \mathbb{R}^d$ and then applying diffusion maps to learn a parameterization $\{\psi'_1, \psi'_2, \dots, \psi'_k\}$ of a manifold $\mathcal{M}' \subset \mathbb{R}^k \subset \mathbb{R}^d$. Under technical conditions on v , d , and τ discussed in the main text, Takens' Delay Embedding Theorem asserts that the dynamical evolution of $\mathbf{y}(t)$ is C^1 -equivalent to that of $\mathbf{r}(t)$ and that the manifold $\mathcal{M}'$ is related to $\mathcal{M}$ via diffeomorphic (i.e., smooth, invertible, and bijective) transformation $\Theta : \mathcal{M}' \rightarrow \mathcal{M}$. We learn the transformation Θ from the training data using a simple artificial neural network. (e,f) The manifold $\mathcal{M} \subset \mathbb{R}^k$ contains a low-dimensional projection of $\mathbf{r}(t) \in \mathbb{R}^{3N}$ into $\{\psi_i\}_{i=1}^k$ from which molecular configurations $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ may be approximately reconstructed. We perform the reconstruction by first learning the pairwise distances between all atoms $\mathbf{d} \in \mathbb{R}^{N(N-1)/2}$ using an artificial neural network (panel e) and then using classical multidimensional scaling to deterministically transform this into the reconstructed atomic coordinates $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ (panel f). All molecular renderings in this work were constructed using VMD Humphrey et al. [1996]. 16

- 2.2 Application of STAR to $C_{24}H_{50}$ polymer chain. (a) Synthetic idealized smFRET time trace of the head-to-tail distance $h2t(t)$ between the terminal united atoms computed over a 100 ns MD trajectory with frames saved every 0.2 ps. The first 20 ns are used for training (orange) and the remaining 80 ns for testing (blue). The molecule undergoes hundreds of folding and unfolding events over the course of the simulation. (b) The intrinsic manifold $\mathcal{M}$ spanned by the dMap CVs $\{\psi_1(t), \psi_2(t)\}$ and supporting the smFES $\beta F(\psi_1, \psi_2)$. The Gibbs free energy F is dedimensionalized by the inverse temperature $\beta = 1/k_B T$. The arbitrary zero of F is specified to lie at the global free energy minimum. (c) RMSD reconstruction accuracy as a function of location on the intrinsic manifold $\mathcal{M}$. (d) Molecular reconstructions using the trained STAR pipeline of five representative points A-E in the $h2t(t)$ time series spanning the test set. Reconstructions $\hat{\mathbf{r}}$ (red) are superposed on the corresponding true configurations $\mathbf{r}$ (blue) extracted directly from the MD simulation. The head-to-tail distance of the true configuration and the RMSD under translational, rotational, mirror, and head-tail alignment between the true and reconstructed configurations are reported under each image. The STAR prediction of the location of each point on the smFES is illustrated in panel b and the corresponding dimensionless free energy βF reported under each image. 32
- 2.3 Application of STAR to the 10-residue engineered mini-protein Chignolin. (a) Synthetic idealized smFRET time trace of the head-to-tail distance $h2t(t)$ between the terminal heavy atoms computed over a 20 μ s MD trajectory with frames saved every 200 ps. The first 5 μ s are used for training (orange) and the remaining 15 μ s for testing (blue). The molecule undergoes dozens of folding and unfolding events over the course of the trajectory. (b) The smFES $\beta F(\psi_1, \psi_2)$ supported by the intrinsic manifold $\mathcal{M}$ has its arbitrary zero of F specified to lie at the global free energy minimum. (c) RMSD reconstruction accuracy as a function of location on the intrinsic manifold $\mathcal{M}$. Molecular reconstructions using the trained STAR pipeline of five representative points A-E selected from the testing $h2t(t)$ time series rendered in (d) ribbon and (e) licorice representations. Reconstructions $\hat{\mathbf{r}}$ (red) are superposed on the corresponding true configurations $\mathbf{r}$ (blue) extracted directly from the MD simulation. The head-to-tail distance of the true configuration and the RMSD under translational, rotational, and mirror alignment between the true and reconstructed configurations are reported under each pair of images. The STAR prediction of the location of each point on the smFES is illustrated in panel b and the corresponding dimensionless free energy βF reported under each image. 35

3.1	Cartoon of the layout of Single Molecule TAKens Reconstruction (STAR). STAR takes time series in one or more coarse-grained observables of the molecular system such as an intramolecular distance furnished by smFRET $v(t) \in \mathbb{R}^1$ and reconstructs a molecular trajectory within the Cartesian coordinate space of N atoms $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ through the three-step pathway $b \rightarrow d \rightarrow c \rightarrow e$. Each panel corresponds to a different representation of a molecular trajectory and the arrows between the panels correspond to learning tasks to convert one representation to another. (a,b) For the purposes of training the STAR pipeline we collect all-atom MD trajectories recording the Cartesian coordinates of the N constituent atoms $\mathbf{r}(t) \in \mathbb{R}^{3N}$ and from these construct synthetic smFRET time series in a single intramolecular distance $v(t) \in \mathbb{R}^1$. Once the STAR pipeline is trained, we no longer need any additional MD simulation data and the pipeline can operate on new synthetic (or experimental) smFRET data collected under similar conditions. Following the prescription of Takens' Theorem we construct a delay embedding of the scalar time series as $\mathbf{y}(t) = [v(t), v(t - \tau), v(t - 2\tau) \dots v(t - (d - 1)\tau)] \in \mathbb{R}^d$, where τ is the delay time and d is the delay dimensionality. (c) The k -dimensional manifold containing the all-atom simulation trajectory $\mathcal{M} \subset \mathbb{R}^k \in \mathbb{R}^{3N}$ and spanned by the nonlinear collective variables $\{\Psi'_1, \Psi'_2, \dots, \Psi'_k\}$. Collective couplings between the molecular degrees of freedom leads to an emergent low-dimensionality wherein $k \ll 3N$. We learn $\mathcal{M}$ from the MD simulation trajectory $\mathbf{r}(t)$ using diffusion maps unsupervised nonlinear manifold learning. (d) The k -dimensional manifold containing the smFRET delay embedding $\mathcal{M}' \subset \mathbb{R}^k \in \mathbb{R}^d$ spanned by the nonlinear collective variables $\{\Psi'_1, \Psi'_2, \dots, \Psi'_k\}$. We learn $\mathcal{M}'$ from the delay embedding $\mathbf{y}(t)$ using diffusion maps. Takens' Theorem asserts $\mathcal{M}'$ is an image of $\mathcal{M}$ and, under some technical conditions on symmetries and periodicities, they are related by a diffeomorphism (i.e., a smooth and invertible transformation) $\Theta : \mathcal{M}' \rightarrow \mathcal{M}$ that we learn by training a simple fully-connected feedforward artificial neural network as a flexible and expressive nonlinear function approximator. (e) Reconstruction of the atomistic molecular trajectory within the Cartesian coordinate space of N atoms $\hat{\mathbf{r}} \in \mathbb{R}^{3N}$. We learn the molecular reconstruction from the manifold $\mathcal{M}$ using a second fully-connected feedforward artificial neural network. All molecular renderings were made using Visual Molecular Dynamics Humphrey et al. [1996].	47
3.2	Heavy atom RMSD reconstruction accuracies for Chignolin as a function of training data volumes Dv , time bin resolution λ , and signal-to-noise ratio I_D	56

- 3.3 Manifolds $\mathcal{M}$ and position-dependent heavy atom RMSD reconstruction accuracy for Chignolin as a function of time bin resolution λ . (a) All-atom manifolds $\mathcal{M}$ within STAR pipelines trained at $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ for each value of $\lambda = \{0.2, 1, 2, 5, 10\}$ ns. Manifolds are represented as scatterplots showing the embedding of each of the $Dv = 4 \times 10^4$ MD configurations $\mathbf{r}$ into the two diffusion map eigenvectors $\{\Psi_1, \Psi_2\}$ spanning each manifold. Each point is colored by the associated free energy $\beta F(\Psi_1, \Psi_2)$ of that point computed by collecting histograms over the empirical probability distribution at a bin size of $\{\Delta\Psi_1 = 5 \times 10^{-7}, \Delta\Psi_2 = 5 \times 10^{-7}\}$. The arbitrary zero of free energy in each panel is specified by subtracting computed energy of the minimum free energy point on the landscape from all other points. Five representative configurations A-E are selected from the MD test trajectory and projected onto each manifold: A and B reside in the folded macrostate, C in the transition region, and D and E in the unfolded ensemble. (b) Visualizations of the reconstructions $\hat{\mathbf{r}}$ of each configuration A-E at each time resolution $\lambda = 0.2$ ns (blue), 1 ns (green), 2 ns (cyan), 5 ns (magenta), and 10 ns (yellow). For visual clarity, configurations are represented as ribbons tracing the backbone of the protein and are superposed upon the true configuration extracted from the MD test trajectory $\mathbf{r}$ (red). (c) Heavy atom RMSD reconstruction corresponding to each state are listed next to each image employing the same color-coding as the reconstructions. All molecular visualizations are constructed using VMD Humphrey et al. [1996]. 59
- 3.4 Heavy atom RMSD reconstruction accuracies for Villin as a function of training data volumes Dv , time bin resolution λ , and signal-to-noise ratio I_D 62
- 3.5 Manifolds $\mathcal{M}$ and heavy atom RMSD reconstruction accuracies for Villin as a function of time bin resolution λ . (a) All-atom manifolds $\mathcal{M}$ within STAR pipelines trained at $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ for each value of $\lambda = \{0.2, 1, 2, 10, 20\}$ ns. Manifolds are represented as scatterplots showing the embedding of each of the $Dv = 4 \times 10^4$ MD configurations $\mathbf{r}$ projected into the two leading diffusion map eigenvectors $\{\Psi_1, \Psi_2\}$ spanning each manifold. Zoomed in cutouts are provided for the sub-panels B-E. Color distributions characterize associated free energy $\beta F(\Psi_1, \Psi_2)$ of each point computed by collecting histograms over the empirical probability distribution at a bin size of $\{\Delta\Psi_1 = 5 \times 10^{-7}, \Delta\Psi_2 = 5 \times 10^{-7}\}$. The arbitrary zero of free energy in each panel is specified by subtracting computed energy of the minimum free energy point on the landscape from all other points. Five representative configurations A-E are selected from the MD test trajectory and projected onto each manifold. A and B reside in the folded macrostate while the C, D and E in the unfolded ensemble. (b) Visualizations of the reconstructions $\hat{\mathbf{r}}$ of each configuration A-E at each time resolution $\lambda = 0.2$ ns (blue), 1 ns (green), 2 ns (cyan), 10 ns (yellow), and 20 ns (black). For visual clarity, configurations are represented as ribbons tracing the backbone of the protein and are superposed upon the true configuration extracted from the MD test trajectory $\mathbf{r}$ (red). (c) RMSD reconstruction corresponding to each state are listed next to each image employing the same color-coding as the reconstructions. All molecular visualizations are constructed using VMD Humphrey et al. [1996]. 64

3.6	Variation of σ_r/r defining the relative uncertainty in r as a function of the donor-acceptor distance r at various values of donor intensity I_D (Eqn. 3.13). For the purposes of this calculation we adopt prototypical values of $\phi_D = \phi_A = 1$ for the donor and acceptor quantum yields and $R_0 = 5$ nm for the FRET radius of the donor-acceptor pair.	72
3.7	Eigenvalue spectrum resulting from the application of diffusion maps to the MD trajectory $\mathbf{r}(t) \in \mathbb{R}^{3N}$ for Chignolin. A gap in the eigenvalue spectrum after the second non-trivial eigenvalue λ_2 is identified using the L-method of Salvador and Chan Salvador and Chan [2004] and informs extraction of a 2D intrinsic manifold $\mathcal{M}$ spanned by $\{\Psi_1, \Psi_2\}$.	80
3.8	Eigenvalue spectrum resulting from the application of diffusion maps to the MD trajectory $\mathbf{r}(t) \in \mathbb{R}^{3N}$ for Villin. A gap in the eigenvalue spectrum after the third non-trivial eigenvalue λ_3 is identified using the L-method of Salvador and Chan Salvador and Chan [2004] and informs extraction of a 3D intrinsic manifold $\mathcal{M}$ spanned by $\{\Psi_1, \Psi_2, \Psi_3\}$.	81
3.9	All-atom manifolds $\mathcal{M}$ for Villin within STAR pipelines trained at $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ for each value of $\lambda = \{0.2, 1, 2, 10, 20\}$ ns. The 3D manifolds in (Ψ_1, Ψ_2, Ψ_3) are represented as scatterplots showing the embedding of each of the $Dv = 4 \times 10^4$ MD configurations $\mathbf{r}$ into the (a) $\{\Psi_1, \Psi_3\}$ and (b) $\{\Psi_2, \Psi_3\}$ projections. Color distributions characterize associated free energy $\beta F(\Psi_i, \Psi_j)$ of each point computed by collecting histograms over the empirical probability distribution at a bin size of $\{\Delta\Psi_1 = 5 \times 10^{-7}, \Delta\Psi_3 = 5 \times 10^{-7}\}$ and $\{\Delta\Psi_2 = 5 \times 10^{-7}, \Delta\Psi_3 = 5 \times 10^{-7}\}$. The arbitrary zero of free energy in each panel is specified by subtracting computed energy of the minimum free energy point on the landscape from all other points. The markers A-E indicate five representative configurations selected from the MD test trajectory and for which molecular visualizations are presented in Fig. 5b.	82
4.1	Schematic of the TAR pipeline. (a) TAR is given an input time series. (b) The time series is structured into either univariate, multivariate or multi-temporal Takens delay vectors generating our Takens embedding manifold. (c) Latent space extraction using diffusion maps manifold learning to learn latent space embedding $\mathbf{I}, \mathcal{M}'$. (e) Extraction of a latent space embedding, $\mathcal{M}$, from a full-dimensional training trajectory. This step is only required if a training trajectory is available. (d) Reconstruction by backmapping to the latent space representation of the full-dimensional system learned from the full-dimensional system observations in (e). If no full-dimensional training trajectory is available, we learn the map from (c) to (e) directly via an artificial neural network.	95
4.2	Lotka Volterra model with TAR reconstructions using a variety of time delays. (A) Phase space diagram of X, Y populations. (B) Reconstruction and true populations of species Y over a 25 time unit period from the test set with $\tau = [0.001, 0.01, 0.1, 1, 10]$ with $(2m\tau) = 11$. (C) Reconstruction error in phase space by time delay, τ and embedding dimension, $2m + 1$, used to construct delay embedding vectors. We compare these errors to a benchmark of the minimal root mean squared distance point in the training population trajectory compared with all other frames ≈ 68 population units.	100

4.3	Reconstructions of the latent space representation of the straight chain hydrocarbon $C_{24}H_{50}$ as a function of atom index, n . (A) Visualization of the $C_{24}H_{50}$ including a label of the 1^{st} , n^{th} and 24^{th} atom in the molecule. We take as our set of available observables the distance between the 1^{st} and n^{th} atom. The rendering was performed using Avogadro Hanwell et al. [2012b] (B) Mean reconstruction errors for TAR models trained on time series of 1^{st} and n^{th} atom distance plotted against Kolmogorov-Sinai Entropies of each time series. Reconstruction errors correspond to the test set and can be compared to a benchmark of 0.25 nm corresponding to the minimum RMSD frame against all other frames. We plot 3 types of TAR models; In cyan we have TAR models trained over 1 000 points with delay dimensionality $(2m + 1)$ chosen using the E1(d) method of Cao Cao et al. [1998]. In yellow, we present TAR models trained over 20 000 points with delay dimensionality $(2m + 1)$ chosen using the E1(d) method of Cao Cao et al. [1998]. In red, we present TAR models trained over 20 000 points with delay dimensionality $(2m + 1 = 5)$ corresponding to $m = k = 2$, the intrinsic dimensionality of the attractor.	105
4.4	Schematic of multi-temporal TAR applied to the protein Villin with reconstruction errors corresponding to delay time choice. (A) The multi-temporal TAR pipeline. Beginning with a time series, we construct p sets of time delay vectors with $\tau_1, \tau_2 \dots \tau_p$ as the time delays and construct p latent space manifolds. We learn the map from $\mathcal{M}_1, \mathcal{M}_2 \dots \mathcal{M}_p$ to $\mathcal{M}_{total}$ via an ANN and reconstruct the full-dimensional molecule as in univariate TAR. (B) is a VMD rendering of the protein Villin (PDB:2F4K) Humphrey et al. [1996]. (C) describes mean reconstruction errors of the full-dimensional atomic trajectory over the course of the test set for each of the tested observables with benchmark line in red corresponding to the minimum RMSD frame against all other frames. Uncertainties are bootstrapped by dividing the delay vectors and trajectory in to 3 equal length portions, training TAR on a subset of each and testing on the others. On the far left we have a univariate model with $\tau = 1 \mu s$ chosen from the first minimum of autocorrelation. All other models are multi-temporal including $\tau_1 = 200 ps$ and all τ up to τ_p listed (i.e. $\tau = 3$ corresponds to a multi-temporal model using τ_1, τ_2 and τ_3).	110
4.5	Vanguard S&P 500 ETF (VOO) returns from January 2024 Services [2024] for 5 second, 1 minute and 1 hour time delays returns. (A) measures hurst exponents over a 100 ms to 2 hour time period. (B) relays the L1 norm between the true returns distributions of VOO and (i) TAR models (ii) AR(1) (autoregressive) model (iii) Gaussian distribution. (C) Histogram of returns over $\tau = 5$ s, $\tau = 1$ min and $\tau = 1$ h time periods with both AR(1) and Gaussian fits using all available data. (D) histogram of returns in TAR produced phase space for both a train and test set with training performed on $\min[5000, \text{len}(\text{delay_vectors})//2]$	114
4.6	(a) Kolmogorov-Sinai Entropies (b) Leading Lyapunov Exponents as a function of number of points used in training for time series of the distance between the first and i^{th} atom in $C_{24}H_{50}$	125

- 6.1 Illustrations of the target PFAS analyte, interferent, and molecular search space over which the probes with optimal sensitivity and selectivity are discovered. (a) 3D structure of the target PFAS variant (perfluorooctanesulfonic acid, PFOS), (b) 3D structure of the template interferent (sodium dodecyl sulfate, SDS), and (c) search space comprising linear hydrogenated or halogenated chains of three to fourteen carbons with and without amine or phosphine based head groups. All semi-halogenated probes considered contain contiguous blocks of halogenated methylene groups for experimental characterizability. We explored primary, secondary, tertiary and quaternary states for the amine based head group and only the primary, secondary and tertiary states of the phosphine head groups. In total, our search space comprises 3850 molecules over which we employ our computational active learning and high-throughput screening approach to discover optimal molecular probes. The 3D structures are rendered using VMDHumphrey et al. [1996], where carbon, hydrogen, oxygen, nitrogen, sulfur, fluorine are colored in cyan, white, red, blue, yellow, and pink, respectively. 151
- 6.2 Illustration of the computational active learning approach employed to identify linear molecular probes with optimal sensitivity and selectivity. (a) Set up of all-atom and explicit molecular dynamics simulations and enhanced sampling using parallel bias metadynamics to estimate sensitivity ($-\Delta G_{PFOS}$) and selectivity ($-\Delta\Delta G_{PFOS-SDS}$), (b) embedding sensitivity and selectivity information generated through (b.i) performance of these enhanced sampling simulations and (b.ii) an encoding of molecular probes into a smooth low-dimensional latent space using a pre-trained variational autoencoder, (c) construction of Gaussian process regression (GPR) surrogate models over the latent space to predict probe sensitivity and selectivity, (d) multi-objective Bayesian optimization (BO) using random scalarizations to guide selection of the next molecular probes within the design space for molecular simulation. 154
- 6.3 Progress of the active learning campaign in advancing the sensitivity-selectivity frontier of PFOS probes within each active learning cycle. (a) Mean distance $d_{all}^{(j)}$ reporting the cumulative average over the d_i values of all n probes sampled so far up to and including cycle j . (b) Pareto distance $d_{Pareto}^{(j)}$ reporting an average over the d_i values of those probes constituting the Pareto frontier at the end of round j , including contributions to the frontier from all prior rounds. (c) Scatter plots of the sampled probes within the sensitivity-selectivity design space over the course of the active learning campaign. The sensitivity and selectivity increase in the positive x- and y-directions, respectively. The probes sampled in the present round are indicated by red markers and those accumulated from all previous rounds by grey markers. The Pareto frontier is indicated by the black line. The green shaded region highlights the shift in the Pareto frontier relative to cycle 0. The molecules in cycle 0 contains the 45 molecules used to train the initial GPR model. 164

6.4	Change in the distribution of sensitivity and selectivity of probes to PFOS with each active learning cycle. Violin plots of the (a) sensitivity ($-\Delta G_{PFOS}$) and (b) selectivity ($-\Delta\Delta G_{PFOS-SDS}$) as a function of the active learning cycle. Chemical structures of molecules with maximum sensitivity and selectivity among all the probes studied up to a given cycle are shown at the cycles in which they were discovered. The black solid line shows the change in the maximum sensitivity or selectivity of probes to PFOS with each active learning cycle. Both sensitivity and selectivity increase with increase in positive direction of y-axis.	166
6.5	Chemical substructures within the 252 molecular probes considered within our active learning search identified by LASSO regression to be the leading determinants of PFOS (a) sensitivity and (b) selectivity. Large LASSO weights w identify substructures that are the leading determinants of sensitivity or selectivity. The chemical substructures possessing the largest LASSO weights are illustrated in green on representative probe molecules containing these substructures. The vertical arrows on the y-axes indicate the direction of favorable LASSO weights.	168
6.6	Effect of probe length upon PFOS (a) sensitivity and (b) selectivity for fully hydrogenated (blue) and fully fluorinated (red) probes. Markers and error bars represent means and standard errors estimated by three-fold block averaging over a 900 ns production run. Standard propagation of errors is employed to obtain the standard errors on selectivity. The vertical arrows on the y-axes indicate the direction of favorable sensitivity and selectivity.	170
6.7	Effect of amine head group upon PFOS sensitivity and selectivity for fully hydrogenated (blue) and fully fluorinated (red) probes with a backbone length of (a) four, (b) eight, and (c) 12 single bonds. Markers and error bars represent means and standard errors estimated by three-fold block averaging over a 900 ns production run. Standard propagation of errors is employed to obtain the standard errors on selectivity. The vertical arrows on the y-axes indicate the direction of favorable sensitivity and selectivity.	172
7.1	Illustration of the cyclodextrin-centric coordinate system and PBMetaD CVs within this reference frame for the α -cyclodextrin and PFOS system. The XY plane is defined by the carbon and oxygen atoms that are colored in green in the narrow-side plane of each 6 sided D-glucose units in each cyclodextrin probe and the Z axis is defined as the normal vector to that plane. The three PBMetaD CVs correspond to (i) z : the Z coordinate of the COM of the target molecule, (ii) r : the radial distance of the COM of the target molecule away from the Z axis in the XY plane, and (iii) θ : the planar angle between a vector drawn between the tail and head of the target molecule and the Z axis. Atoms are color coded; Carbon, hydrogen, oxygen atoms are colored in cyan, white, and red, respectively. Something along these lines will be nice. Molecular renderings were produced with VMD Humphrey et al. [1996]	182

7.2	Illustration of convergence of the funnel well-tempered parallel bias metadynamics calculations for the α -cyclodextrin – PFOS system. (a-c) CVs z, r and θ (from left to right) as a function of simulation time.(d-f) Bias heights deposited as a function of time for CVs z, r and θ (from left to right). (g-i) ΔG s and related quantities computed from the simulation by using trajectory of length equal to the reported simulation time on x -axis. Points are computed at 10 ns intervals of simulation length. (g) Observed ΔG computed over the entire trajectory up until the reported simulation time. (h) Change in ΔG computed for each marginal 10 ns of simulation. (i) Standard error computed from 3 fold block averaging of ΔG over the trajectory of length equal to the reported simulation time on x -axis.	184
7.3	Univariate free energy landscapes $PMF(s)$ projected into the COM separation of the target and probe $s = \sqrt{r^2 + z^2}$ illustrating that we reach the effectively non-interacting regime in the far field where the free energy ceases to be a function of s . Landscapes are adjusted by the entropic correction accounting for the trivial increase in configurational phase space volume associated with increasing s and which physically corresponds to subtracting out the $PMF(s)$ corresponding to non-interacting point particles. (a) Describes $PMF(s)$ for PFOS. (b) Describes $PMF(s)$ for TCAA.	187
7.4	Molecular simulations of PFOS analyte and TCAA interferent binding with α -, β -, and γ -cyclodextrin probes.(a) Illustration of the cyclodextrin-centric coordinate frame. Potential of mean force (PMF) landscapes $PMF(r, z)$ are constructed in the radial separation r and axial separation z of the molecular centers of mass. PMFs for (a) α -cyclodextrin–PFOS (b) β -cyclodextrin–PFOS (c) γ -cyclodextrin–PFOS (d) α -cyclodextrin–TCAA (e) β -cyclodextrin–TCAA (f) γ -cyclodextrin–TCAA. For visualization convenience, the arbitrary additive constant for each PMF reflecting our	192
7.5	Calculated values of the standard state (a) binding free energy and (b) binding constant for α -, β -, and γ -cyclodextrin probes with the PFOS and TCAA target molecules. . . .	194
7.6	PMF landscapes and selectivity of β -CD with PFOS, SDS and TCAA. (a) PMF landscape of β -CD with PFOS. (b) PMF landscape of β -CD with SDS. (c) PMF landscape of β -CD with TCAA. Renderings of each system in its strongest bound state are presented above (a-c). d) diagram describing construction of r and z collective variables as a function of probe and analyte positioning. (e) Spatially resolved selectivity of β -CD with PFOS relative to SDS. (f) Spatially resolved selectivity of β -CD with PFOS relative to TCAA.	196

LIST OF TABLES

5.1	Types of deformations of the manifold $\mathcal{M}'$ under C^1 equivalence as guaranteed by Takens' Theorem that may vary with observable choice Hirsch et al. [2003], Takens [1981], Sauer et al. [1991].	134
7.1	Binding constants K_b for the PFOS analyte and TCAA interferent with the α -, β -, and γ -cyclodextrin probes calculated using enhanced sampling all-atom classical MD simulations. The selectivity of each cyclodextrin probe for PFOS relative to TCAA is quantified by the ratio of the binding constants $K_b^{\text{PFOS}}/K_b^{\text{TCAA}}$	194
7.2	Binding constant K_b computed under cutoff regimes from $0.8 \text{ nm} \leq s_{\text{bound}} \leq 1.6 \text{ nm}$. .	195
7.3	Binding free energies $\Delta G^\ominus$ computed under cutoff regimes from $0.8 \text{ nm} \leq s_{\text{bound}} \leq 1.6 \text{ nm}$	195

ACKNOWLEDGMENTS

Parashat Masei (Numbers 33) records the journey of the Israelites from Egypt to Israel. Though the journey should only take 40 days in principle, in practice, it is spread over 40 years with 42 stops along the way. It seems odd then that it be named masei, journeys, when the passage chronicles a largely static experience. In one of his sichos on the topic, Rebbe Menachem Mendel Schneerson explains that this naming has a philosophical origin. Journeys are multi-modal and development is complex. While we may be parked in a station of life at any given moment we continue to learn and grow. At each stage we learn different things and in this sense while we may be stopped in a new place, our growth in character remains in motion.

However, some lessons take longer than others... The PhD is a six year stop. Wandering in this desert for the past six years any sense of advancement is at best mercurial. And yet I have learned a thousand little lessons along the way, some portion of which I hope are conveyed in this manuscript, all of which are attributable to those around me who played a critical role in my academic and personal development. I cannot, in this brief space, thank all of the people who helped me upon this journey but I will take a crack at it.

My parents have, no doubt, been the most formidable and continuing source of support I have had over the years. My father taught me that satisficing is not the same as satisfying. While reading through 40+ year old math texts and being interested in problems that do not seem to be popular in my department or in lab can be lonely or challenging, this lesson kept me pursuing those problems that I found most intellectually rewarding. When in doubt, there is always one of his countless aphorisms to rely on; “mona se vista de seda, se queda mona”, “el golpe avisa”, “el chicharrón no es carne”, “veremos dijo el ciego” ... My mother taught me that there is also time in life for things outside of work, a lesson that I highly encourage those starting a PhD to avail themselves of. While their lessons are unenumerable, I am only able to employ them because of the hard work and sacrifice my grandparents put in leaving Cuba and making a new life in a foreign country. My

grandmother, Berta Topel, died just over a year before the publication of this dissertation but this work exists in no small part thanks to her actions and dedication to her family and our success.

From the University to the climbing gym and beyond, my friends have played the largest role in my personal growth over these past six years. I would like to take this opportunity to acknowledge a few who have had the biggest impacts. Those in my PhD cohort, particularly, Reginaldo Gomes, Efren Muñoz and Ben Soloway taught me more science than the all of my coursework at the University of Chicago put together. Waleed Al Rawi has helpfully questioned everything I have done since I got to the University of Chicago, a quality which has all but cured my woeful lack of introspection and helped me clarify both my scientific investigations and life goals. Bradley Tongco has been the soundest of sounding boards for almost a decade now by being the spitting image of the rational actor straight from an economics textbook (and is also single-handedly responsible for me occasionally putting in more hours on the lakefront trail than in lab). Zofia Caes, my running partner, my climbing partner and tireless supporter of my work ... there are no words.

Beyond my family and friends, I am most fortunate to have found a strong and vibrant Jewish community during my time here in Chicago. Rabbi Mendel Rapoport, Chabad and the Jewish Grad Student group are responsible for a cultural and spiritual learning I had neglected during those COVID years. While antisemitism is never far, 2023-2025 were particularly unpleasant years to be a Jew on a college campus. JGrads reminded me that my Judaism should be defined by learning, debate and community and not by the ignorance and bigotry of others, no matter how loud. From Shabbat dinners to Simchat Torah hakafot, Mendel and this community made Hyde Park home.

And now, as contractually obligated, I must thank my beloved University of Michigan, its Physics and Judaic Studies departments and those professors who showed me what learning for the sake of learning looks like. Professor Zvi Gitelman found it within himself to overlook my poor writing and guided me towards the stacks and the world of Jewish Studies. This set me on the academic path in a way my engineering classes had not. Professor Bryan Roby, then newly minted faculty in

the Judaic Studies Department, gave me my first research assistant role and showed me the depths of Humanities research. It was at this point I realized that academic research is a thing that people do. Professionally. I must admit I am only here because Professor Jean Krisch told me what a PhD was sometime in my third year of college. Professor Nicholas Kotov gave me my first real science research role, and it was at this point that I realized that at least, in principle, I could do some calculation on a piece of paper or on a computer that could inform a complex and intricate experiment. Incidentally, I asked him once if I should try something more sophisticated than my pen and paper toy models like say, this really popular software I found on the internet, Gromacs. He told me it was not necessary and I should stay away. This was good advice.

Finally, the academic work in these pages is a product of the laboratories I have worked in and with over the course of this PhD. My advisor, Andrew Ferguson, gave me the tools, guidance and patience required for me to get oriented in this machine learning and embedding theorems space. It probably took three or four years before I really understood what questions I was personally interested in answering and where my work would fit in to the literature. I really do appreciate ability to do all of the wandering that I did and for the foresight that he had in putting me on this embedding theorems project. I really cannot think of any other subfield I would have rather landed on. Dr. Siva Dasetty taught me more about computers than anyone else. While much of that was in the context of Gromacs, his truly undying patience is responsible for much of my understanding of molecular dynamics. Additionally, the depth and scope of his scientific pursuits and continual learning has made him a role model in my own journey to become a good scientist. I would also like to thank Professors Michael Benzaquen and Jean-Philippe Bouchaud for their generosity in hosting me at the EconophysiX lab so that I could pursue these ideas in embedding theorems in economic contexts.

Perhaps the most fitting and consequently overused quote in acknowledgments and often attributed

to Newton;

“If I see further it is by standing on the shoulders of giants”

. Richard Southern attributes the aphorism to Bernard of Chartres, translating his version as “the modern scholar compared with the ancients was as a dwarf standing on the shoulders of a giant.” Southern [1970]. Just as he was humbled, recognizing that even Abelard was dwarfed by the vast depths of inherited knowledge, I now leave the University of Chicago with an ever deeper appreciation for the extra millennium of giants upon whose shoulders I stand. To those great academics who built the fields of dynamical systems, embedding theorems and mechanics - and to everyone who helped me on this journey - thank you all.

CHAPTER 1

INTRODUCTION

1.1 On the Reconstruction of Dynamical Systems

Dynamical Systems include any system evolving in time according to a set of rules — whether deterministic, probabilistic, or a combination of both. Consequently, there is very little we observe that the field of dynamical systems does not touch. With that incredible diversity comes a wide variety of information representation techniques. We typically write down the laws of physics as differential equations. Economic performance tends to be measured as time series while food webs are often represented as temporal networks. In each case, we aim to model the evolution of a system in time yet each approach comes with tradeoffs.

In physics, we are partial to differential equations whose forms can be used to describe how the physical laws governing the world propagate a system through phase space. This clear and explicit formulation can be scalable; with sufficient computing power and well-tested potentials, we can simulate large and complex interactions, such as the folding of proteins or the binding of molecules. This is the premise of Molecular Dynamics (MD), a computational technique whereby each atom or coarse graining of atoms in a molecule is propagated in time according to an interaction potential and Newton's laws of motion.

However, at such large scales, the individual bases of these spaces lose their interpretability. For instance, the utility of knowing, the x-axis coordinate of a particular carbon in a protein of thousands of atoms, surrounded by hundreds of thousands of atoms of solvent, is unclear. Nor can we derive a fundamental understanding of the dynamics from the millions of coupled differential equations that propagate this system. Moreover, for many systems, these differential equation representations are unavailable, as we know neither the correct basis to represent our system nor the correct forces to

propagate each degree of freedom in space.

Experimental observations offer a different window in to dynamics. While an experimental observation may suffer from stochastic or deterministic forcing of unknown origin, the recorded phenomena must have been produced by some physical means. Noise or *a priori* unknown forcing may be a part of the story, but some underlying dynamics generate the time series.

It is, perhaps, the primary goal of dynamical systems theory to write down real world observations in a mathematically rigorous and physically interpretable framework. In many ways, this is the central challenge of all of science! It is then wholly remarkable that such works by the likes of Takens, Whitney, Eckmann, Stark, Yorke, Ruelle and others have not better known by the broader scientific community. Their efforts laid the mathematical foundations and identified the practical limitations of recovering a “full picture” (a complete representation of every basis coordinate) of a dynamical system from partial observations. They showed that, when properly structured, observations embed within them all that is required to describe a system’s dynamics in a mathematically rigorous fashion. It is then appropriate that the rules describing these relations are called *embedding theorems*.

While elegant and robust, these theories lacked something that scientists and engineers value: applicability to real-world systems. This limitation arose because these tools were only able to provide us with embeddings describing the dynamics of full dimensional systems in a distorted way. And while the path back to an undistorted representation was guaranteed by these theorems, how to go about learning that mapping has been an open question.

The solution to this problem emerged in the form of universal functional approximators from statistical learning. With the advent of neural networks, it is now possible to learn these *a priori* unknown mappings between embeddings of dynamics data extracted from time series and those derived from the full-dimensional description of a system. The objective of this dissertation is to contribute to the closing of that loop, by creating a pipeline from time series observables to

full-dimensional reconstruction.

Over the last six years I have been a grateful and humbled beneficiary of the profound and comprehensive works of the scientists who defined how to represent ground truth geometrically. I am fortunate to come to this literature at a time where the modeling tools required to recover these full-dimensional representations of dynamics from embeddings are available and easily deployable. Thanks to these advances, I have been able to make my own modest contribution to the field of dynamics and embeddings that I have come so deeply appreciate.

The central objective of this dissertation is to provide *a* solution to the problem of learning full-dimensional dynamics from time series, addressing a key challenge in the field of embedding theorem. We offer key insights into how arbitrary time series from a dynamical system can be used to reconstruct a complete representation of its dynamics. In Chapter 2, we develop and validate this approach for molecular systems. In Chapter 3, we demonstrate the robustness of this formulation to experimental noise. In Chapter 4, we extend this work to systems beyond molecular dynamics and introduce the TAR software package. In Chapter 5, we explore the optimization of feature selection for modeling of complex systems. Chapters 6 and 7 cover a separate project from my PhD, where I designed probes for the removal of PFAS contaminants from water, using both linear and cyclic probes, respectively.

1.2 Takens' Reconstruction of Single Molecules

Protein engineering and design is a field of growing interest due to the therapeutics, reaction engineering and fundamental science problems relying on the structure and function of these biologically ubiquitous molecules. Proteins are, however, still molecules and understanding their behavior is then a statistical mechanics problem. While the highest spatial resolution experimental techniques available, cryo-electron microscopy and X-ray crystallography are capable of describing the structures of these molecules frozen in time with root mean squared deviations (RMSDs) on the

order of 0.1 nm, these proteins are not static. They fold and unfold over time following dynamical trajectories through some high dimensional phase space spanned by their atomic coordinates and velocities. To sample these dynamics, we have to sacrifice atomistic resolution. State of the art techniques like single molecule fluorescence resonance energy transfer (smFRET) furnish univariate or multivariate time series observations of proteins. Typically, smFRET trajectories measure head to tail distances along the backbone of a protein which is sufficient to describe if a protein is folded or unfolded. It may not, however, resolve different folded or misfolded states or the dynamical trajectories that bring them there. It is then highly desirable to reconstruct atomistic representations of these trajectories to resolve these differences and better understand the dynamics of proteins.

Takens' Delay Embedding Theorem provides a proves that single variate time series observations can be used to construct manifolds that embed within them the full dimensional dynamics of the systems that they describe under deterministic and stochastic forcing Takens [1981], Stark [1999], Stark et al. [2003], Broomhead and King [1986]. Chapters 2-3, comprise content reported in two articles published in The Journal of Chemical Physics Topel and Ferguson [2020] and The Journal of Chemical Theory and Computation Topel et al. [2023]. The second work was performed in collaboration with **Prof. Allison Squires (UChicago)** who provided us with the noise forms characteristic of smFRET time series critical to our ability to produce realistic synthetic smFRET trajectories.

In Chapter 2, we propose and computationally validate a pipeline taking in synthetic univariate smFRET trajectories and reconstructing atomistic representations of every frame in these trajectories with a pipeline entitled Single-molecule TAKens Reconstruction (STAR). This work was performed with molecular dynamics simulations of the molecule $C_{24}H_{50}$ and the mini-protein Chignolin Lindorff-Larsen et al. [2011] from which we extract synthetic smFRET time series measured from the distance between the first and final backbone C_{α} carbons of each molecule. In this pipeline, we first extract manifolds embedding the full dimensional dynamics of molecules from their smFRET

time series via Takens’ Delay Embedding Theorem. Next, we reduce the dimensionality of these manifolds using the diffusion maps technique to extract a phase space spanned by the minimal number of collective variables to describe the dynamics of the molecules we study, $\mathcal{M}'$. This reduction is performed using a training portion of each trajectory with the remainder of the points projected on to the latent space via Nystrom extension. We perform a second dimensionality reduction from the configurational states of the full dimensional trajectory to recover intrinsic manifold $\mathcal{M}$. We then employ a fully connected feed-forward neural network to map from $\mathcal{M}'$ to $\mathcal{M}$ trained over points from the first quarter of each trajectory. A second fully connected feed-forward neural network are used to learn a pairwise distance matrix between each atom in each frame while final reconstructions are performed using classical multidimensional scaling. This work provides computation validation of the STAR technique by demonstrating reconstruction qualities with accuracies on the order 0.5-2 angstroms, in line with state of the art static experimental techniques. Moreover, it provides access to representations of free energy landscapes of molecules from univariate timeseries.

In Chapter 3, we validate the STAR technique under noise characteristic of smFRET experiments. The smFRET technique relies on the emissions of photons from acceptor and donor fluorophores tagged to individual atoms in the molecules we are studying. The measure of intramolecular distance is a function of the relative photon emission rates of these dyes. As photon emission statistics are Poisson distributed, a key source of noise in these systems comes from the discrete emissions of these photons. Using experimental noise forms, we perform STAR on synthetic smFRET time series with synthetic noise corresponding to different numbers of photon emissions per smFRET time bin for the proteins Chignolin and Villin. Additionally, in order to address larger proteins, we modify the STAR framework to map directly from latent space manifolds to full dimensional dynamics with an artificial neural network rather than through classical multidimensional scaling. We demonstrate that a trained STAR model can reconstruct time-resolved molecular trajectory directly from experimental data. Trained models achieve heavy atom RMSD reconstruction accuracies

over a hold-out molecular dynamics test set of 0.1-0.4 nm, commensurate with the ~ 0.1 nm accuracies attainable by cryo-electron microscopy and X-ray crystallography Schüttelkopf and van Aalten [2004], Chang and Rosenthal [2011]. The STAR models trained over molecular simulation trajectories may be applied to novel smFRET data. We demonstrate that we can construct STAR models for two fast-folding mini-proteins, Chignolin ($\tau_{\text{fold}} = 0.6 \mu\text{s}$) and Villin ($\tau_{\text{fold}} = 2.8 \mu\text{s}$) under conditions of trajectory length, time resolution, and signal-to-noise ratio (i.e., dye intensity) bridging computational and experimental viability. While these proteins fold too quickly to be accessible with existing smFRET technology, this work lays foundations for and specifies experimental conditions necessary to perform STAR on real proteins.

1.3 Takens' Reconstruction of Dynamics from Arbitrary Time Series

Takens' Theorem is applicable to arbitrary dynamical systems and consequentially, STAR is not limited to molecular systems. In this section, we extend this work to a variety of contexts from ecology to economics. We introduce the TAKens' Reconstruction (TAR) software package, a multipurpose pipeline capable of in-taking univariate or multivariate time series, reconstructing phase spaces of their dynamics and backmapping to full dimensional dynamics where appropriate. In Chapters 4 and 5, we perform the critical work of i) generalizing STAR to TAR, ii) providing a computational framework that can be easily imported by any scientist with automated TAR parameter generation iii) exploring and optimizing data structuring through modified Takens' delay vector construction iv) providing theoretical explanations for empirically derived optimization results. With these challenges addressed, the main contribution of the works presented in this section is to democratize this class of models and extend profound results obtained in the field of dynamical systems to other fields. This work was performed in the **Econophysix Lab** partially at the École Polytechnique in Paris under the guidance of **Jean-Phillipe Bouchaud** and **Michael Benzaquen** who helped me to understand and model dynamics of economic systems.

In Chapter 4 we introduce the TAR software package and present alternative formulations of the TAR pipeline allowing for the analysis of systems without ground truth full dimensional training trajectories. This work presents a highly interoperable framework that integrates dynamical systems theory with machine learning models to optimize time series analysis in a physically robust manner. We introduce a pipeline that extracts phase space representations of time series data in a low-dimensional format while preserving high-dimensional dynamics through Takens' Theorem and manifold learning. Additionally, we provide a method for mapping these representations back to full-dimensional system descriptions when a training trajectory is available. We show that TAR can be used to reconstruct arbitrary dynamical systems, starting with Lotka Volterra. Furthermore, we demonstrate that dynamical systems theoretic measurements taken directly from timeseries can be used to predicatively select observation streams to optimize TAR models using molecular simulations of $C_{24}H_{50}$. In a Villin protein system, we showed that multi-temporal TAR can access dynamics across timescales and integrate the information to improve upon univariate and multivariate TAR constructions. Finally, we demonstrate how TAR can be used in the field of economics, by reproducing the probability distributions of returns on short timescales for liquid securities.

In Chapter 5 we provide a mathematical treatment of the observable optimization problem. Takens' Theorem is silent on the choice of observable for time series reconstruction provided we have sampled a time series for a sufficiently long period of time. However, we observe empirical differences in reconstruction quality as a function of observable choice. The Kolmogorov-Sinai Entropy (KSE) is a measure of the rate of information production in a chaotic system. Takens' Theorem only guarantees that a diffeomorphism exists between the manifolds embedding the time series and full dimensional dynamics, so reconstructed manifolds may have smoothly deformed topography corresponding to stretching, squashing, bending, local warping, twisting, smooth rotation and shearing. *Intrinsic* geometrical deformations will alter distances on the manifold while *extrinsic* deformations leave distances unchanged. Here we prove that intrinsic deformations of

an extracted manifold corresponding are directly related to the KSE measurement. Moreover, we motivate why minimization of KSEs should maximize reconstruction quality as minimally chaotic trajectories are less sensitive to the accuracy of time series observations.

1.4 Molecular Probe Design For Per- and polyfluoroalkyl Substance Removal

Per- and polyfluoroalkyl (PFAS) are so-called “forever chemicals” that have become ubiquitous in the water supply. While the United States Environmental Protection Agency has reduced to zero with an enforceable limit on the order of parts per trillion, mass detection and removal techniques have not been able to keep up with increased water standards. Consequently, PFAS molecules pose a profound yet unaddressed hazard to humans with risks from low birth weights to various cancers. Sequestering these molecules in the face of a wide range of competing interferents is an important question in the world of public health. The shape and nature of this broad class of molecules poses two key challenges. The first is that as straight chain fluorinated compounds, they look and behave similarly to other similar molecules such as Sodium dodecyl sulfate (SDS). The second is that PFAS compounds encompass on the order of thousands of molecules, meaning that it is crucial that any method of sequestering one type of PFAS be tunable to address other types of the molecule. In Chapters 6 and 7 we explore molecular designs capable of sequestering these molecules in a selective fashion. This work was a collaboration with **Prof. Junhong Chen (UChicago)**.

In Chapter 6, we combine all-atom molecular dynamics simulations, enhanced sampling, deep representational learning, and Bayesian optimization to perform high-throughput virtual screening for highly sensitive and selective molecular probes. Our molecular design space consists of 3850 linear hydrocarbon chains with varying degrees of halogenation with and without amine and phosphine-based head groups. By employing a data-driven search process, we efficiently explore the molecular design space to optimize sensitivity to perfluorooctanesulfonic acid (PFOS) as a

prototypical PFAS analyte and selectivity relative to a sodium dodecyl sulfate (SDS) interferent. We calculate 504 Gibbs free energies of probe-analyte and probe-interferent interactions and identify probes with PFOS association free energies of up to $(-\Delta G_{PFOS}) = (9.8 \pm 0.2)$ kJ/mol and selectivities relative to SDS of $(-\Delta \Delta G_{PFOS-SDS}) = (3.1 \pm 1.5)$ kJ/mol. A $C_{11}Br_{23}P(CH_3)_2$ probe containing 11 backbone brominated carbons and a tertiary phosphine head group possesses the most sensitive binding constant to PFOS within the defined search space of $K_b^{PFOS} = (177.4 \pm 12.7) M^{-1}$, and a semibrominated probe $C_5H_{11}C_7Br_{14}N(CH_3)_2$ containing twelve backbone carbons and a tertiary amine head group possesses the highest selectivity relative to SDS $K_b^{PFOS}/K_b^{SDS} = (4.6 \pm 1.7)$. A retrospective analysis of our data to extract interpretable design rules reveals that the sensitivity of linear hydrogenated probes increases by approximately 1 kJ/mol per C-C bond. Addition or removal of halogen atoms and amine or phosphine head groups produces non-monotonic changes in both sensitivity and selectivity with changes to sensitivity of up to 2.5 kJ/mol. This work places empirical limitations on the performance of a wide range of linear probes for PFOS detection, and offers a generic strategy for the high-throughput computational screening to promote selective and sensitive binding.

In Chapter 7, we provide computational validations of a β -cyclodextrin based PFOS sensor. This ultrasensitive sensing platform for perfluorooctane sulfonic acid (PFOS) detection in tap water with a reporting limit (250 ppq) lower than the maximum contaminant level (4 ppt) set by the U.S. Environmental Protection Agency, using remote gate field-effect transistor (RGFET) sensors featuring β -cyclodextrin (β -CD)-modified reduced graphene oxide as the sensing membrane. We performed enhanced sampling simulations of α , β and γ cyclodextrins with PFOS as well as interferents SDS and TCAA (trichloroacetic acid). The all-atom molecular simulations provide molecular-level insight into the structural and thermodynamic behaviors of the target-probe complex binding. Computed binding constants are in good agreement with experimental measurements, reflecting the same rank order in preferential binding. Our calculations suggest that the superior sensitivity and specificity of β -cyclodextrin for PFOS over TCAA can be understood as a good

size match of the β -cyclodextrin cavity for the PFOS molecule that results in favorable dispersion interactions of the hydrophobic chain within the cavity while enabling the hydrophilic head group solvent exposed. Our results also suggest potential routes for rational engineering of the rim of the β -cyclodextrin to promote favorable electrostatic interactions with the PFOS head and further elevate the binding constant.

1.5 Publications

Chapters 2, 3, 6 and 7 include work from published articles. Chapters 4 and 5 are adapted from working manuscripts whose work I have presented in a variety of venues including The American Physical Society March Meeting and Network Society annual conferences. Completed versions are being prepared for submission. The citations are enumerated according to the corresponding chapters. Note that Chapter 7 only contains the computational portion of the work. Equal contribution co-first authors are indicated by an asterisk.

2. **Maximilian Topel** and Andrew L. Ferguson. Reconstruction of protein structures from single-molecule time series. *The Journal of Chemical Physics*, 153(19):194102, 11 2020. doi:10.1063/5.0024732.
3. **Maximilian Topel**, Ayesha Ejaz, Allison Squires, and Andrew L. Ferguson. Learned reconstruction of protein folding trajectories from noisy single-molecule time series. *Journal of Chemical Theory and Computation*, 19(14):4654–4667, 2023. doi:10.1021/acs.jctc.2c00920.
4. **Maximilian Topel** and Andrew L. Ferguson. Optimizing data structure to learn dynamics from time series data, Manuscript
5. **Maximilian Topel** and Andrew L. Ferguson. Kolmogorov Sinai Entropies for identification of optimal observables for prediction and dynamics reconstruction, Manuscript

6. Siva Dasetty*, **Maximilian Topel***, Yifeng Oliver Tang, Yuqin Wang, Eric Jonas, Seth B. Darling, Junhong Chen, and Andrew L. Ferguson. Data-driven discovery of linear molecular probes with optimal selective affinity for pfas in water. *Journal of Chemical and Engineering Data*, 68(12): 3148–3161, 2023. doi:10.1021/acs.jced.3c00404.
7. Yuqin Wang, Hyun-June Jang, **Maximilian Topel**, Siva Dasetty, Yining Liu, Mohamed Ateia, Aaron Tam, Vepa Rozyyev, Ellie Ouyang, Wen Zhuang, Haihui Pu, Sang Soo Lee, Jeffrey Elam, Andrew Ferguson, Seth Darling, Junhong Chen. “Reversible Ppt-Level Detection of Perfluorooctane Sulfonic Acid in Tap Water Using Field-Effect Transistor Sensors.” *ChemRxiv*, 2024, doi:10.26434/chemrxiv-2024-6q6q5

CHAPTER 2

RECONSTRUCTION OF PROTEIN STRUCTURES FROM SINGLE-MOLECULE TIME SERIES

2.1 Abstract

Single-molecule experimental techniques track the real-time dynamics of molecules by recording a small number of experimental observables. Following these observables provides a coarse-grained, low-dimensional representation of the conformational dynamics but does not furnish an atomistic representation of the instantaneous molecular structure. Takens' Delay Embedding Theorem asserts that, under quite general conditions, these low-dimensional time series can contain sufficient information to reconstruct the full molecular configuration of the system up to an *a priori* unknown transformation. By combining Takens' Theorem with tools from statistical thermodynamics, manifold learning, artificial neural networks, and rigid graph theory, we establish an approach Single-molecule TAKens Reconstruction (STAR) to learn this transformation and reconstruct molecular configurations from time series in experimentally-measurable observables such as intramolecular distances accessible to single molecule Förster resonance energy transfer. We demonstrate the approach in applications to molecular dynamics simulations of a $C_{24}H_{50}$ polymer chain and the artificial mini-protein Chignolin. The trained models reconstruct molecular configurations from synthetic time series data in the head-to-tail molecular distances with atomistic root mean squared deviation accuracies better than 0.2 nm. This work demonstrates that it is possible to accurately reconstruct protein structures from time series in experimentally-measurable observables and establishes the theoretical and algorithmic foundations to do so in applications to real experimental data.

2.2 Introduction

The molecular structure of a protein is defined by a vector in $\mathbb{R}^{3N}$ specifying the Cartesian coordinates of the N constituent atoms. Molecular dynamics (MD) simulation is a computational algorithm that propagates the $3N$ -dimensional position vectors through time under the action of a force field specifying the interatomic interaction potential Frenkel and Smit [2002]. At each step of the MD simulation the full $3N$ -dimensional configurational state of the system is available. Sophisticated experimental techniques such as X-ray crystallography and cryo-electron microscopy can solve protein structures to near atomic resolution in crystalline or vitrified samples Schüttelkopf and van Aalten [2004], Chang and Rosenthal [2011], Roy et al. [2008]. Proteins, however, are typically not functional under these conditions and these structures cannot capture transitions between metastable states. For example, the *Mycobacterium tuberculosis* protein tyrosine phosphatase PtpB exhibits dynamical transitions between “closed” and “open” states in which a pair of α -helices transiently cover the active site and dynamically protects the active site from oxidative inactivation Flynn et al. [2010]. Single-molecule experimental techniques such as single molecule Förster resonance energy transfer (smFRET) can follow protein dynamics by tracking small numbers of experimentally observable intramolecular distances between fluorescent probes conjugated to the target molecule Chang and Rosenthal [2011], Roy et al. [2008], Zerze et al. [2014]. There are currently no experimental techniques available to follow the dynamical evolution in atomistic detail.

Takens’ Delay Embedding Theorem is a result from dynamical systems theory that asserts a sufficiently long and frequently sampled time series of a single system observable can contain sufficient information to reconstruct the full-dimensional state of the system up to an *a priori* unknown diffeomorphism (i.e., smooth and invertible bijective transformation) Takens [1981], Sauer et al. [1991], Packard et al. [1980], Broomhead and King [1986], Cao et al. [1998], Stark [1999], Stark et al. [2003], Vialar [2009], Kantz and Schreiber [2005], Ye et al. [2015]. This theoretical result opens the door to reconstructing the molecular coordinates of a protein from single-molecule

experimental measurements such as smFRET. We have previously applied Takens’ Theorem to synthetic single-molecule time series extracted from molecular dynamics simulations of polymers and proteins to estimate single molecule free energy surfaces (smFES) Wang and Ferguson [2016, 2018a]. Since the simulations also furnish the full molecular configurations we also estimated the “true” smFES from the atomistic molecular simulation trajectory to numerically verify the existence of the *a priori* unknown diffeomorphism and place empirical bounds on the degree of perturbation to the true smFES induced by this transformation. In this work, we build upon these foundations using tools from rigid graph theory and artificial neural networks to learn this transformation from the data and also approximate the inverse transformation from the low-dimensional smFES back up to the molecular configuration space. This calibrates a functional approximation mapping a time series in a single system observable to the atomistic molecular configuration, thereby enabling reconstruction of atomistic molecular structures from experimentally-measurable observables. We term this approach Single-molecule TAKens Reconstruction (STAR).

The structure of this paper is as follows. In the next section we describe the methodological details of STAR. We detail the mathematical formulation and numerical solution of each step of the learning problem combining principles and tools from statistical thermodynamics, manifold learning, and rigid graph theory. In Section 2.4 we present applications of STAR to MD simulations of a $C_{24}H_{50}$ polymer chain and the 10-residue artificial mini-protein Chignolin. We train STAR to reconstruct molecular configurations from univariate time series in the head-to-tail distance as a synthetic and idealized smFRET time trace. In Section 2.5 we present our conclusions and opportunities for future work.

2.3 Methods

2.3.1 Principles of STAR

A cartoon schematic of STAR is presented in Fig. 2.1. The objective of this work is to train STAR as an accurate and generalizable model to predict molecular configurations from time series data in experimental observables through the four-step pathway $b \rightarrow d \rightarrow c \rightarrow e \rightarrow f$. Each panel in the figure corresponds to a different representation of the molecular system and the arrows between them correspond to mathematical operations to convert one representation to another. Red arrows correspond to unsupervised learning problems, typically nonlinear dimensionality reduction, blue arrows to supervised learning problems requiring learning of a nonlinear function linking the inputs and outputs, and grey arrows to deterministic operations. In order to learn and approximate each of these functions we require training data for which the molecular configurations (Fig. 2.1a) corresponding to each point in the time series in the experimental observable (Fig. 2.1b) is known. We obtain the former from atomistic MD simulations and the latter by computing the experimentally-measurable observable corresponding to each frame in the trajectory. In this work, we adopt the head-to-tail ($h2t$) distance as an experimental observable that could, in principle, be measured by smFRET Roy et al. [2008], Zerze et al. [2014]. We emphasize that the MD simulation data is only required to train the model and once the model is trained there is no requirement for any additional molecular calculations. The trained model predicts molecular configurations (Fig. 2.1f) from novel time series data (Fig. 2.1b) that was not present in the training data via the pathway $b \rightarrow d \rightarrow c \rightarrow e \rightarrow f$. In this work, we generate the novel time series data from additional MD simulations, but, in principle, the STAR model trained and calibrated over MD data could be applied to experimental measurements. We detail the additional sophistications that would be necessary to achieve that goal in the final section of the paper. A separate STAR model must be trained for each molecular system.

What is the origin of the multi-step pathway in STAR illustrated in Fig. 2.1? Why not predict

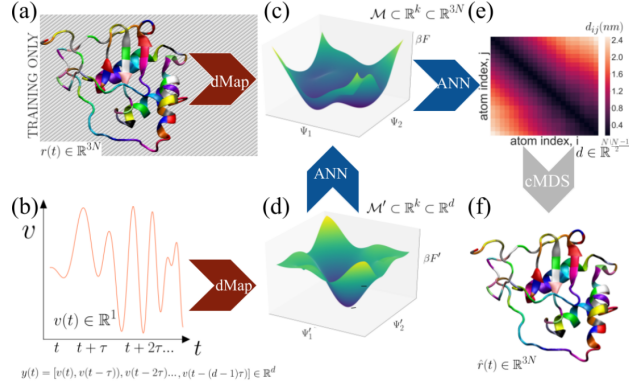


Figure 2.1: Cartoon schematic of Single-molecule TAKens Reconstruction (STAR). The trained STAR model predicts molecular configurations of a protein $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ from univariate time series in an experimentally-measurable observable of the system $v(t) \in \mathbb{R}^1$ through the four-step pathway $b \rightarrow d \rightarrow c \rightarrow e \rightarrow f$. Each panel corresponds to a different representation of the molecular system and arrows between them to mathematical operations. Red arrows represent unsupervised nonlinear manifold learning problems (i.e., nonlinear dimensionality reduction), blue arrows represent supervised learning problems (i.e., function approximation), and grey arrows represent deterministic operations. (a,b) Molecular dynamics simulations sufficiently long to sample the thermally-relevant configurational space provide training trajectories of molecular configurations $\mathbf{r}(t) \in \mathbb{R}^{3N}$ (panel a) and a scalar time series in an experimentally-measurable system observable $v(t) \in \mathbb{R}^1$ (panel b). In this work, we set $v(t)$ to be the head-to-tail ($h2t$) intramolecular distance that is, in principle, measurable by smFRET. (c) Interactions between atoms constrain the molecular trajectory $\mathbf{r}(t) \in \mathbb{R}^{3N}$ to a k -dimensional manifold $\mathcal{M} \subset \mathbb{R}^k \subset \mathbb{R}^{3N}$ with effective dimensionality $k \ll 3N$. The collective variables $\{\psi_1, \psi_2, \dots, \psi_k\}$ parameterizing the manifold are extracted from the simulation trajectory using diffusion maps. The manifold $\mathcal{M}$ supports the smFES $\beta F(\psi_1, \psi_2, \dots, \psi_k)$. (d) An image of $\mathcal{M}$ and the smFES is obtained from the time series data $v(t) \in \mathbb{R}^1$ by forming a d -dimensional delay embedding $\mathbf{y}(t) = [v(t), v(t - \tau), v(t - 2\tau), \dots, v(t - (d - 1)\tau)] \in \mathbb{R}^d$ and then applying diffusion maps to learn a parameterization $\{\psi'_1, \psi'_2, \dots, \psi'_k\}$ of a manifold $\mathcal{M}' \subset \mathbb{R}^k \subset \mathbb{R}^d$. Under technical conditions on v , d , and τ discussed in the main text, Takens' Delay Embedding Theorem asserts that the dynamical evolution of $\mathbf{y}(t)$ is C^1 -equivalent to that of $\mathbf{r}(t)$ and that the manifold $\mathcal{M}'$ is related to $\mathcal{M}$ via diffeomorphic (i.e., smooth, invertible, and bijective) transformation $\Theta : \mathcal{M}' \rightarrow \mathcal{M}$. We learn the transformation Θ from the training data using a simple artificial neural network. (e,f) The manifold $\mathcal{M} \subset \mathbb{R}^k$ contains a low-dimensional projection of $\mathbf{r}(t) \in \mathbb{R}^{3N}$ into $\{\psi_i\}_{i=1}^k$ from which molecular configurations $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ may be approximately reconstructed. We perform the reconstruction by first learning the pairwise distances between all atoms $\mathbf{d} \in \mathbb{R}^{N(N-1)/2}$ using an artificial neural network (panel e) and then using classical multidimensional scaling to deterministically transform this into the reconstructed atomic coordinates $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ (panel f). All molecular renderings in this work were constructed using VMD Humphrey et al. [1996].

molecular configurations directly from the $h2t$ time series in a single step $b \rightarrow f$? In this work, we exploit the generic low effective dimensionality of molecular systems along with theoretical guarantees offered by Takens’ Delay Embedding Theorem to formulate a succession of simpler and typically lower-dimensional learning problems with firm theoretical underpinnings that we solve using appropriate numerical tools. An additional benefit of this perspective is that the approach also explicitly learns the smFES. The trained STAR model therefore furnishes both a prediction of the molecular structure and its location and thermodynamic stability on the smFES. We now detail each step of the STAR approach presented in Fig. 2.1.

Molecular dynamics training data $(\mathbf{r}(t), v(t))$

MD simulations of sufficient duration to comprehensively sample the thermally-relevant configurational space of the molecular system are required to provide the initial training data for STAR. Data that do not sample all relevant conformational states and transitions will result in models overfitted to the training data and a poorly calibrated STAR model that is unable to generalize to new data. For the small macro and biomolecular systems considered in this work, simulations of a few hundred ns to a few μ s were sufficient to provide good sampling and generalizability. The simulation trajectory provides a temporally-ordered sequence of snapshots $\mathbf{r}(t) \in \mathbb{R}^{3N}$ (Fig. 2.1a). For each snapshot we compute the instantaneous value of an experimentally-accessible observable to define a 1D time series $v(t) \in \mathbb{R}^1$ (Fig. 2.1b). In the present work, we set $v(t)$ to be the head-to-tail molecular distance $h2t(t)$ that is, in principle, measurable by smFRET by conjugating the termini of the molecules with fluorescent probes. These time series can therefore be regarded as idealized synthetic smFRET time traces with no measurement noise and arbitrarily high time resolution. These training data are used to train all steps of the STAR pipeline.

Learning the atomistic manifold $\mathcal{M}$

Interactions between the constituent atoms of the molecular system generically constrain molecular trajectories $\mathbf{r}(t) \in \mathbb{R}^{3N}$ within the $3N$ -dimensional Cartesian coordinate space to a so-called intrinsic manifold $\mathcal{M} \in \mathbb{R}^k \subset \mathbb{R}^{3N}$ of effective dimensionality $k \ll 3N$ Ferguson et al. [2010], García [1992], Amadei et al. [1993], Hegger et al. [2007], Zhuravlev et al. [2009], Das et al. [2006], Belkin and Niyogi [2003] (Fig. 2.1c). We learn this manifold from the simulation trajectory training data using the diffusion maps (dMaps) nonlinear manifold learning technique Coifman et al. [2008a], Ferguson et al. [2010], Das et al. [2006], Ferguson et al. [2011], Coifman et al. [2005b], Coifman and Lafon [2006], Nadler et al. [2006b,a], Coifman et al. [2005b], Coifman and Lafon [2006]. Applying dMaps identifies both the dimensionality k of the latent manifold and the CVs $\{\psi_1, \psi_2, \dots, \psi_k\}$ parameterizing it as nonlinear functions of the atomic coordinates Ferguson et al. [2010, 2011], Ferguson [2017], Sidky et al. [2020b]. The dMap CVs are the leading eigenvectors of a discrete random walk constructed over the $3N$ -dimensional snapshots comprising the simulation trajectory. The bandwidth ε of the Gaussian kernel used to construct the random walk can be tuned automatically based on the structure of the data Ferguson et al. [2010], Coifman et al. [2008b] and the dimensionality of the embedding k defined by a gap in the eigenvalue spectrum Ferguson et al. [2010, 2011], Wang and Ferguson [2018b], Coifman and Lafon [2006], Coifman et al. [2005b].

Assuming that the distance metric used to measure pairwise similarities between molecular configurations is a good proxy for the kinetic proximity of the configurations over short time scales, these CVs furnish a dynamically meaningful low-dimensional embedding into the slowest relaxing modes of a diffusion process over the data. In particular, Euclidean distances in the embedding correspond to diffusion distances in configurational space that measure the kinetic proximity of any pair of configurations under the action of the random walk Coifman et al. [2005b], Nadler et al. [2006a]. This dynamic interpretability of dMap CVs make them an excellent choice for parameterization of the intrinsic manifold $\mathcal{M}$, but other nonlinear dimensionality reduction or manifold learning techniques such as Isomap Tenenbaum et al. [2000], Weinberger and Saul [2006], Li et al. [2006],

Wang [2011], locally linear embeddings (LLE) Roweis and Saul [2000b], Zhang and Wang [2006], local tangent space alignment Wang [2011] may also be employed.

The only input to the diffusion map algorithm is the definition of a distance metric $k(\mathbf{r}_i, \mathbf{r}_j)$ with which to measure the dissimilarity of pairs of configurations $(\mathbf{r}_i, \mathbf{r}_j)$ in the trajectory. In order to guarantee the existence of the diffeomorphism, it is critical that this metric be appropriately symmetrized to eliminate any *spatial symmetries* that cannot be distinguished by the choice of experimentally-measurable observable $v(t)$. Our choice of observable $v(t) = h2t(t)$ is invariant under translation, rotation, mirror inversion, and – for chemically symmetric molecules (e.g., simple polymer chains) – head-tail inversion of the molecular configuration $\mathbf{r} \in \mathbb{R}^{3N}$. Any representation of the system based on $v(t)$ sacrifices the ability to distinguish changes in the system state under any of these transformations. It is possible to recover some or all of these symmetries under different choices for $v(t)$ or by multiplexed simultaneous measurements $\mathbf{v}(t) = \{v_1(t), v_2(t), \dots\}$. As is standard practice in measuring molecular similarity, we select $k(\mathbf{r}_i, \mathbf{r}_j)$ to be the rotationally and translationally aligned root mean squared deviation (RMSD), which naturally mods out the rototranslational symmetry using the Kabsh algorithm Kabsch [1976], and we eliminate the discrete symmetries by minimizing under mirror inversion,

$$k(\mathbf{r}_i, \mathbf{r}_j) = \min_{\text{mirror inversion}} \text{RMSD}(\mathbf{r}_i, \mathbf{r}_j). \quad (2.1)$$

or both mirror and head-tail inversion for chemically symmetric molecules Wang and Ferguson [2018a],

$$k(\mathbf{r}_i, \mathbf{r}_j) = \min_{\text{head-tail inversion}} \min_{\text{mirror inversion}} \text{RMSD}(\mathbf{r}_i, \mathbf{r}_j). \quad (2.2)$$

Failure to eliminate all such spatial symmetries induced by the choice of observable $v(t)$ violates Takens’ Theorem and can cause the manifolds $\mathcal{M}$ and $\mathcal{M}'$ to no longer be diffeomorphisms related by a well defined and learnable transformation.

The dMap CVs $\{\psi_i\}_{i=1}^k$ define a data-driven parameterization of the k -dimensional manifold $\mathcal{M}$. Projection of the full simulation trajectory into these CVs defines an empirical probability distribution $P(\psi_1, \psi_2, \dots, \psi_k)$. An estimate of the smFES mapping out the free energy F over $\mathcal{M}$ is obtained via the statistical mechanical relationship $\beta F(\psi_1, \psi_2, \dots, \psi_k) = -\ln P(\psi_1, \psi_2, \dots, \psi_k) + C$, where C is an arbitrary additive constant and the free energy is de-dimensionalized by the reciprocal temperature $\beta = 1/k_B T$. New molecular configurations $\mathbf{r}_{\text{new}}$ not contained within the data used to construct $\mathcal{M}$ may be projected onto the manifold using the Nyström extension Sunday et al. [2009], Laing et al. [2007], Long and Ferguson [2019b].

Learning the Takens' manifold $\mathcal{M}'$

The previous section detailed a procedure to estimate the intrinsic manifold $\mathcal{M}$ and the smFES it supports by applying manifold learning to a simulation trajectory of the atomistic molecular configuration $\mathbf{r}(t) \in \mathbb{R}^{3N}$. Takens' Delay Embedding Theorem Takens [1981], Sauer et al. [1991], Packard et al. [1980], Broomhead and King [1986], Cao et al. [1998], Stark et al. [2003], Vialar [2009], Kantz and Schreiber [2005], Ye et al. [2015] provides a means to reconstruct an image of the intrinsic manifold $\mathcal{M}'$ (Fig. 2.1d) by applying similar operations to delay embeddings of a time series in a single coarse-grained observable $v(t) \in \mathbb{R}^1$ (Fig. 2.1b). Takens' Theorem asserts that (i) a $(d \geq (2k + 1))$ -dimensional delay embedding $\mathbf{y}(t) = [v(t), v(t - \tau), v(t - 2\tau), \dots, v(t - (d - 1)\tau)] \in \mathbb{R}^d$ in a generic observable $v(t)$ constructed at a delay time τ uniquely specifies the instantaneous state of the system, (ii) the dynamical evolution of $\mathbf{y}(t)$ is C^1 -equivalent (i.e., identical under a smooth and continuous mapping) to that of $\mathbf{r}(t)$, and (iii) the evolution of $\mathbf{y}(t)$ lies on a manifold $\mathcal{M}'$ that is related to $\mathcal{M}$ by a diffeomorphism (i.e., smooth, invertible, and bijective transformation) $\Theta : \mathcal{M}' \rightarrow \mathcal{M}$ with inverse $\Theta^{-1} : \mathcal{M} \rightarrow \mathcal{M}'$ Wang and Ferguson [2018a], Takens [1981], Sauer et al. [1991], Packard et al. [1980], Broomhead and King [1986], Cao et al. [1998], Stark et al. [2003], Vialar [2009], Kantz and Schreiber [2005], Zerze et al. [2014]. The diffeomorphism is *a priori* unknown but is guaranteed only to stretch and squash the manifold and not tear it or stitch

it together in new ways. As such, $\mathcal{M}'$ is a topologically identical image of $\mathcal{M}$ that preserves its continuity and connectivity and possesses the same dimensionality Takens [1981], Sauer et al. [1991], Packard et al. [1980], Broomhead and King [1986], Vialar [2009], Kantz and Schreiber [2005].

Takens' Theorem holds for any generic observable of the system v that does not contain any spurious symmetries that are not present in the system itself (i.e., the system is invariant in the observable under particular symmetries), for any delay time τ that does not introduce temporal aliasing (i.e., is a multiple of a period of the dynamical motion), and for any delay embedding dimensionality d greater than twice the intrinsic dimension of the system k Takens [1981], Sauer et al. [1991], Kantz and Schreiber [2005], Letellier et al. [1998], Cross and Gilmore [2010], Letellier and Gouesbet [1996], Letellier et al. [1998], Wang and Ferguson [2016, 2018a]. In practice, better results are obtained for observables v that are strong functions of all system degrees of freedom and which respond sensitively to the important dynamical motions of the system Wang and Ferguson [2018a], for delay times τ estimated as the first minimum of the autocorrelation or mutual information of $v(t)$ Fraser and Swinney [1986a], and for delay dimensionalities d estimated using the E_1 method of Cao Cao [1997], Villani and Zaldivar Comenges [2000], Kennel et al. [1992]. Takens' Theorem still holds when applied to observables v that do contain symmetries not present in the system, but the system may only be reconstructed up to those symmetric operations (*vide supra*), and also to observables of subsystems and under stochastic or deterministic forcing Stark [1999], Stark et al. [2003]. The latter two generalizations are relevant to the present work because we adopt the head-to-tail molecular distance as our observable $v(t) = h_2 t(t)$, which is both an observation of the solute subsystem within the full solute-solvent system and is subject to dynamical coupling with solvent and deterministic or stochastic forcing by any attached thermostats, barostats, or other external constraints. Finally, Takens' Theorem may also be applied to multiplexed measurements of several simultaneous observables $\mathbf{v}(t) = \{v_1(t), v_2(t), \dots\}$ such as multi-channel smFRET measuring multiple intramolecular distances Roy et al. [2008], Cao et al. [1998].

The manifold $\mathcal{M}' \subset \mathbb{R}^k \subset \mathbb{R}^d$ is estimated by applying dMaps to the delay embedding $\mathbf{y}(t) = [v(t), v(t-\tau), v(t-2\tau), \dots, v(t-(d-1)\tau)]$ parameterized by the k CVs $\{\psi'_1, \psi'_2, \dots, \psi'_k\}$. We apply dMaps under a distance metric $k'(\mathbf{y}_i, \mathbf{y}_j)$ measuring the dissimilarity of pairs of delay vectors that we select to be a simple Euclidean distance metric. This metric requires modification due to the introduction of a spurious symmetry into the system by the temporal ordering of $v(t)$ within the $\mathbf{y}$ vectors induced by the Takens' delay embedding. This spurious symmetry can be understood by considering a hypothetical microstate transition of the system $\mathbf{r}_A(t-\tau) \rightarrow \mathbf{r}_B(t) \rightarrow \mathbf{r}_C(t+\tau)$ and its reverse $\mathbf{r}_C(t-\tau) \rightarrow \mathbf{r}_B(t) \rightarrow \mathbf{r}_A(t+\tau)$, with associated $d=3$ -dimensional delay vectors $\mathbf{y}_{\text{fwd}} = [v_A, v_B, v_C]$ and $\mathbf{y}_{\text{bkwd}} = [v_C, v_B, v_A]$. Adopting a convention that defines a mapping between system microstates and delay vectors based on the central microstate, both of these delay vectors are associated with microstate $\mathbf{r}_B$ but the delay vectors themselves are not identical. By observing the past and future of $\mathbf{r}_B$ we can distinguish whether it was occupied as part of a forward or reverse transition, and the “forward” $\mathbf{r}_B$ and “backward” $\mathbf{r}_B$ appear in the Takens' delay embedding as distinct identifiable states. For equilibrium systems obeying detailed balance, the forward and backward transitions between any pair of microstates have equal fluxes and we should not be able to identify whether occupancy of a particular microstate $\mathbf{r}_B$ resulted from the forward or reverse transition in any such pair. To prevent the occurrence of this symmetry breaking in $\mathcal{M}'$ that remains unbroken in $\mathcal{M}$ we must symmetrize $k'(\mathbf{y}_i, \mathbf{y}_j)$ to eliminate this *temporal symmetry*. This assures that Takens' Theorem is not violated and the manifolds $\mathcal{M}$ and $\mathcal{M}'$ remain diffeomorphic. We have previously demonstrated the importance of eliminating this symmetry in applications of Takens' Theorem to equilibrium molecular systems Wang and Ferguson [2016, 2018a]. As such, we minimize the distance metric under time reversal, which is equivalent to minimizing under inversion of one of the delay vectors,

$$k'(\mathbf{y}_i, \mathbf{y}_j) = \min [\|\mathbf{y}_i - \mathbf{y}_j\|_2, \|\mathbf{y}_i - \text{flip}(\mathbf{y}_j)\|_2] \quad (2.3)$$

The dMap CVs $\{\psi'_i\}_{i=1}^k$ define a data-driven parameterization of the k -dimensional manifold $\mathcal{M}'$ over which we construct the smFES $\beta F'(\psi'_1, \psi'_2, \dots, \psi'_k) = -\ln P'(\psi'_1, \psi'_2, \dots, \psi'_k) + C'$ by compiling the empirical probability distribution $P'(\psi'_1, \psi'_2, \dots, \psi'_k)$ of the delay vectors projected onto the manifold. New delay vectors $\mathbf{y}_{\text{new}}$ not contained within the data used in the construction of $\mathcal{M}'$ may be projected onto the manifold using the Nyström extension Sunday et al. [2009], Laing et al. [2007], Long and Ferguson [2019b]. We exploit this out-of-sample extension when applying the trained STAR model to new data that was not present during training.

Learning the diffeomorphism from $\mathcal{M}'$ to $\mathcal{M}$

Takens' Theorem guarantees the smooth manifolds $\mathcal{M}$ and $\mathcal{M}'$ are related by a diffeomorphism $\Theta : \mathcal{M}' \rightarrow \mathcal{M}$ Wang and Ferguson [2018a], Takens [1981], Sauer et al. [1991], Packard et al. [1980], Broomhead and King [1986], Cao et al. [1998], Stark et al. [2003], Vialar [2009], Kantz and Schreiber [2005], Zerze et al. [2014]. The theoretical guarantees on the existence of this mapping and its low dimensional nature are a valuable advantage of the multi-step STAR pathway that would be lost by formulating a direct reconstruction of the molecular configurations from the univariate time series. We adopt a convention associating each delay vector $\mathbf{y}(t) = [v(t), v(t - \tau), v(t - 2\tau), \dots, v(t - (d - 1)\tau)]$ with the configurational microstate corresponding its central element $\mathbf{r}(t - ((d - 1)/2)\tau)$. We assert that d be odd in order to make this association unambiguous. This mapping means that the configurations in the leading $t < (d - 1)\tau/2$ and trailing $t > (d - 1)\tau/2$ periods of the molecular simulation trajectory are not associated with any delay vector and are eliminated from all analyses.

Having defined the associations between the projections of the delay vectors $\mathbf{y}(t)$ on $\mathcal{M}'$ represented as $\{\psi'_i(t)\}_{i=1}^k$ and the projections of the molecular configurations $\mathbf{r}(t)$ on $\mathcal{M}$ represented as $\{\psi_i(t)\}_{i=1}^k$, we define a supervised learning problem between pairs of data points $\left(\{\psi'_i(t)\}_{i=1}^k, \{\psi_i(t)\}_{i=1}^k\right)$ to perform data-driven estimation of the k -dimensional to k -dimensional diffeomorphism $\Theta : \mathcal{M}' \rightarrow \mathcal{M}$ (Fig. 2.1d $\rightarrow$ c). There are many ways to learn and approximate

this function, including k-nearest neighbors Cover and Hart [1967], kernel methods Schölkopf and Smola [2002], or local Jacobians Wang and Ferguson [2016, 2018a], Hauser and Ray [2017]. In this work we employ simple fully-connected feedforward artificial neural networks (ANN) as an easy to train and flexible function approximator Hassoun et al. [1996]. We typically find that networks comprising 4-8 hidden layers each containing $\sim 10 \times k$ neurons are adequate to furnish high accuracy mappings.

Learning the reconstruction $\hat{\mathbf{r}}(t)$

The final step is to learn approximate reconstructions of the molecular configurations $\mathbf{r}(t) \in \mathbb{R}^{3N}$ from their projections $\{\psi_i(t)\}_{i=1}^k \in \mathbb{R}^k$ onto the intrinsic manifold $\mathcal{M}$. If the effective dimensionality of the simulation trajectory is less than or equal to k and the dMap CVs have been properly learned from the simulation trajectory, then the k -dimensional subspace spanned by $\{\psi_i(t)\}_{i=1}^k$ is expected to preserve the important configurational variance in $\mathbf{r}(t)$ Ferguson et al. [2010, 2011]. As such, the location on the intrinsic manifold $\mathcal{M}$ should contain sufficient information to approximately reconstruct the configurational state of the system $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ up to any symmetries that have been modded out in its construction (*vide supra*). In this sense, the process $\mathbf{r}(t) \rightarrow \{\psi_i(t)\}_{i=1}^k \rightarrow \hat{\mathbf{r}}(t)$ may be viewed as the concatenation of a low-dimensional encoder furnished by diffusion maps and a decoder to be learned from the data. Again we have the one-to-one mapping of data pairs $\left(\{\psi_i(t)\}_{i=1}^k, \mathbf{r}(t)\right)$ and can formulate and solve this as a supervised learning problem.

We split this learning problem into two steps and instead of learning to predict the Cartesian coordinates of the atoms $\mathbf{r}(t)$ directly from $\{\psi_i(t)\}_{i=1}^k$, we first learn the ordered vector of pairwise distances between all atoms $\mathbf{d}(t) \in \mathbb{R}^{N(N-1)/2}$. Rigid graph theory asserts that specification of all $\binom{N}{2}$ pairwise distances defines the absolute coordinates of the N points up to translation, rotation, and mirror inversion Dokmanic et al. [2015], Singer [2008]. In practice, calculation of the coordinates from the pairwise distances is easily accomplished using classical multidimensional

scaling (cMDS) to form the Gram matrix and compute its eigendecomposition Dokmanic et al. [2015], Crippen [1978]. In the present application, the translational, rotational, and mirror inversions (and head-tail inversion, where applicable) have all been modded out of the construction of $\mathcal{M}$ and so there is no (additional) information loss in formulating the supervised learning problem as first approximating the functional mapping $\{\psi_i(t)\}_{i=1}^k \rightarrow \mathbf{d}(t)$ (Fig. 2.1c $\rightarrow$ e) and then making the deterministic transformation $\mathbf{d}(t) \rightarrow \hat{\mathbf{r}}$ using cMDS (Fig. 2.1e $\rightarrow$ f). Unlike the one-step formulation of the reconstruction $\{\psi_i(t)\}_{i=1}^k \rightarrow \hat{\mathbf{r}}$, the two-step formulation eliminates the need to perform any alignment of the molecular configurations with respect to rotation, translation, and mirror inversion since these symmetries are naturally modded out within the pairwise distance matrix. The omission of these alignment operations, either mutually or to some fixed reference structure, carries advantages in avoiding the introduction of noise and approximations into the fitting problem.

We use simple ANNs to learn $\{\psi_i(t)\}_{i=1}^k \rightarrow \mathbf{d}(t)$ from the training data. Typically we find that networks comprising 4-8 hidden layers each containing on the order of up to $\sim 1000 \times k$ neurons are adequate to furnish high accuracy mappings. The molecular configurations $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ are reconstructed only up to translational, rotational, and mirror inversion symmetries (and head-tail inversion, for chemically symmetric molecules), and so comparisons of the reconstruction accuracy between $(\hat{\mathbf{r}}, \mathbf{r})$ pairs must be performed by mutual alignment under these transformations. The alignment problem corresponds to an orthogonal Procrustes problem Schönemann [1966] that we efficiently solve using the Kabsch algorithm Kabsch [1976].

Deploying the trained STAR model

The STAR pipeline trained for a particular molecular system may then be used to predict molecular reconstructions $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ from new time series data $v(t) \in \mathbb{R}^1$ through the four-step pathway $b \rightarrow d \rightarrow c \rightarrow e \rightarrow f$ (Fig. 2.1b-f): Takens’ delay vectors $\mathbf{y}$ are constructed from the time series $v(t)$, projected onto the Takens’ manifold $\mathcal{M}'$ using the Nyström extension Sunday et al. [2009], Laing et al. [2007], Long and Ferguson [2019b], mapped onto the atomistic manifold $\mathcal{M}$ using the trained

ANN, converted into the atomistic pairwise distances matrix $\mathbf{d}$ using another trained ANN, and then finally transformed into the reconstructed molecular configurations $\hat{\mathbf{r}}$ using cMDS. Provided the training data were sufficiently rich to span the thermally-relevant metastable states and transitions in the system and the supervised learning problems are not overfitted, then the trained STAR model should be able to generalize to new time series data from the system that were not included in model training. Importantly, deployment of the trained model does not require any additional molecular simulation data during the deployment phase (Fig. 2.1a).

We validate the trained model by extracting time traces of the molecular heat-to-tail distance $v(t) = h2t(t)$ from new MD simulation trajectories that were not present in the training data and testing the capability of the trained STAR model to reconstruct the molecular configurations $\mathbf{r}(t)$ in the trajectory. A well-trained model should be able to predict molecular configurations from novel testing data with similar accuracies to that in the training data. In applications to real experimental smFRET data, the true molecular configurations would not be available to make this comparison and the model predictions would have to be validated by indirect means. In all cases we seek only to reconstruct the molecular configuration of the solute and do not attempt to predict the coordinates of the water solvent or any counter ions. In principle, this information is – again subject to the relevant symmetries – present within the Takens’ delay embedding, but is much more challenging to recover due to the permutational fungibility of the water molecules and the relatively weaker influence of solvent coordinates on our choice of solute-centric observable.

2.3.2 *Molecular Dynamics Simulations*



MD simulations of $\text{C}_{24}\text{H}_{50}$ were conducted using the Gromacs 4.6 simulation suite van der Spoel et al. [2005] using the TraPPE potential Martin and Siepmann [1998] that models each CH_3 and CH_2 group as a united atom and the all-atom SPC model of water Berendsen et al. [1981]. Chain

topologies were constructed using the PRODRG2 server Schüttelkopf and van Aalten [2004]. Lennard Jones interactions were smoothly set to zero at a cutoff of 1.4 nm and Lorentz-Berthelot combining rules used to determine dispersion interactions between unlike atoms Allen and Tildesley [1989]. Electrostatic interactions were treated using particle mesh Ewald Essmann et al. [1995] with a real-space cutoff of 1.4 nm and a 0.12 nm reciprocal-space grid spacing that were optimized during runtime. Simulations were conducted in a $5 \times 5 \times 5 \text{ nm}^3$ cubic box with periodic boundary conditions that was sufficiently large to prevent self-interactions of the chain through the periodic walls even in its fully-extended configuration. Initial system configurations were generated by placing an initially elongated chain into an empty box and solvating with 4117 water molecules to a density of 1.0 g/cm^3 . High energy overlaps were eliminated by steepest descent energy minimization to a force threshold of 2000 kJ/mol.nm. Initial particle velocities were sampled from a Maxwell-Boltzmann distribution at 298 K. Simulations were performed in the NPT ensemble at 298 K and 1 bar using a Nosé-Hoover thermostat Nosé [1984] and an isotropic Parrinello-Rahman barostat Parrinello and Rahman [1981]. Equations of motion were integrated using the leap-frog algorithm Hockney and Eastwood [2010] with a 2 fs time step. LINCS constraints were used to fix bond lengths to their equilibrium distances for computational efficiency and as required by the TraPPE and SPC potentials Nadler et al. [2006a]. Systems were equilibrated for 1 ns before conducting a 100 ns production run during which system configurations were saved every 0.2 ps. Head-to-tail distances were computed for each frame of the 500,000 frame simulation trajectory to furnish the univariate time series $v(t) = h2t(t)$ in addition to the atomistic simulation trajectory $\mathbf{r}(t)$. The first 20 ns (100,000 frames) of these trajectories provided the $(\mathbf{r}(t), v(t) = h2t(t))$ training data used to train the STAR model, and the remaining 80 ns (400,000 frames) reserved for testing. Input files for these simulations are provided in the Supplementary Material.

Chignolin

MD simulations of the 10-residue (166 atom) engineered mini-protein Chignolin (GYDPETGTWG, PDB ID: 1UAO) Honda et al. [2004] were performed by D.E. Shaw Research using the Desmond simulation suite Bowers et al. [2006] on the Anton supercomputer Shaw et al. [2010] and reported in Ref. Lindorff-Larsen et al. [2011]. The peptide was modeled using the the CHARMM22 $\times$ force field Piana et al. [2011] and solvated by ~ 1900 water molecules Jorgensen et al. [1983b] in a $4 \times 4 \times 4$ nm³ cubic box with periodic boundary conditions. Two Na⁺ ions were added to maintain charge neutrality. Lennard-Jones interactions were treated with a 0.95 nm cutoff and electrostatic interactions treated by Gaussian Split Ewald Shan et al. [2005] employing a 0.95 nm real-space cutoff and a $32 \times 32 \times 32$ cubic grid. Equations of motion were integrated with a 2.5 fs time step. Equilibration runs were conducted in the NPT ensemble at 340 K and maintained by a Nosé-Hoover thermostat Nosé [1984], Hoover [1985]. The 106 μ s production run was conducted in the NVT ensemble employing a Nosé-Hoover thermostat Nosé [1984], Hoover [1985] and system configurations harvested every 200 ps. The first 5 μ s (25,000 frames) of these trajectories were used to train the STAR model, and the next 15 μ s (75,000 frames) employed for testing.

2.4 Results and Discussion

We demonstrate and validate STAR in applications to molecular dynamics simulations of a C₂₄H₅₀ polyethylene chain and the β -hairpin engineered mini-protein Chignolin. Simulation data constituting the training and testing data are collected as described in Section 2.3.2. The STAR pipeline was trained and deployed as described in Section 2.3.1 with details specific to each system provided below. In each case we achieve RMSD reconstruction accuracies on the hold-out test data of better than 0.2 nm.

2.4.1 $C_{24}H_{50}$

STAR Training

The training portion of the $C_{24}H_{50}$ simulation trajectory comprised 100,000 frames saved at 0.2 ps intervals recording the Cartesian coordinates $\mathbf{r}(t) \in \mathbb{R}^{72}$ of the $N = 24$ united atoms of the polymer chain (cf. Fig. 2.1a). In order to run dMaps into local memory, the simulation trajectory was subsampled with a stride of 20 and the excluded frames projected into the manifold *post hoc* using the the Nyström extension Sunday et al. [2009], Laing et al. [2007], Long and Ferguson [2019b]. The stride length was empirically tuned to balance computational efficiency and comprehensive sampling of the configurational phase space to achieve robust embeddings and reconstruction accuracies. Application of spatially-symmetrized dMaps to these data employing a kernel bandwidth of $\varepsilon = \exp(-3)$ nm exposed a $k=2$ -dimensional intrinsic manifold $\mathcal{M} \subset \mathbb{R}^2$ spanned by nonlinear collective variables of the united atom coordinates $\{\psi_1, \psi_2\}$ and supporting the smFES $\beta F(\psi_1, \psi_2)$ (cf. Fig. 2.1c). The eigenvalue spectrum exhibiting a spectral gap separating the second non-trivial eigenvalue from the continuum of higher-order eigenvalues is presented in Fig. S1a in the Supplementary Material.

A scalar time series in the head-to-tail distance between the two terminal united atoms $v(t) = h2t(t) \in \mathbb{R}^1$ was computed from the MD training trajectory as a synthetic and idealized smFRET time series over the 100,000 frames separated by intervals of 0.2 ps (cf. Fig. 2.1b). Takens' delay vectors $\mathbf{y}(t) = [h2t(t), h2t(t - \tau), h2t(t - 2\tau), \dots, h2t(t - (d - 1)\tau)]$ were constructed at a delay time of $\tau = 30$ ps (150 time steps) computed as the first minimum in the mutual information of $h2t(t)$ Fraser and Swinney [1986a] and delay dimensionality $d = 5$ computed using the E_1 method of Cao Cao [1997], Villani and Zaldivar Comenges [2000], Kennel et al. [1992]. This resulted in the construction of 99,400 delay vectors, each associated with a frame in the MD trajectory containing the molecular structure from which the central element in the delay vector was computed. The initial and terminal $(d - 1)\tau/2 = 60$ ps (300 frames) of the MD trajectory that were unassociated

with any delay vector were dropped from further analyses. Application of temporally-symmetrized dMaps to the delay embedding trajectory, subsampled with a stride of 20 in order to fit into local memory and employing a kernel bandwidth of $\varepsilon = 1$ nm, defined a $k = 2$ -dimensional manifold $\mathcal{M}' \subset \mathbb{R}^2$ spanned by $\{\psi'_1, \psi'_2\}$. The eigenvalue spectrum exhibiting a spectral gap after the second non-trivial eigenvalue is presented in Fig. S1b. The frames strided over in the application of dMaps were projected into $\mathcal{M}'$ using the Nyström extension and used to estimate the smFES $\beta F'(\psi'_1, \psi'_2)$ (cf. Fig. 2.1d).

The diffeomorphism $\Theta : \mathcal{M}' \subset \mathbb{R}^2 \rightarrow \mathcal{M} \subset \mathbb{R}^2$ linking the two manifolds was learned using a simple 2-25-25-25-25-2 fully-connected feedforward ANN comprising four hidden layers of 25 neurons (cf. Fig. 2.1d $\rightarrow$ c). The ANN employed tanh activations and was trained using Adam Kingma and Ba [2014] with a batch size of 500 and a learning rate of 1×10^{-4} . Training was terminated after 250 epochs upon plateauing of the validation error.

The function linking the projection of each frame in the MD trajectory into the intrinsic manifold $\{\psi_1(t), \psi_2(t)\}$ to the $N(N-1)/2 = 276$ -dimensional united atom pairwise distance vectors $\mathbf{d}(t) \in \mathbb{R}^{276}$ corresponding to that configuration was learned using a 2-4-187-370-552-276 fully-connected feedforward ANN employing tanh activations and trained using Adam Kingma and Ba [2014] with a batch size of 500 and a learning rate of 1×10^{-3} over 150 epochs (cf. Fig. 2.1c $\rightarrow$ e). Molecular reconstructions of the united atom coordinates $\hat{\mathbf{r}}(t) \in \mathbb{R}^{72}$ were computed deterministically from each pairwise distance vector using cMDS (cf. Fig. 2.1e $\rightarrow$ f). The reconstruction accuracy of the trained STAR pipeline (cf. Fig. 2.1b $\rightarrow$ d $\rightarrow$ c $\rightarrow$ e $\rightarrow$ f) is assessed by computing the RMSD of the predicted $\hat{\mathbf{r}}(t) \in \mathbb{R}^{72}$ and true $\mathbf{r}(t) \in \mathbb{R}^{72}$ configurations under translational, rotational, mirror, and head-tail alignment by the Kabsch algorithm Kabsch [1976].

STAR Deployment

We illustrate the application of the trained $\text{C}_{24}\text{H}_{50}$ STAR model in Fig. 2.2. The trained STAR pipeline enables us to associate each value of $h2t(t)$ in the synthetic smFRET time trace with a molecular reconstruction and also a location and stability on the smFES supported by the low-dimensional intrinsic manifold Ferguson et al. [2010], Wang and Ferguson [2016]. The mean RMSD reconstruction accuracy of the trained pipeline applied to the 20 ns training trajectory is $\text{RMSD}_{\text{train}} = 8.4 \times 10^{-2}$ nm. In an application to the remaining 80 ns testing trajectory that was not part of the training ensemble, the mean reconstruction accuracy degrades only slightly to $\text{RMSD}_{\text{test}} = 8.6 \times 10^{-2}$ nm and is at worst 0.19 nm, indicating that the STAR model is well trained and has good generalizability to novel data. Histograms illustrating the complete RMSD distribution over the testing trajectory are presented in Fig. S2. We illustrate for five selected points A-E in the $h2t(t)$ time trace (Fig. 2.2a) their projection onto the smFES $\beta F(\psi_1, \psi_2)$ (Fig. 2.2b), RMSD reconstruction accuracy as a function of position on the intrinsic manifold (Fig. 2.2c), and reconstructed $\hat{\mathbf{r}}(t) \in \mathbb{R}^{72}$ and true $\mathbf{r}(t) \in \mathbb{R}^{72}$ molecular structures (Fig. 2.2d). Movie S1 in the Supplementary Material presents an animation of molecular reconstructions and smFES projections for all test data in the time series in Fig. 2.2a.

Configuration A corresponds to an elongated chain configurations with the preponderance of the backbone dihedrals in the *trans* state. Configuration C corresponds to a similarly elongated conformation but with a small number of gauche defects that lead to a small degree of curvature in the contour of the chain. Configuration C resides within $0.3 k_B T$ of the global free energy minimum at $\beta F = 0$, while Configuration A lies slightly higher at $\beta F = 2.7$. Configurations B and D correspond to partially collapsed twisted structures $\sim 3 k_B T$ less stable than the global free energy minimum that lie along the transition pathway between the global minimum at $(\psi_1 \sim 1.2, \psi_2 \sim 2.0)$ containing elongated chains and the weak local minimum at $(\psi_1 \sim 3.2, \psi_2 \sim -6.0)$ containing hydrophobically collapsed coiled chains. Configuration E is a metastable hydrophobically collapsed coil that lies slightly outside the local minimum at $\beta F = 4.8$. We note that Configurations B

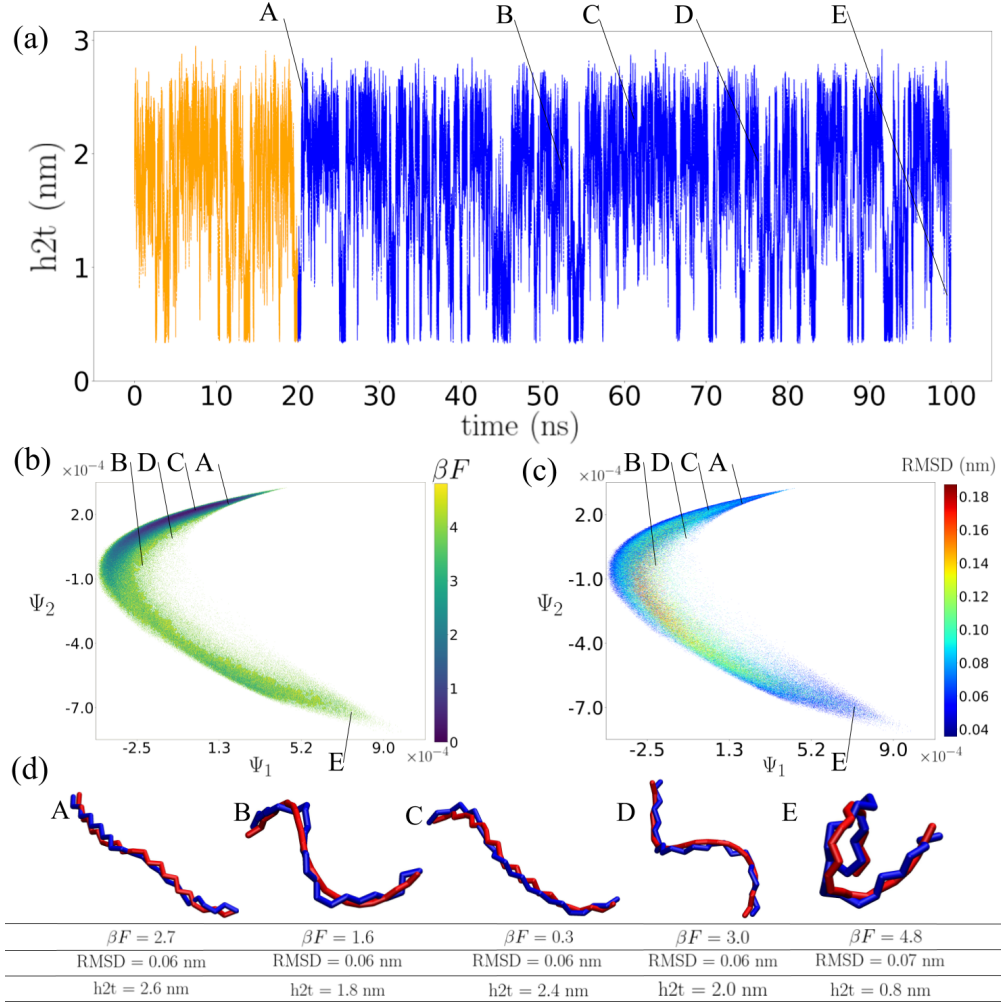


Figure 2.2: Application of STAR to $C_{24}H_{50}$ polymer chain. (a) Synthetic idealized smFRET time trace of the head-to-tail distance $h2t(t)$ between the terminal united atoms computed over a 100 ns MD trajectory with frames saved every 0.2 ps. The first 20 ns are used for training (orange) and the remaining 80 ns for testing (blue). The molecule undergoes hundreds of folding and unfolding events over the course of the simulation. (b) The intrinsic manifold $\mathcal{M}$ spanned by the dMap CVs $\{\psi_1(t), \psi_2(t)\}$ and supporting the smFES $\beta F(\psi_1, \psi_2)$. The Gibbs free energy F is dedimensionalized by the inverse temperature $\beta = 1/k_B T$. The arbitrary zero of F is specified to lie at the global free energy minimum. (c) RMSD reconstruction accuracy as a function of location on the intrinsic manifold $\mathcal{M}$. (d) Molecular reconstructions using the trained STAR pipeline of five representative points A-E in the $h2t(t)$ time series spanning the test set. Reconstructions $\hat{\mathbf{r}}$ (red) are superposed on the corresponding true configurations $\mathbf{r}$ (blue) extracted directly from the MD simulation. The head-to-tail distance of the true configuration and the RMSD under translational, rotational, mirror, and head-tail alignment between the true and reconstructed configurations are reported under each image. The STAR prediction of the location of each point on the smFES is illustrated in panel b and the corresponding dimensionless free energy βF reported under each image.

($h2t = 1.8\text{nm}$) and D ($h2t = 2.0\text{ nm}$) possess similar values of the $h2t$ distance but correspond to very distinct molecular configurations residing at different locations on the smFES that are both accurately reconstructed by STAR. This illustrates the value of using Takens’ Theorem to reconstruct molecular configurations that cannot be distinguished from the instantaneous value of the observable alone.

2.4.2 Chignolin

STAR Training

The training portion of the Chignolin trajectory comprised 25,000 frames saved at 200 ps intervals recording the Cartesian coordinates $\mathbf{r}(t) \in \mathbb{R}^{279}$ of the $N = 93$ heavy atoms of the protein. A $k = 2$ -dimensional intrinsic manifold $\mathcal{M} \subset \mathbb{R}^2$ spanned by $\{\psi_1, \psi_2\}$ was constructed by applying dMaps with a kernel bandwidth of $\varepsilon = \exp(-3)$ nm to a subsampling of this trajectory with a stride of 2. The eigenvalue spectrum exhibiting a spectral gap separating the second non-trivial eigenvalue from the continuum of higher-order eigenvalues is presented in Fig. S3a. Again the stride length was tuned to balance computational cost and robust training of the STAR pipeline. The excluded frames were projected into the manifold using the the Nyström extension. Takens’ delay embeddings were constructed from the synthetic smFRET time series recording the distance between the terminal heavy atoms at a delay time $\tau = 200$ ps (1 time step) Fraser and Swinney [1986a] and delay dimensionality $d = 11$ Cao [1997], Villani and Zaldivar Comenges [2000], Cao [1997], Kennel et al. [1992]. This results in the construction of 24,990 delay vectors. Applying temporally-symmetrized dMaps to the delay embedding trajectory, sub-sampled with a stride of 2 in order to fit into local memory, $\mathbf{y}(t) \in \mathbb{R}^{11}$ with a kernel bandwidth of $\varepsilon = 1$ nm defined a $k = 2$ -dimensional manifold $\mathcal{M}' \subset \mathbb{R}^2$. The eigenvalue spectrum exhibiting a spectral gap after the second non-trivial eigenvalue is presented in Fig. S3b. Again, the excluded frames were projected into the manifold using the the Nyström extension. The diffeomorphism Θ mapping $\mathcal{M}'$ to $\mathcal{M}$ was approximated by a 2-25-25-25-25-2 ANN trained using Adam Kingma and Ba [2014] with a

batch size of 500 and learning rate of 1×10^{-4} over 250 epochs. The function mapping locations on $\mathcal{M}$ to the $N(N-1)/2 = 4278$ -dimensional heavy atom pairwise distances vectors $\mathbf{d}(t) \in \mathbb{R}^{4278}$ was learned and approximated by a 2-4-2855-5706-8556-4278 ANN using Adam Kingma and Ba [2014] with a batch size of 500 and learning rate of 1×10^{-5} over 150 epochs. Predictions of heavy atom molecular configurations were computed deterministically from the pairwise distance vectors using cMDS.

STAR Deployment

Application of the trained Chignolin STAR model is illustrated in Fig. 2.3. The mean heavy atom reconstruction accuracy over the $5 \mu\text{s}$ training trajectory is $\text{RMSD}_{\text{train}} = 0.12 \text{ nm}$. The mean accuracy is only slightly diminished to $\text{RMSD}_{\text{test}} = 0.14 \text{ nm}$ over the $15 \mu\text{s}$ test trajectory reaching a worst case maximum of 0.29 nm , again illustrating good generalizability of the trained model. Histograms illustrating the complete RMSD distribution over the testing trajectory are presented in Fig. S4. We illustrate for five selected points A-E in the $h2t(t)$ time trace (Fig. 2.3a) their projection onto the smFES $\beta F(\psi_1, \psi_2)$ (Fig. 2.3b), RMSD reconstruction accuracy as a function of position on the intrinsic manifold (Fig. 2.3c), and reconstructed $\hat{\mathbf{r}}(t) \in \mathbb{R}^{279}$ and true $\mathbf{r}(t) \in \mathbb{R}^{279}$ molecular structures (Fig. 2.3d,e). Movie S2 presents an animation of molecular reconstructions and smFES projections for all test data in the time series in Fig. 2.3a.

Configurations A and D correspond to the native hairpin state of the protein in which all native hydrogen bonds are intact and which lie within the deep global free energy minimum. The reconstruction accuracy of the densely sampled native fold is extremely good as indicated by RMSDs of $\sim 0.05 \text{ nm}$. Configurations B, C, and E represent a sampling of the unfolded ensemble containing a diversity of random coiled states with some or all of the native hydrogen bonds broken. Despite the relatively sparser sampling and larger configurational diversity of the unfolded ensemble, the RMSD reconstruction accuracy is still better than 0.2 nm . This indicates that the training data, despite containing only four folding/unfolding transitions, provides a sufficiently

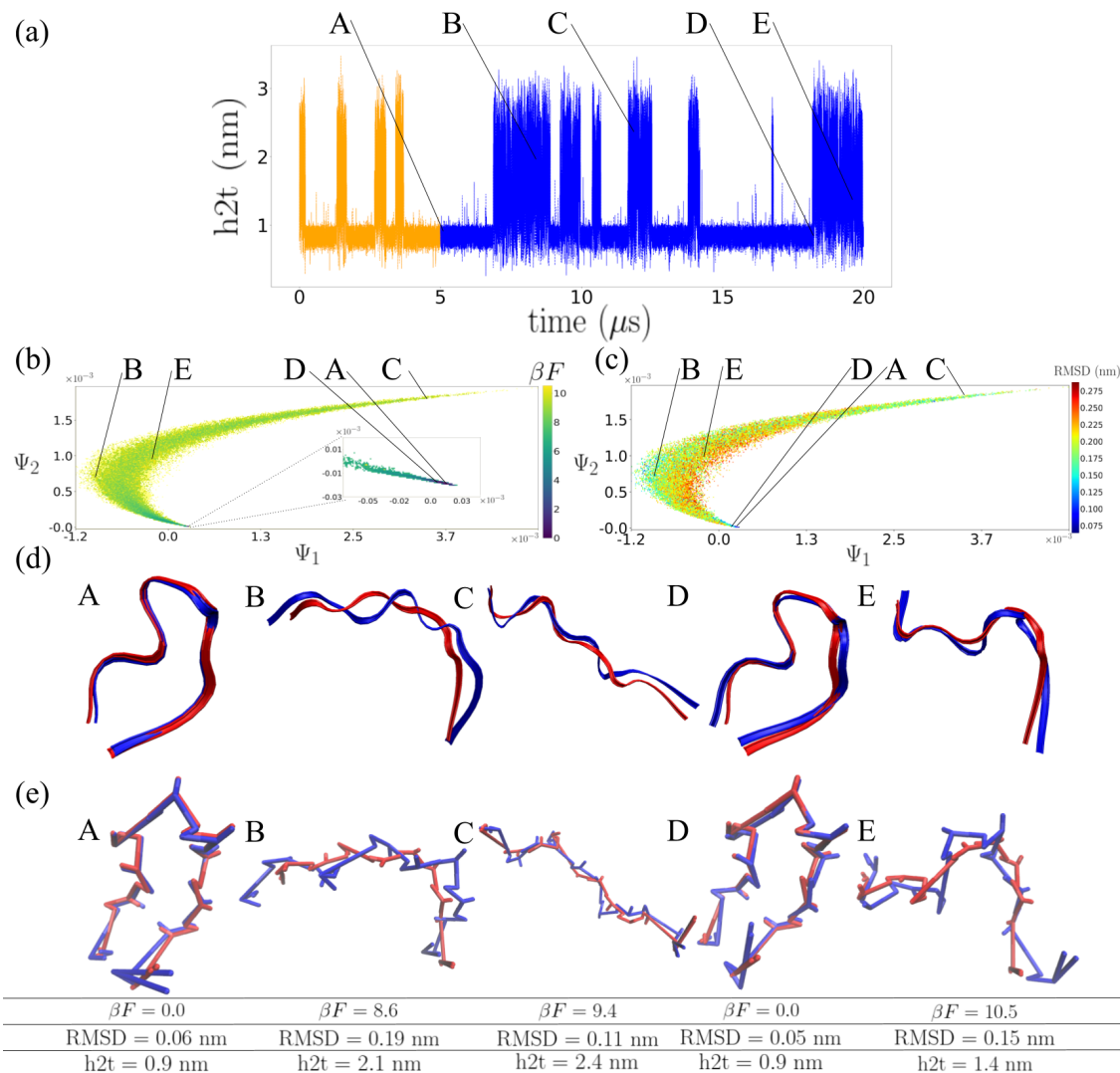


Figure 2.3: Application of STAR to the 10-residue engineered mini-protein Chignolin. (a) Synthetic idealized smFRET time trace of the head-to-tail distance $h2t(t)$ between the terminal heavy atoms computed over a 20 μ s MD trajectory with frames saved every 200 ps. The first 5 μ s are used for training (orange) and the remaining 15 μ s for testing (blue). The molecule undergoes dozens of folding and unfolding events over the course of the trajectory. (b) The smFES $\beta F(\psi_1, \psi_2)$ supported by the intrinsic manifold $\mathcal{M}$ has its arbitrary zero of F specified to lie at the global free energy minimum. (c) RMSD reconstruction accuracy as a function of location on the intrinsic manifold $\mathcal{M}$. Molecular reconstructions using the trained STAR pipeline of five representative points A-E selected from the testing $h2t(t)$ time series rendered in (d) ribbon and (e) licorice representations. Reconstructions $\hat{\mathbf{r}}$ (red) are superposed on the corresponding true configurations $\mathbf{r}$ (blue) extracted directly from the MD simulation. The head-to-tail distance of the true configuration and the RMSD under translational, rotational, and mirror alignment between the true and reconstructed configurations are reported under each pair of images. The STAR prediction of the location of each point on the smFES is illustrated in panel b and the corresponding dimensionless free energy βF reported under each image.

dense and representative sampling of configurational space to enable accurate reconstruction of even transiently visited molecular conformations. It is unsurprising that configurations lying at high free energies (e.g., Configuration E, $\beta F = 10.5$, RMSD = 0.15 nm) that are very sparsely sampled in the training data have poorer reconstruction accuracies than the densely sampled native configurations (e.g., Configuration A, $\beta F = 0.0$, RMSD = 0.06 nm). We propose that adaptive sampling techniques to perform targeted sampling of infrequently-visited regions of the manifold may be beneficial in providing more training data to the STAR pipeline and improving the reconstruction accuracy of higher free energy configurations. Configurations B ($h2t = 2.1$ nm) and C ($h2t = 2.4$ nm) possess similar values of the $h2t$ distance but constitute two configurationally and thermodynamically distinct members of the Chignolin unfolded ensemble. Again, we observe that STAR accurately reconstructs configurations with similar instantaneous values of the scalar observable but which correspond to structurally different configurations that lie in very different regions of the smFES.

2.5 Conclusions

This work presents the theoretical underpinnings and numerical implementation of an approach Single-molecule TAKens Reconstruction (STAR) to reconstruct molecular configurations from time series in a single experimentally-measurable observable. The basis for the approach rests upon the integration of Takens’ Delay Embedding Theorem with tools from manifold learning, statistical thermodynamics, artificial neural networks, and rigid graph theory to extract a representation of the system state from the scalar time series and learn the *a priori* unknown mapping to the molecular configuration from molecular dynamics simulation training data. The trained STAR model can then be applied to novel time series data to predict both the corresponding molecular configurations and their location and stability on the single molecule free energy surface. We have demonstrated and validated the approach in applications to a $C_{24}H_{50}$ polyethylene chain and the 10-residue engineered β -hairpin mini-protein Chignolin. In both cases we demonstrate that trained STAR models can robustly reconstruct the molecular configurations from time series data in the head-to-tail

distance with mean RMSD reconstruction accuracies better than 0.2 nm. This is comparable to the approximately 0.4 nm resolution of cryo-electron microscopy Righetto et al. [2019], Cheng [2015] or 0.1 nm resolution of X-ray crystallography Blakeley et al. [2015].

In this work, we adopt the head-to-tail distance as an experimental observable that can, in principle, be measured by a single molecule experimental technique such as smFRET. The head-to-tail time traces in this work are extracted from MD simulation trajectories in order to test our approach in applications where the ground truth molecular configurations are explicitly available from the simulations. These time traces can therefore be considered to represent synthetic and idealized smFRET data at arbitrarily high temporal resolution and subject to no measurement error or noise. The present work reports a computational proof-of-principle demonstration of STAR in this idealized limit. Applying STAR to real experimental time series must engage a number of concerns surrounding the experimental realities of smFRET measurements including millisecond limits in sampling frequency; shot noise, uncertainties, and unpredictable trajectory lengths due to photobleaching; degraded measurement reliabilities for fluorophores outside of the 2-8 nm range; and conformational perturbations induced by conjugation of the fluorescent probes Chang and Rosenthal [2011], Roy et al. [2008]. Furthermore, the STAR model must be trained and calibrated on MD training data and so the accuracy of the molecular potential functions and degree of sampling of the thermally-relevant configurational space will limit the quality to the trained model. A STAR model calibrated on MD simulation data employing a poor force field or which does not sample all of the experimentally-accessible states and transitions will not perform well when deployed on real experimental time series. In future work, we propose to engage these practical issues empirically by adding noise to the observables extracted from our simulation trajectories, limiting the accessible time resolution, limiting the length of the time series, and exploring the transferability of the trained models between molecular force fields.

We would also like to explore technical innovations to determine optimal experimental observables

(e.g., optimal FRET fluorophore placement) for high-accuracy molecular reconstruction Mittal and Shukla [2018], the extension of STAR to multiplexed measurements Cao et al. [1998], and the potential to reconstruct not just the molecular configuration but also the location and orientation of proximate solvent molecules using permutationally invariant representations of the solvent coordinates Pietrucci [2020], Han et al. [2018]. It would also be of interest to explore the transferability of reconstructions learned under one set of conditions (e.g., temperature, pressure, salt concentration, mutations from wild type) to reconstruct those at another. Takens' Theorem does require time-ordered data, but this may come from many short trajectories rather than a single long one. As such, we would also like to develop adaptive sampling capabilities wherein the deployed model can identify regions of configurational space where it does not have sufficient training data to make accurate predictions and performs automated targeting of additional on-the-fly molecular simulations to supplement training in under-sampled regions. This is anticipated to be particularly important for applications of STAR to large proteins where it is challenging to comprehensively sample the important configurational space. MD simulations are limited by the short time step required for stability of the numerical integration to millisecond time scales on even the fastest computational hardware. This adaptive parameterization protocol would enable efficient training of robust STAR models using a discontinuous ensemble of short MD trajectories on the microsecond time scale, and then deployment of the trained model to reconstruct atomistic trajectories from smFRET data on the time scale of seconds or longer. We would additionally like to also experiment with replacing the ANN and cMDS approach to molecular reconstruction with a conditional Wasserstein generative adversarial network (cWGAN) Arjovsky et al. [2017], Gulrajani et al. [2017] as a more powerful reconstruction tool that, due to its intrinsic driving by noise, has the potential to reintroduce the richness and diversity of molecular structures in the degrees of freedom that are not explicitly captured within the low-dimensional intrinsic manifold. Finally, we also see potential applications of STAR in other applications where it is of interest to reconstruct the state of a dynamical system where it is challenging or impossible to obtain complete information on its state.

As such, we foresee potential applications of the approach in diverse fields and areas of dynamical systems modeling such as fluid mechanics, meteorology, climatology, epidemiology, and ecology.

Supplementary Material

See supplementary material in Topel and Ferguson [2020] for simulation input files for the $C_{24}H_{50}$ molecular dynamics simulations, Jupyter notebooks implementing each stage of the STAR pipeline, Figure S1 showing the $C_{24}H_{50}$ diffusion map eigenvalue spectra, Figure S2 showing the distribution of $C_{24}H_{50}$ RMSD reconstruction accuracies, Figure S3 showing the Chignolin diffusion map eigenvalue spectra, Figure S4 showing the distribution of Chignolin RMSD reconstruction accuracies, Movie S1 showing an animation of molecular reconstructions and smFES projections for $C_{24}H_{50}$, Movie S2 showing an animation of molecular reconstructions and smFES projections for Chignolin.

Author Contributions

A.L.F. conceived and designed the research and provided funding acquisition; A.L.F. and M.T. performed the calculations, analyzed data, and wrote and edited the paper.

Acknowledgments

We thank Dr. Gül H. Zerze for fruitful discussions. This material is based upon work supported by the National Science Foundation under Grant No. DMS-1841810. We are grateful to D.E. Shaw Research for sharing the Chignolin simulation trajectories. A.L.F. is a consultant of Evozyne and a co-author of US Provisional Patents 62/853,919 and 62/900,420 and International Patent Applications PCT/US2020/035206 and PCT/US20/50466.

Data Availability Statement

Input files for the $C_{24}H_{50}$ molecular dynamics simulations and Jupyter notebooks implementing each stage of the STAR pipeline are provided in the supplementary material. The Chignolin simulation trajectories reported in Ref. Lindorff-Larsen et al. [2011] were obtained by request from D.E. Shaw Research.

CHAPTER 3

LEARNED RECONSTRUCTION OF PROTEIN FOLDING TRAJECTORIES FROM NOISY SINGLE-MOLECULE TIME SERIES

3.1 Abstract

Single molecule Förster Resonance Energy Transfer (smFRET) is an experimental methodology to track the real-time dynamics of molecules using fluorescent probes to follow one or more intramolecular distances. These distances provide a low-dimensional representation of the full atomistic dynamics. Under mild technical conditions, Takens’ Delay Embedding Theorem guarantees that the full three-dimensional atomistic dynamics of a system are diffeomorphic (i.e., related by a smooth and invertible transformation) to a time-delayed embedding of one or more scalar observables. Appealing to these theoretical guarantees, we employ manifold learning, artificial neural networks, and statistical mechanics to learn from molecular simulation training data the *a priori* unknown transformation between the atomic coordinates and delay-embedded intramolecular distances accessible to smFRET. This learned transformation may then be used to reconstruct atomistic coordinates from smFRET time series data. We term this approach Single-molecule TAKens Reconstruction (STAR). We have previously applied STAR to reconstruct molecular configurations of a $C_{24}H_{50}$ polymer chain and the mini-protein Chignolin with accuracies better than 0.2 nm from simulated smFRET data under noise free and high time resolution conditions. In the present work, we investigate the role of signal-to-noise ratio, data volume, and time resolution in simulated smFRET data in order to assess the performance of STAR under conditions more representative of experimental realities. We show that STAR can reconstruct the Chignolin and Villin mini-proteins to accuracies of 0.12 nm and 0.42 nm, respectively, and place bounds on these conditions for accurate reconstructions. These results demonstrate that it is possible to reconstruct dynamical trajectories of protein folding from time series in noisy, time binned experimentally-measurable observables, and

lay the foundations for the application of STAR to real experimental data.

3.2 Introduction

Imaging single molecule dynamics is crucial for understanding protein folding and misfolding Schüttelkopf and van Aalten [2004], Chang and Rosenthal [2011], Roy et al. [2008], Zerze et al. [2014], Knowles et al. [2014]. The behavior of a single protein can be characterized by recording the positions of each of its N constituent atoms over time in a time series of R^{3N} dimensional vectors called a molecular trajectory. Computationally, molecular dynamics (MD) simulations can generate all-atom trajectories of molecules by solving Newton’s equations of motion under an interatomic potential defined by an appropriate force field Frenkel and Smit [2002]. MD simulations are subject to numerical and finite precision errors Wong-ekkabut and Karttunen [2016] and can become prohibitively expensive for the simulation of large or slow folding proteins due to the short time steps required to propagate the simulations accurately. Experimentally, cryo-electron microscopy and X-ray crystallography can provide static reconstructions of protein structures at root mean square deviation (RMSD) accuracies of the position of each atom of ~ 0.1 nm Schüttelkopf and van Aalten [2004], Chang and Rosenthal [2011], Roy et al. [2008], Zerze et al. [2014]. The structure of proteins within these crystalline or vitrified states may not correspond to the native functional structure and the static reconstructions cannot shed light on the nature of the dynamical fluctuations and transitions between metastable states Childers and Daggett [2017]. Study of protein dynamics can be critical for understanding their function or dysfunction, with more than 50 known disorders related to misfolding of functional peptides and proteins Knowles et al. [2014]. Single-molecule Fluorescence Resonance Energy Transfer (smFRET) is a popular technique to experimentally follow the dynamics of biomolecular motions by optically recording energy transfer between fluorescent dyes grafted to particular sites on the molecule Roy et al. [2008], Lankiewicz et al. [1997]. This technique permits an observer to track protein dynamics in a coarse-grained sense by following one or more simple geometrical parameters such as the intramolecular distance between the two dyes.

No experimental techniques are currently available to track the single molecule dynamical evolution of a protein with full atomistic resolution.

Takens' Delay Embedding Theorem is a result from dynamical systems theory which asserts that time series data recording a single scalar observable of a dynamical system can, under some weak technical conditions, contain sufficient information to reconstruct the state of the full-dimensional system up to a diffeomorphic (i.e., smooth and invertible) transformation Takens [1981], Sauer et al. [1991], Packard et al. [1980], Broomhead and King [1986], Cao et al. [1998], Stark [1999], Stark et al. [2003], Vialar [2009], Kantz and Schreiber [2005], Ye et al. [2015]. In the context of the present application, the theorem asserts that smFRET measurements can contain sufficient information to reconstruct the full three-dimensional atomic dynamics of the protein via an *a priori* unknown transformation.

In previous work, we showed that it is possible to learn this transformation from MD simulation training data and then use this learned model to reconstruct molecular trajectories from intramolecular distances accessible to smFRET Topel and Ferguson [2020]. We refer to this approach as Single molecule TAKens' Reconstruction (STAR). Our application of STAR to computer simulations of a $C_{24}H_{50}$ polymer chain and the Chignolin mini-protein demonstrated reconstruction accuracies from simulated smFRET data of better than 0.2 nm RMSD accuracy. Although this work served as proof of principle of the technique, it was only validated for long idealized synthetic smFRET time series that were sampled at extremely high time resolution and free of sampling noise.

The motivation of the present work is to test the ability of STAR to accurately reconstruct protein dynamics under experimentally realistic constraints on the smFRET time series associated with its temporal resolution, trajectory length, and presence of sampling noise. The temporal resolution is limited by the need to accumulate sufficient numbers of photons over a specified time bin in order to extract a reliable reading of the intramolecular distance between the fluorescent probes Roy et al. [2008]. The length of a trajectory is primarily limited by photobleaching or photoblinking of the

fluorophores Salem et al. [2019], Demchenko [2020], Ha and Tinnefeld [2012]. Photobleaching is an irreversible chemical process whereby changes in a fluorophore’s electronic structure render it permanently non-emissive Gensch et al. [2005]. Photoblinking is intermittent emissivity of a fluorophore that may result from temporary changes in electronic structure such as electric charge or isomerization, or from being trapped in a triplet state Gensch et al. [2005], Demchenko [2020]. The Poisson statistics regulating photon emission combined with the ~ 1 ns lifetime of the (emitting) excited singlet state Gensch et al. [2005] define this limitation. Sampling noise within the time series arises due to systematic effects and shot-noise (i.e., Poisson statistics) Lankiewicz et al. [1997], Roy et al. [2008]. Theoretical guarantees on the applicability of Takens’ theorem in the presence of limited data, low sampling resolution, and measurement noise are not available. Empirical testing of STAR under physical constraints on these three key variables will assess the degree to which STAR can accurately reconstruct molecular configurations from experimentally realistic smFRET trajectories. In this work we lay the foundations for the application of STAR to real data by applying it to synthetic smFRET data generated from computational MD trajectories for which the ground truth atomic coordinates are exactly known and for which we can precisely control the length, resolution, and noise of the smFRET trajectories.

The primary outcome of this work is to show that STAR is capable of accurate recovery of molecular configurations to accuracies of 0.12 nm and 0.42 nm for, respectively, the Chignolin and Villin fast-folding mini-proteins under conditions that bridge computationally tractable simulations to experimentally realistic FRET conditions. These accuracies are achieved using simulated smFRET trajectories of an aggregated length of approximately $0.7\text{-}3.3\times$ the characteristic protein folding time, temporal resolutions of $1/120\text{-}1/280\times$ the folding time, and sampling noise commensurate with collection of $\sim 10^5$ photons per time bin. For Chignolin ($\tau_{\text{fold}} = 0.6 \mu\text{s}$) this corresponds to MD trajectories of approximately $2 \mu\text{s}$ and time bins shorter than 5 ns, and for Villin ($\tau_{\text{fold}} = 2.8 \mu\text{s}$) to trajectories of $2 \mu\text{s}$ and time bins shorter than 10 ns Lindorff-Larsen et al. [2011]. Extrapolation of these constraints on trajectory length, time resolution, and signal-to-noise ratio suggest that STAR

may currently be deployed upon proteins with characteristic folding times exceeding approximately 100-1000 μ s using state-of-the-art photon-by-photon single-molecule instruments with dye pairs like Cy3/Cy5 that have temporal resolutions of approximately 1-10 μ s Pirchi et al. [2016], Phelps et al. [2017], Oikawa et al. [2018].

The structure of this paper is as follows. In Sec. II, we summarize the previously reported STAR pipeline and discuss implementation of noise, data restriction and time resolution limitations on STAR input data. In Sec. III, we present applications of STAR to MD simulations of the 10-residue artificial mini-protein Chignolin and 35-residue protein Villin under a variety of trajectory lengths, time resolutions, and signal-to-noise ratios. In Sec. IV, we discuss the impact of this work and scope for future development and applications of STAR.

3.3 Methods

3.3.1 Principles of STAR

A schematic illustration of STAR technique is presented in Fig. 3.1. In this subsection we provide a brief overview of the approach that we previously reported in Ref. Topel and Ferguson [2020]. Details of the mathematical underpinnings and numerical implementations including algorithms, training protocols, and (hyper)parameters are provided in the Supporting Information. Template Jupyter notebooks implementing the STAR pipeline are available at <https://github.com/Ferguson-Lab/Limits-of-single-molecule-Takens-reconstruction-notebooks>.

Takens' Delay Embedding Theorem is a proven result from dynamical systems theory asserting that, under some mild technical conditions, there exists a diffeomorphism (i.e., a smooth and invertible transformation) between the the full dimensional state of the system and a so-called delay embedding of one or more coarse-grained observables of the system Takens [1981], Sauer et al. [1991], Packard et al. [1980], Broomhead and King [1986], Cao et al. [1998], Stark [1999],

Stark et al. [2003], Vialar [2009], Kantz and Schreiber [2005], Ye et al. [2015]. In the context of protein folding, this theorem implies that smFRET time series may contain sufficient information to reconstruct the all-atom configuration of the molecule. It is the fundamental principle of STAR to learn this *a priori* unknown transformation from MD simulation training data, and then apply the learned transformation to “upgrade” smFRET time series into trajectories of a molecule tracking its Cartesian coordinates. In the illustration of the STAR pipeline in Fig. 3.1, each panel corresponds to a different representation of a molecular trajectory and the arrows between the panels correspond to learning tasks to convert one representation to another.

The STAR algorithm employs a combination of manifold and nonlinear learning tools to convert a smFRET time series $v(t) \in \mathbb{R}^1$ (Fig. 3.1b) through a three-step pathway ($b \rightarrow d \rightarrow c \rightarrow e$) to a reconstruction of the atomistic molecular trajectory within the Cartesian coordinate space of N atoms $\hat{\mathbf{r}} \in \mathbb{R}^{3N}$ (Fig. 3.1e). Depending on the application and data quality, one may choose to reconstruct all atoms in the molecule or just a subset, for example, the heavy or backbone atoms. In principle, Takens’ Theorem asserts that one could learn this transformation in a single step (i.e., $b \rightarrow e$). In practice, we make use of the generically low effective dimensionality of molecular systems Ferguson et al. [2010, 2011] to recover a $k \ll 3N$ dimensional manifold containing the smFRET trajectory $\mathcal{M}'$ (Fig. 3.1d) and learn the transformation to the analogous k -dimensional manifold containing the all-atom trajectory $\mathcal{M}$. This is beneficial in defining a lower-dimensional and better posed mapping that must be learned from the smFRET to atomistic data. It also furnishes an informative and interpretable k -dimensional free energy landscape supported on the manifold $\mathcal{M}$ that provides a wealth of information on the metastable states and transition pathways of the molecular system.

The existence of the transformation ($d \rightarrow c$) is guaranteed by Takens’ Theorem but its expression is initially unknown and it must be learned from training data. We perform all-atom molecular dynamics simulations to furnish both an all-atom molecular trajectory $r \in \mathbb{R}^{3N}$ (Fig. 3.1a) and a

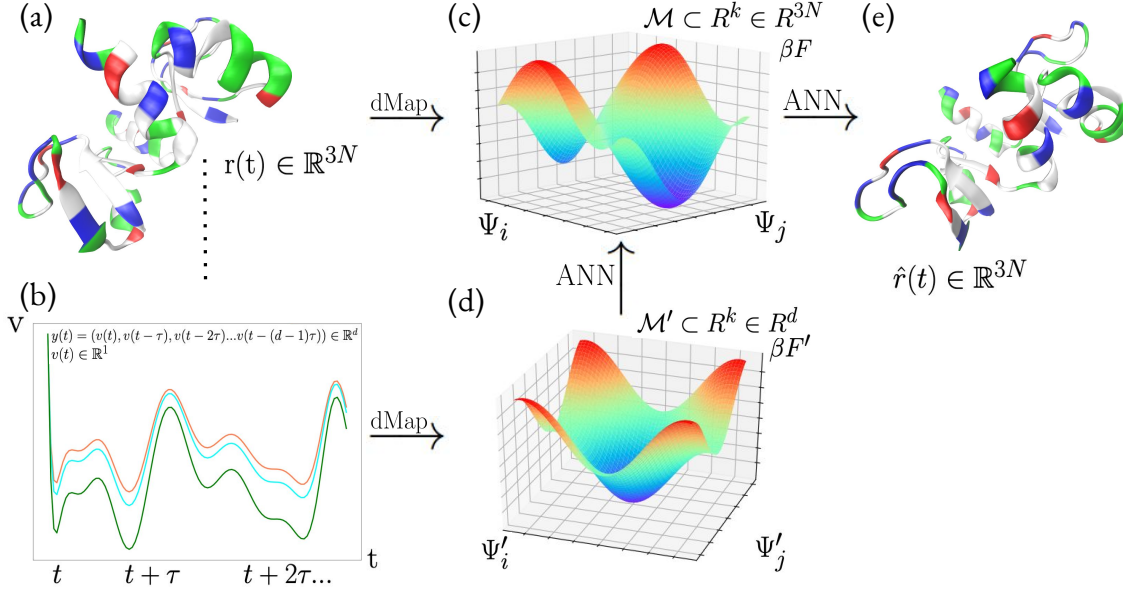


Figure 3.1: Cartoon of the layout of Single Molecule TAKens Reconstruction (STAR). STAR takes time series in one or more coarse-grained observables of the molecular system such as an intramolecular distance furnished by smFRET $v(t) \in \mathbb{R}^1$ and reconstructs a molecular trajectory within the Cartesian coordinate space of N atoms $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ through the three-step pathway $b \rightarrow d \rightarrow c \rightarrow e$. Each panel corresponds to a different representation of a molecular trajectory and the arrows between the panels correspond to learning tasks to convert one representation to another. (a,b) For the purposes of training the STAR pipeline we collect all-atom MD trajectories recording the Cartesian coordinates of the N constituent atoms $\mathbf{r}(t) \in \mathbb{R}^{3N}$ and from these construct synthetic smFRET time series in a single intramolecular distance $v(t) \in \mathbb{R}^1$. Once the STAR pipeline is trained, we no longer need any additional MD simulation data and the pipeline can operate on new synthetic (or experimental) smFRET data collected under similar conditions. Following the prescription of Takens' Theorem we construct a delay embedding of the scalar time series as $\mathbf{y}(t) = [v(t), v(t-\tau), v(t-2\tau) \dots v(t-(d-1)\tau)] \in \mathbb{R}^d$, where τ is the delay time and d is the delay dimensionality. (c) The k -dimensional manifold containing the all-atom simulation trajectory $\mathcal{M} \subset \mathbb{R}^k \in \mathbb{R}^{3N}$ and spanned by the nonlinear collective variables $\{\Psi'_1, \Psi'_2, \dots, \Psi'_k\}$. Collective couplings between the molecular degrees of freedom leads to an emergent low-dimensionality wherein $k \ll 3N$. We learn $\mathcal{M}$ from the MD simulation trajectory $\mathbf{r}(t)$ using diffusion maps unsupervised nonlinear manifold learning. (d) The k -dimensional manifold containing the smFRET delay embedding $\mathcal{M}' \subset \mathbb{R}^k \in \mathbb{R}^d$ spanned by the nonlinear collective variables $\{\Psi'_1, \Psi'_2, \dots, \Psi'_k\}$. We learn $\mathcal{M}'$ from the delay embedding $\mathbf{y}(t)$ using diffusion maps. Takens' Theorem asserts $\mathcal{M}'$ is an image of $\mathcal{M}$ and, under some technical conditions on symmetries and periodicities, they are related by a diffeomorphism (i.e., a smooth and invertible transformation) $\Theta: \mathcal{M}' \rightarrow \mathcal{M}$ that we learn by training a simple fully-connected feedforward artificial neural network as a flexible and expressive nonlinear function approximator. (e) Reconstruction of the atomistic molecular trajectory within the Cartesian coordinate space of N atoms $\hat{\mathbf{r}} \in \mathbb{R}^{3N}$. We learn the molecular reconstruction from the manifold $\mathcal{M}$ using a second fully-connected feedforward artificial neural network. All molecular renderings were made using Visual Molecular Dynamics Humphrey et al. [1996].

synthetic smFRET time series in a single intramolecular distance $v(t) \in \mathbb{R}^1$ (Fig. 3.1b). In this work, we take v to be the distance between two selected hypothetical fluorophore attachment positions. A more realistic approximation would explicitly model the FRET fluorophores within the MD simulation Zerze et al. [2014]. These trajectories constitute the training data necessary to construct the k -dimensional embedding of the all-atom simulation trajectory ($a \rightarrow c$), construct the k -dimensional embedding of the synthetic smFRET trajectory ($b \rightarrow d$), learn the transformation between them ($d \rightarrow c$), and learn the reconstruction of the molecular configuration from the low-dimensional all-atom manifold ($c \rightarrow e$). Once all steps in the pipeline are learned from the training data, STAR can be applied to reconstruct the molecular trajectories from new synthetic (or real) smFRET trajectories collected under the same conditions without conducting any additional molecular simulations via the pathway $b \rightarrow d \rightarrow c \rightarrow e$.

The primary focus of the present work is to test the application of STAR to smFRET trajectories with constraints on the number of individual smFRET measurements within the trajectory (i.e., data volume, Dv), the temporal resolution of the smFRET trajectory (i.e., bin size, λ), and under the presence of sampling noise that is controlled by the brightness of the FRET donor fluorophore (I_D). We train STAR models on computational training data with different values of (Dv, λ, I_D) and then test the models on novel synthetic smFRET data to assess the reconstruction accuracy. Importantly, we intentionally test our approach on synthetic smFRET data for which we can explicitly control Dv , λ , and I_D , and for which we possess the ground truth atomistic molecular simulation trajectories against which we can test the performance of our STAR reconstruction. Having defined the regimes of these three critical parameters within which STAR is determined to perform well for two fast-folding mini-proteins computationally amenable to simulation, we then prospectively identify protein systems and experimentally realistic FRET conditions capable of meeting these constraints and to which future applications of STAR to real experimental smFRET data may be successful.

Training data: Molecular dynamics simulations $\mathbf{r}(t)$ and synthetic smFRET time series $v(t)$)

We use all-atom MD simulations to furnish the training data necessary to learn the transformations within the STAR pipeline. The MD trajectory provides a time series of coordinates $\mathbf{r}(t) \in \mathbb{R}^{3N}$ (Fig. 3.1a) from which we generate the synthetic smFRET trajectory corresponding to a scalar time series of an intramolecular distance between two hypothetical FRET fluorophores $v(t) \in \mathbb{R}^1$ (Fig. 3.1b). In this work, we assume the fluorophores to be placed at the beginning and end of the linear proteins such that $v(t)$ corresponds to the molecular head-to-tail distance. We train STAR models on the synthetic smFRET data extracted from the MD trajectory at particular trajectory lengths Dv , time bin resolutions λ , and signal-to-noise ratios I_D . Models are trained over subsamples of the first 80% of each MD trajectory (*vide infra*) and the remaining 20% held out as a test partition. To assure good configurational diversity in the training data, we bin the training data into 10 equally spaced bins in the molecular head-to-tail distance and randomly sample $Dv/10$ configurations from each bin. As detailed below, Takens’ Theorem requires access to the immediate time history of each point and so we also collect the configurations preceding each selected point as far back as required, which, for the two protein systems considered in this work, lies in the range 2-220 ns. The training data may therefore be conceived of as an ensemble of short contiguous trajectories of the molecule distributed over a variety of head-to-tail distances to assure good sampling of its full configurational space. Hyperparameters are tuned for each component of the training pipeline using noiseless data at high time resolution and large data volumes and then applied to each (Dv, λ, I_D) triplet. We note that we train a single STAR model at a particular choice of parameters and apply it transferably to all other parameter regimes. An alternative strategy would be to train independent STAR models for each choice of these three parameters to better mimic within the training data the conditions of the testing set during model deployment. In our experience, this did not lead to significant improvements in performance.

Time resolution, λ . The operating principle of smFRET is to label two locations on a molecule with a pair of fluorophores whose absorption and emissions are distinct but overlapping, so that they may non-radiatively exchange energy upon excitation. This exchange of energy is governed by the relative geometry and spectra of the two fluorophores so that energy flows from the higher-energy donor fluorophore to the lower-energy acceptor at distances on the scale of 2-10 nm. The experimentally observed relative fluorescence of the donor and acceptor under donor-only excitation reports the efficiency of energy transfer and can be used to directly estimate the distance separating the two fluorophores as a “molecular ruler” Roy et al. [2008], Stryer and Haugland [1967]. Photon emission statistics are Poisson distributed, so FRET efficiencies are computed over finite time bins to mitigate the effects of noise. By recording the number of donor and acceptor photons over the course of a single time bin a single intramolecular distance is computed corresponding to an estimate of the average donor-acceptor distance over the time bin. Bins of $1\text{-}10^5 \mu\text{s}$ have been used in practice Oikawa et al. [2018], Phelps et al. [2017], Roy et al. [2008], with larger bins possessing improved signal-to-noise ratios but sacrificing temporal resolution. To mimic time binning in our synthetic smFRET trajectories, we bin the $v(t)$ time series into a sequence of bins of length λ and report for each bin the mean value of $v(t)$ recorded between time t and $(t + \lambda)$ as $v(t \rightarrow t + \lambda) = (1/\lambda) \int_t^{t+\lambda} v(t) dt$, where we approximate the integral as a discrete sum at the resolution of the synthetic time series. We perform an analogous operation for the corresponding MD training data $\mathbf{r}(t)$ wherein the Cartesian coordinates of each atom are averaged over the time bin λ . Typical values of λ for smFRET are on the order of ms for confocal FRET setups Hellenkamp et al. [2018], Roy et al. [2008], but can be pushed down to 1-10 μs for state-of-the-art photon-by-photon single-molecule instruments Phelps et al. [2017], Oikawa et al. [2018], Pirchi et al. [2016]. Training STAR models for different values of λ enable us to assess how the temporal resolution affects the reconstruction accuracy of the trained pipeline. In this work, we consider $\lambda = \{0.2, 1, 2, 5, 10\}$ ns for Chignolin and $\lambda = \{0.2, 1, 2, 10, 20\}$ ns for Villin as appropriate to capture the dynamics of these ultrafast-folding mini-proteins with characteristic folding times of

$\tau_{\text{fold}} = 0.6$ and $2.8 \mu\text{s}$, respectively Lindorff-Larsen et al. [2011]. These systems were selected for this work as sufficiently small and fast-folding to be amenable to good sampling with unbiased MD simulations. In the analysis of our results, we focus on the reconstruction accuracy as a function of the ratio $\lambda/\tau_{\text{fold}}$, which enables us to extrapolate our predictions to larger, slower-folding proteins that are much more challenging to sample using MD but are more typically studied by experimental smFRET.

Trajectory length, D_V . Training of the STAR pipeline requires MD training data for the protein of interest in order to learn the mappings denoted by the arrows in Fig. 3.1. MD simulations are typically limited to time scales of μs to ms, even on high performance and bespoke computational hardware Wolf et al. [2020], Lindorff-Larsen et al. [2011]. To assess the influence of training data volume upon the reconstruction accuracy of the trained STAR model, we consider a variety of training data volumes defined by the number of synthetic smFRET distance measurements D_V within the training ensemble. In this work, we select $D_V = \{10^3, 10^4, 2 \times 10^4, 4 \times 10^4\}$ observations using our training data selection criteria. If less than $D_V/10$ points are available in each head-to-tail decile, the total number of training points selected may be slightly less than D_V . As a point of comparison, experimental smFRET trajectories can vary in length from milliseconds to tens of seconds Hellenkamp et al. [2018], Roy et al. [2008], Asher et al. [2021] and employ temporal resolutions of several μs to ms Oikawa et al. [2018], Phelps et al. [2017], Pirchi et al. [2016], meaning that an experimental trajectory can contain 10^2 - 10^7 individual distance measurements.

Noise, I_D . The signal-to-noise ratio in smFRET time series are largely controlled by the intensity of the donor fluorophore: brighter fluorophores produce higher signal-to-noise ratios, whereas dimmer ones suffer more from the effects of noise. There are a number of sources of noise in smFRET emission measurements stemming from thermal fluctuations, biases in dye orientations or spatial distributions relative to the labeling sites, fluctuations in dye photophysical properties such as quantum yield due to local chemical environments, and fast blinking or other sub-time resolution

kinetics Roy et al. [2008]. Due to the discrete nature of photon counts, the noise in these counts can be modeled as shot or Poisson noise Roy et al. [2008], Nir et al. [2006]. Application of propagation of uncertainties to the relationship between FRET efficiency and donor-acceptor distance v allows us to derive a closed-form model for the relative uncertainty σ_v/v in this distance (see derivation in the Appendix),

$$\left(\frac{\sigma_v}{v}\right)^2 = \frac{1 + \left(\frac{R_0}{v}\right)^6}{36I_D(\lambda)} \left[1 + \left(\frac{v}{R_0}\right)^6 \left(\frac{\phi_D}{\phi_A}\right) \right], \quad (3.1)$$

where ϕ_D is the quantum yield of the donor, ϕ_A is the quantum yield of the acceptor, R_0 is the characteristic FRET radius for the donor-acceptor pair, and $I_D(\lambda)$ is the intensity of the donor channel (i.e., number of photons collected over the time bin λ) under direct excitation by the laser without acceptor present in the system. We then sample from this noise distribution to artificially corrupt the idealized bin-averaged distances extracted from our MD simulation trajectories,

$$v(t) \leftarrow v(t) + \mathcal{N}(0, \sigma_v^2), \quad (3.2)$$

where $\mathcal{N}(0, \sigma_v^2)$ is a random Gaussian variable with mean zero and variance σ_v^2 . A Gaussian noise model is a good continuous approximation for the underlying discrete Poisson statistics for sufficiently large photon counts, which, as described below, is the regime in which we operate.

In this work, we assume the FRET fluorophores to have ideal quantum efficiencies $\phi_A = \phi_D = 1$ and a typical FRET radius of $R_0 = 5$ nm. The uncertainty in the donor-acceptor distance v is then fully specified by the intrinsic brightness of the donor dye and the time bin λ over which photons are collected, which together define I_D . Brighter donors produce more photons in the donor and acceptor channels for a given distance v and larger time bins increase the absolute number of photons collected in the detector. Both of these effects therefore improve the signal-to-noise ratio in the calculated donor-acceptor distance. By deploying our trained STAR models on synthetic smFRET time series with different I_D , we can quantify how the reconstruction accuracy depends on

the signal-to-noise ratio. Experimental setups employing sophisticated burst-search algorithms Nir et al. [2006] can track on the order of 1,000-10,000 photon bursts per measurement Hellenkamp et al. [2018]. In this work, we consider $I_D = \{10^3, 10^4, 10^5, 10^6, \infty\}$, where $I_D = \infty$ corresponds to the hypothetical limit of an infinitely bright dye and noiseless conditions.

3.3.2 *Molecular Dynamics Simulations*

The MD simulations of the mini-protein Chignolin and the actin-binding protein Villin conducted by D.E. Shaw Research were used for STAR training and validation Lindorff-Larsen et al. [2011]. Simulations were performed on the supercomputer Anton under the CHARMM22* force field Piana et al. [2011] with a compatible modified TIP3P water model Jorgensen et al. [1983b]. Lys, Arg, Asp and Glu residues, N- and C-termini were simulated in their charged states Lindorff-Larsen et al. [2011]. Each system was equilibrated in the NPT ensemble using the Desmond software package on a PC cluster and reversible equilibrium folding simulations were performed on the Anton supercomputer in the NVT ensemble Lindorff-Larsen et al. [2011], Shaw et al. [2010]. The initial structure for the NVT ensemble simulation was chosen as the frame with the volume closest to the average volume. The system was then coupled to a Nose-Hoover thermostat with 1 ps relaxation time Nosé [1984], Hoover [1985]. Equations of motion were integrated at a 2.5 fs time step and frames were recorded every 200 ps Lindorff-Larsen et al. [2011]. All simulations were run in NVT ensemble using a Lennard-Jones potential with a 0.95 nm cutoff distance for short ranged electrostatics and the Gaussian Split Ewald method for long distance electrostatics with a $32 \times 32 \times 32$ cubic grid Shan et al. [2005], Lindorff-Larsen et al. [2011].

Chignolin. The 10-residue 166 atom mini-protein Chignolin peptide (PDB ID: 1UAO) Honda et al. [2004] was solvated in a box with 4 nm sides containing approximately 1900 water molecules Jorgensen et al. [1983b]. The (-2) peptide charge was neutralized with two Na^+ ions. A 106 μs MD simulation in the NVT ensemble was conducted Lindorff-Larsen et al. [2011]. Training data was subsampled from the first 80% of the trajectory and the last 20% used for testing. The characteristic

folding time of Chignolin is $\tau_{\text{fold}} = 0.6 \mu\text{s}$ Lindorff-Larsen et al. [2011].

Villin. The 35-residue 577 atom Villin polypeptide with protonated HIS residue (PDB ID: 2F4K) Kubelka et al. [2006] was solvated in 40 mM NaCl in a cube with 5.4 nm sides containing approximately 4400 water molecules. A 120 μs MD simulation in the NVT ensemble was conducted Lindorff-Larsen et al. [2011]. Training data was subsampled from the first 80% of the trajectory and the last 20% used for testing. The characteristic folding time of Villin is $\tau_{\text{fold}} = 2.8 \mu\text{s}$ Lindorff-Larsen et al. [2011].

3.4 Results and Discussion

We now report the results of our parametric study of STAR reconstruction accuracy as a function of $(D\nu, \lambda, I_D)$ for each of the two proteins Chignolin and Villin. Our general conclusions are that for data volumes of $D\nu \geq 10^4$ observations, time resolutions of $\lambda \leq 5$ ns for Chignolin and $\lambda \leq 10$ ns for Villin ($\lambda/\tau_{\text{fold}} \leq 8.3 \times 10^{-3}$ and 3.5×10^{-3} , respectively), and signal-to-noise ratios corresponding to $I_D \geq 10^5$ photons per bin, we are able to reconstruct Chignolin and Villin structures with heavy atom RMSD accuracies of 0.1-0.4 nm. These reconstruction fidelities lie in the same range as static reconstruction techniques such as x-ray crystallography and cryo-electron microscopy Schüttelkopf and van Aalten [2004], Chang and Rosenthal [2011].

3.4.1 Chignolin

STAR training

Calibration of the STAR hyperparameters is performed over configurations harvested from the first 80% of the 107 μs simulation trajectory (85.6 μs comprising 427,797 frames at 0.2 ns intervals) employing $D\nu = 40,000$, $\lambda = 0.2$ ns, and $I_D = \infty$. We choose to reconstruct the $N = 93$ heavy atoms producing a MD trajectory $\mathbf{r}(t) \in \mathbb{R}^{279}$. Diffusion maps were used to extract the manifold $\mathcal{M}$ from $\mathbf{r}(t)$, constructed with a kernel bandwidth $\varepsilon = e^{-3}$ nm. A gap in the eigenvalue spectrum informed

a 2D embedding spanned by $\{\Psi_1, \Psi_2\}$ (Fig. 3.7). Delay vectors $\mathbf{y}(t)$ were constructed from the time series in the head-to-tail distance $v(t)$ computed between the terminal heavy atoms. A delay dimensionality of $d = 11$ and delay time of $\tau = \lambda$ were employed. Diffusion maps used to extract the manifold $\mathcal{M}'$ from $\mathbf{y}(t)$ were constructed with a kernel bandwidth $\varepsilon = 1$ nm to produce 2D embeddings into $\{\Psi'_1, \Psi'_2\}$. The map from $\mathcal{M}'$ to $\mathcal{M}$ was parameterized by a 2-10-10-10-10-2 ANN trained over 100 epochs with a batch size of 500 and a learning rate of 10^{-4} using the Adam algorithm Kingma and Ba [2014]. The map from $\mathcal{M}$ to $\hat{\mathbf{r}}$ was parameterized by a 2-4-189-374-558-279 ANN trained over 120 epochs with a batch size of 400 and a learning rate of 10^{-5} using the Adam algorithm Kingma and Ba [2014].

STAR deployment and RMSD dependence on (Dv, λ, I_D)

Using the hyperparameters detailed in the previous subsection, we trained 100 independent STAR models at each combination of $\lambda = \{0.2, 1, 2, 5, 10\}$ ns, $Dv = \{10^3, 10^4, 2 \times 10^4, 4 \times 10^4\}$, and $I_D = \{10^3, 10^4, 10^5, 10^6, \infty\}$. The heavy atom RMSD reconstruction accuracies of each model on the 20% hold-out test partition (21.4 μ s comprising 106,946 frames at 0.2 ns intervals) is illustrated in Fig. 3.2. In general, we observe quite good RMSD reconstruction accuracies between 0.10-0.25 nm for all (Dv, λ, I_D) triplets considered with the exception of $I_D = 10^3$ triplets where the signal-to-noise ratio was too low to converge training of an ANN to learn the $\mathcal{M}' \rightarrow \mathcal{M}$ diffeomorphism and results are not reported. We must regard this accuracy with two important caveats. First, Chignolin is a very small protein comprising only 10 residues, so the per residue RMSD lies in the range 0.010-0.025 nm. Second, we can achieve a baseline RMSD accuracy on the test data of 0.155 nm by approximating each configuration by a single configuration from the test trajectory that results in the lowest overall RMSD averaged over all other configurations. The predictive power of the trained STAR models should therefore be viewed in light of improvements beyond this baseline accuracy.

The most obvious trend in Fig. 3.2 is the influence of I_D , for which we see an approximate halving in the RMSD reconstruction error from ~ 0.25 nm to ~ 0.12 nm in moving from $I_D = 10^4$ to 10^5 .

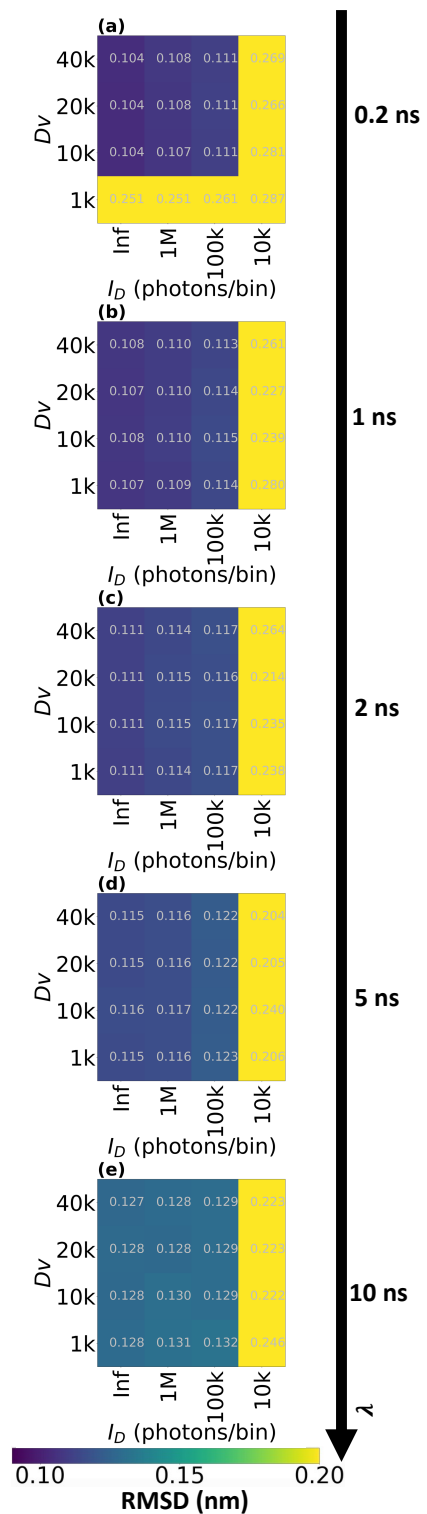


Figure 3.2: Heavy atom RMSD reconstruction accuracies for Chignolin as a function of training data volumes D_v , time bin resolution λ , and signal-to-noise ratio I_D .

Further increasing I_D to 10^6 or ∞ leads to relatively minor improvements in accuracy on the order of 3%. We were unable to learn a mapping between $\mathcal{M}'$ and $\mathcal{M}$ at $I_D = 10^3$. These trends indicate a clear floor on the signal-to-noise ratio necessary for reliable and accurate training of a STAR pipeline. We find reconstruction accuracy to be quite insensitive to data volume Dv . For a particular choice of I_D and λ , modulating Dv over the full range of 10^3 – 4×10^4 observations leads to only 1% changes in the RMSD accuracy except for $\lambda = 0.2$ ns triplets. One exception to this trend occurs for $\{Dv = 10^3, \lambda = 0.2 \text{ ns}, I_D \geq 10^4\}$ (Fig. 3.2a), where we observe substantially poorer reconstruction accuracies than at larger data volumes Dv (Fig. 3.2a) and lower time resolutions λ (Fig. 3.2b-e). We hypothesize that the high temporal resolution of the training data at $\lambda = 0.2$ ns means that larger data volumes are required for STAR to effectively learn the system dynamics than at larger λ values for two interlinked reasons. First, at fixed Dv a larger λ corresponds to a sampling of a longer time $Dv \times \lambda = Dv \times \tau$ that samples a longer period of the evolution of the system. Second, a larger λ coarse-grains over the sub- λ temporal fluctuations that may improve reconstruction accuracy by attenuating the high frequency system dynamics. Finally, the reconstruction error is moderately sensitive to the time resolution λ . For fixed I_D and Dv , increasing λ over the range 0.2-10 ns ($\lambda/\tau_{\text{fold}} = 3.3 \times 10^{-4}$ - 1.6×10^{-2}) leads to a progressive 20-30% degradation in the RMSD.

Taken together, our analysis reveals that for training data volumes $Dv \geq 10^4$ samples, photon counts per bin of $I_D \geq 10^5$, and time resolutions of $\lambda \leq 5$ ns ($\lambda/\tau_{\text{fold}} \leq 8.3 \times 10^{-3}$), we are able to achieve RMSD reconstruction accuracies of 0.12 nm or better, which is 23% better than the RMSD = 0.155 nm baseline. Conversely, in low signal-to-noise ratio (i.e., $I_D \leq 10^4$) regimes, reconstruction accuracy is worse than this baseline or training data is too noisy to permit convergence of a trained model.

Analysis of $\mathcal{M}$

For signal-to-noise ratios produced by $I_D \geq 10^5$ photons per bin, the dependence of the RMSD reconstruction accuracy upon data volume Dv and donor dye brightness I_D is relatively weak and

the primary determinant of reconstruction accuracy is the temporal time binning λ . To understand how the observed trends in the reconstruction accuracy as a function of λ can be attributed to the all-atom manifold $\mathcal{M}$, we present in Fig. 3.3a, the manifold $\mathcal{M}$ recovered from the MD trajectory $\mathbf{r}(t)$ at each λ value for $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ photons per bin. Taking the highest resolution manifold at $\lambda = 0.2$ ns as the ground truth (Fig. 3.3a-A), we observe two metastable free energy minima at $(\Psi_1 \sim 1.6 \times 10^{-6}, \Psi_2 \sim 9.5 \times 10^{-5})$ and $(\Psi_1 \sim 1.6 \times 10^{-4}, \Psi_2 \sim -8.5 \times 10^{-5})$ corresponding, respectively, to the native and unfolded metastable macrostates. The high free energy region connecting them corresponds to the transition paths linking these two states. Prior work has reported two-state or three-state free energy landscapes corresponding to folded, unfolded and misfolded states Wang et al. [2019], Shaffer et al. [2016], with variations attributed to differences in the simulated molecules and simulation protocols. Our results are in good agreement with those of Schaffer *et al.* Shaffer et al. [2016], who report a deep free energy well separated from a second metastable extended state by a relatively high free energy transition state. Taking the highest resolution manifold at $\lambda = 0.2$ ns as the ground truth, we observe that the primary impact of increasing bin size is an attenuation in the sampling of the transition region. This results in more sparse sampling of this region and an elevation in the apparent free energy of the transition pathways. This can be understood as a consequence of the enlarged window over which atomic positions are averaged with increasing λ that reduces the influence of transition states that are only fleetingly occupied relative to the comparatively long-lived metastable states. Conversely, the relative location of the metastable free energy basins are insensitive to the value of λ but become more smeared out and lose definition due with increasing λ due to increased averaging over molecular configurations.

Analysis of changes in the free energy landscape as a function of λ leads us to hypothesize that configurations within the transition region are likely to be more poorly reconstructed than those within the metastable basins due to the relatively poorer sampling of this region that is exacerbated at large λ . To test this hypothesis, we select from our validation trajectory five representative configurations A-E and project them onto each manifold $\mathcal{M}$ in Fig. 3.3a. Configurations A and B

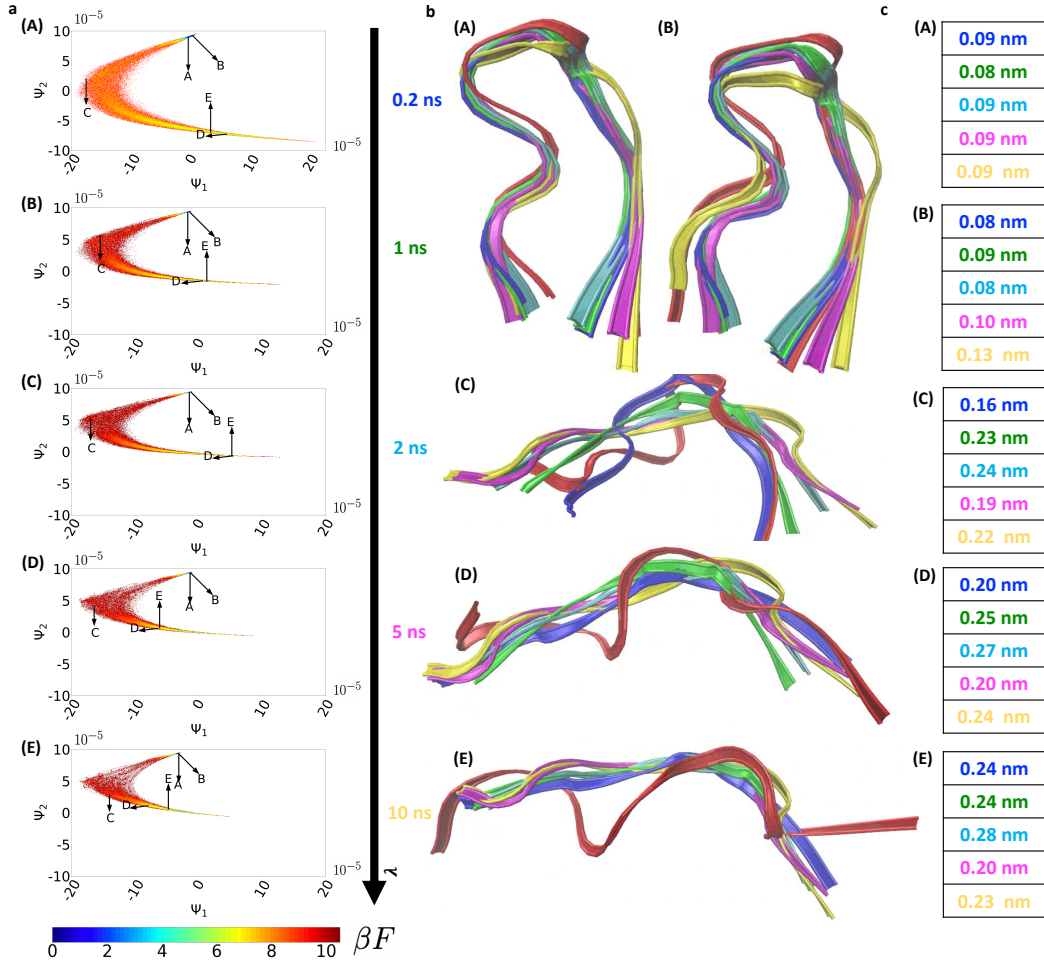


Figure 3.3: Manifolds $\mathcal{M}$ and position-dependent heavy atom RMSD reconstruction accuracy for Chignolin as a function of time bin resolution λ . (a) All-atom manifolds $\mathcal{M}$ within STAR pipelines trained at $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ for each value of $\lambda = \{0.2, 1, 2, 5, 10\}$ ns. Manifolds are represented as scatterplots showing the embedding of each of the $Dv = 4 \times 10^4$ MD configurations $\mathbf{r}$ into the two diffusion map eigenvectors $\{\Psi_1, \Psi_2\}$ spanning each manifold. Each point is colored by the associated free energy $\beta F(\Psi_1, \Psi_2)$ of that point computed by collecting histograms over the empirical probability distribution at a bin size of $\{\Delta\Psi_1 = 5 \times 10^{-7}, \Delta\Psi_2 = 5 \times 10^{-7}\}$. The arbitrary zero of free energy in each panel is specified by subtracting computed energy of the minimum free energy point on the landscape from all other points. Five representative configurations A-E are selected from the MD test trajectory and projected onto each manifold: A and B reside in the folded macrostate, C in the transition region, and D and E in the unfolded ensemble. (b) Visualizations of the reconstructions $\hat{\mathbf{r}}$ of each configuration A-E at each time resolution $\lambda = 0.2$ ns (blue), 1 ns (green), 2 ns (cyan), 5 ns (magenta), and 10 ns (yellow). For visual clarity, configurations are represented as ribbons tracing the backbone of the protein and are superposed upon the true configuration extracted from the MD test trajectory $\mathbf{r}$ (red). (c) Heavy atom RMSD reconstruction corresponding to each state are listed next to each image employing the same color-coding as the reconstructions. All molecular visualizations are constructed using VMD Humphrey et al. [1996].

reside within the folded macrostate, C within the transition region, and D and E within the unfolded ensemble. In Fig. 3.3b we present the reconstruction of each reconstructed configuration $\hat{\mathbf{r}}$ at each value of λ superposed together with the true configuration $\mathbf{r}$. The folded configurations A and B possess RMSD reconstruction accuracies of 0.08-0.13 nm, the unfolded configurations D and E possess accuracies of 0.20-0.28 nm, and the transition configuration C accuracies of 0.16-0.24 nm. Infrequent observation of transition states relative to both folded and metastable extended configurations results in poorer reconstruction of these transitory states, and this effect is indeed amplified at larger λ .

Although it is the case that transitory states are more poorly reconstructed than those residing in the metastable folded and unfolded wells, transition states are occupied by only $\sim 1\%$ of the test trajectory and therefore make only a small contribution to the overall mean RMSD accuracy. The overwhelming determinant of the degradation in the RMSD accuracy with increasing λ is therefore the globally poorer reconstruction of all configurations, even those within the relatively well sampled metastable wells, due to the loss of temporal resolution associated with the more severe degree of temporal averaging associated with the use of larger time bins.

3.4.2 Villin

STAR training

STAR hyperparameters are calibrated over configurations harvested from the first 80% of the 125.6 μs simulation trajectory (100.5 μs with 502,325 frames at 0.2 ns intervals) using $Dv = 37,465$, $\lambda = 0.2$ ns, and $I_D = \infty$. We reconstruct the $N = 287$ heavy atoms of Villin to produce an MD trajectory $\mathbf{r}(t) \in \mathbb{R}^{861}$. Diffusion maps with a kernel bandwidth $\varepsilon = e^{-2.5}$ nm were used to extract the manifold $\mathcal{M}$ from $\mathbf{r}(t)$ and a gap in the eigenvalue spectrum was used to determine a 3D embedding spanned by $\{\Psi_1, \Psi_2, \Psi_3\}$ (Fig. 3.8). Delay vectors $\mathbf{y}(t)$ were constructed from head-to-tail distance time series data $v(t)$ computed between terminal heavy atoms. A delay dimensionality of $d = 11$ and

delay time of $\tau = \lambda$ were employed. Diffusion maps with a kernel bandwidth $\varepsilon = 1$ nm were used to extract the manifold $\mathcal{M}'$ from $\mathbf{y}(t)$, producing 3D embeddings spanned by $\{\Psi'_1, \Psi'_2, \Psi'_3\}$. The map from $\mathcal{M}'$ to $\mathcal{M}$ was parameterized by a 3-15-15-15-15-3 ANN trained over 200 epochs with a batch size of 750 and a learning rate of 10^{-5} using the Adam algorithm Kingma and Ba [2014]. The map from $\mathcal{M}$ to $\hat{\mathbf{r}}$ was parameterized by a 3-6-578-1150-1722-861 ANN trained over 240 epochs with a batch size of 1000 and a learning rate of 5×10^{-7} using the Adam algorithm Kingma and Ba [2014].

STAR deployment and RMSD dependence on (Dv, λ, I_D)

Using the hyperparameters specified in the previous subsection, we trained 100 independent STAR models at each combination of $\lambda = \{0.2, 1, 2, 10, 20\}$ ns, $Dv = \{10^3, 10^4, 2 \times 10^4, 4 \times 10^4\}$, and $I_D = \{10^3, 10^4, 10^5, 10^6, \infty\}$ triplet. The heavy atom RMSD reconstruction accuracies of each model on the 20% hold-out test partition (25.1 μ s comprising 125,581 frames at 0.2 ns intervals) is illustrated in Fig. 3.5. We observe RMSD reconstruction accuracies between 0.35-0.46 nm for all (Dv, λ, I_D) triplets excluding $I_D = 10^3$ triplets where, as was the case for Chignolin, the low signal-to-noise ratio prevented the ANN failed from learning a converged $\mathcal{M}' \rightarrow \mathcal{M}$ diffeomorphic map. Villin is a medium sized protein comprising 35 residues, so the per residue RMSD lies in the range 0.010-0.013 nm. We can achieve a baseline RMSD accuracy from test data of 0.51 nm by computing the RMSD of every frame against the configuration from the test trajectory that results in the lowest overall RMSD averaged over all other configurations. STAR reconstruction performance should be judged relative to this baseline.

As with Chignolin, the clearest trend is the variation in reconstruction accuracy with the signal-to-noise ratio I_D , for which we observe in Fig. 3.4 a nearly 10% increase in RMSD reconstruction error from ~ 0.39 nm to ~ 0.44 nm in moving from $I_D = 10^5$ to 10^4 averaged across λ and Dv values. Further increases of I_D result in $\sim 1\%$ improvements averaging across all tested values of Dv and λ upon reaching the $I_D \rightarrow \infty$ limit. We note that the degree of improvement varies substantially with time resolution, approaching 10-16% in high time resolution regimes but falling to less than 1% at

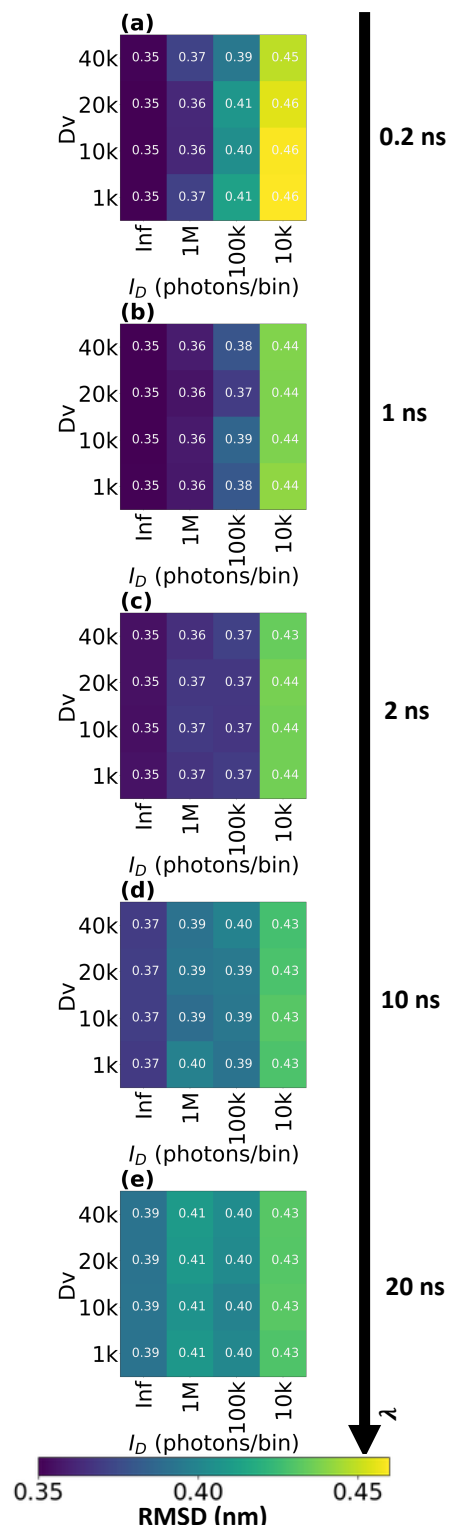


Figure 3.4: Heavy atom RMSD reconstruction accuracies for Villin as a function of training data volumes D_v , time bin resolution λ , and signal-to-noise ratio I_D .

low time resolutions. We were unable to learn a mapping between $\mathcal{M}'$ and $\mathcal{M}$ at $I_D = 10^3$. These trends reflect a floor on the test data signal-to-noise ratio required to reliably train the STAR pipeline. We find reconstruction accuracy to be largely insensitive to data volume Dv . For a particular choice of I_D and λ , modulating Dv over the full range of 10^3 – 4×10^4 leads to an average of less than 1% changes in the RMSD accuracy. Finally for a fixed I_D and Dv , increasing λ over the range 0.2-20 ns ($\lambda/\tau_{\text{fold}} = 3.6 \times 10^{-2}$ – 3.6×10^{-3}) leads to up to 14% degradation in the RMSD for $I_D \geq 10^4$, suggesting reconstruction quality is also quite sensitive to the time resolution λ .

For training data volumes $Dv \geq 10^4$ samples, photon counts per bin of $I_D \geq 10^5$, and time resolutions of $\lambda \leq 10$ ns ($\lambda/\tau_{\text{fold}} \leq 3.6 \times 10^{-3}$), we can achieve heavy atom RMSD reconstruction accuracies of 0.42 nm or better, representing an 18% improvement over the RMSD = 0.51 nm baseline. These conditions are largely the same as for Chignolin. However, for Villin, even in the low signal-to-noise ratio regime (i.e., $I_D = 10^4$), reconstruction accuracy is better than baseline value up until the failure to produce a map at $I_D = 10^3$.

Analysis of $\mathcal{M}$

For $I_D \geq 10^5$ photons per bin, the temporal time binning λ determines the majority of variation in reconstruction accuracy, with the dependence of the RMSD reconstruction accuracy on data volume Dv and donor dye brightness I_D being comparatively weak. We present in Fig. 3.5 the all-atom manifold $\mathcal{M}$ at each λ value for $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ photons per bin. For visual clarity, we consider a 2D projection of the 3D manifold into the Ψ_1 - Ψ_2 plane spanned by the two leading eigenvectors. This 2D projection is sufficient to illuminate the changes in global structure and free energy landscape over the manifold as a function of λ . Additional projections of the manifold are presented in Fig.3.9 in the Supporting Information. Considering the highest resolution manifold at $\lambda = 0.2$ ns as the ground truth (Fig. 3.5a-A), we observe a single free energy minimum at ($\Psi_1 \sim -3.1 \times 10^{-4}$, $\Psi_2 \sim -9.9 \times 10^{-4}$) corresponding to the folded state. The higher free energy region surrounding this region contains members of the unfolded ensemble. Previous

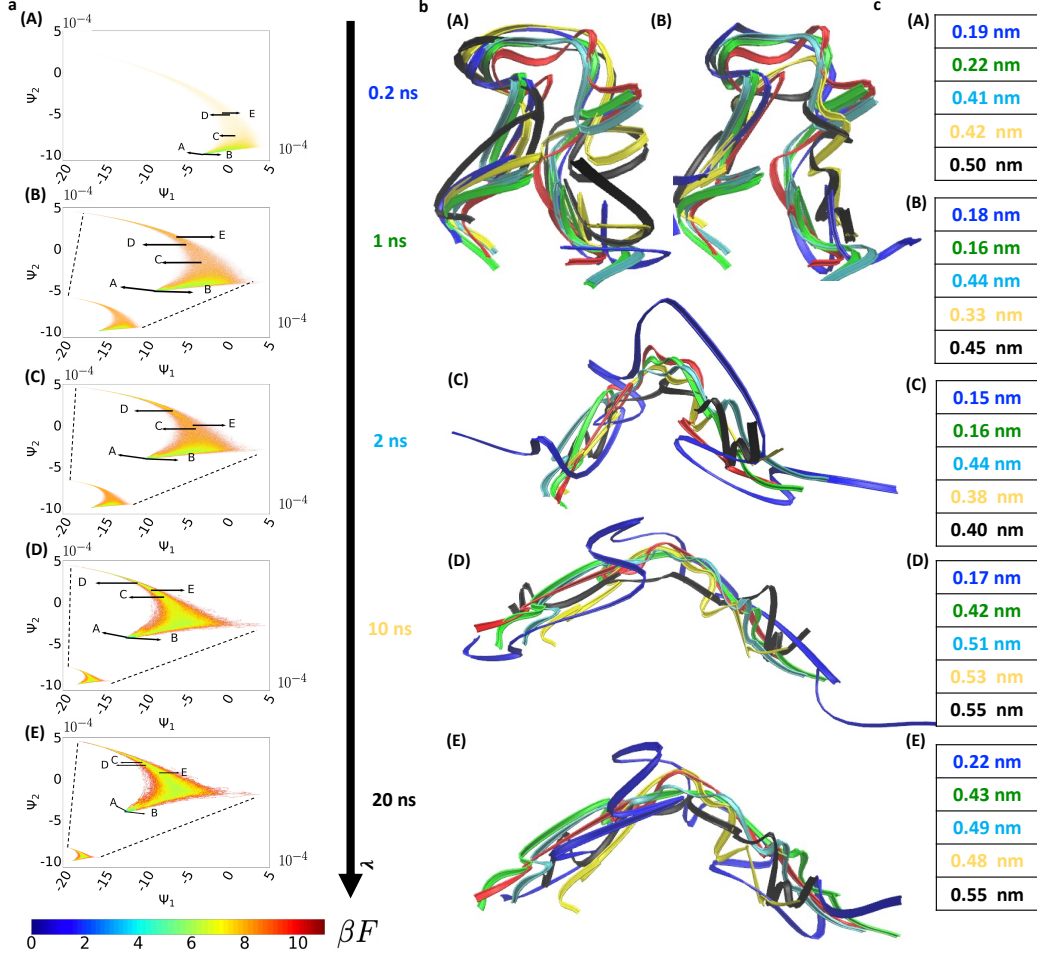


Figure 3.5: Manifolds $\mathcal{M}$ and heavy atom RMSD reconstruction accuracies for Villin as a function of time bin resolution λ . (a) All-atom manifolds $\mathcal{M}$ within STAR pipelines trained at $Dv = 4 \times 10^4$ samples and $ID = 10^5$ for each value of $\lambda = \{0.2, 1, 2, 10, 20\}$ ns. Manifolds are represented as scatterplots showing the embedding of each of the $Dv = 4 \times 10^4$ MD configurations $\mathbf{r}$ projected into the two leading diffusion map eigenvectors $\{\Psi_1, \Psi_2\}$ spanning each manifold. Zoomed in cutouts are provided for the sub-panels B-E. Color distributions characterize associated free energy $\beta F(\Psi_1, \Psi_2)$ of each point computed by collecting histograms over the empirical probability distribution at a bin size of $\{\Delta\Psi_1 = 5 \times 10^{-7}, \Delta\Psi_2 = 5 \times 10^{-7}\}$. The arbitrary zero of free energy in each panel is specified by subtracting computed energy of the minimum free energy point on the landscape from all other points. Five representative configurations A-E are selected from the MD test trajectory and projected onto each manifold. A and B reside in the folded macrostate while the C, D and E in the unfolded ensemble. (b) Visualizations of the reconstructions $\hat{\mathbf{r}}$ of each configuration A-E at each time resolution $\lambda = 0.2$ ns (blue), 1 ns (green), 2 ns (cyan), 10 ns (yellow), and 20 ns (black). For visual clarity, configurations are represented as ribbons tracing the backbone of the protein and are superposed upon the true configuration extracted from the MD test trajectory $\mathbf{r}$ (red). (c) RMSD reconstruction corresponding to each state are listed next to each image employing the same color-coding as the reconstructions. All molecular visualizations are constructed using VMD Humphrey et al. [1996].

work applying dimensionality reduction to molecular dynamics simulations of Villin similarly reflect this single free energy well Sittel et al. [2014]. Other works reflect the emergence of multiple energy wells at higher temperatures or due to denaturing Lei et al. [2007]. The most obvious effect of increasing λ from 0.2-20 ns is the translation and shrinkage of the free energy surface into the lower left corner of the $\Psi_1 - \Psi_2$ projection (Fig. 3.5a). The absolute values and magnitudes of $\Psi_1 - \Psi_2$ between different embeddings are not meaningful, only the relative locations of points within each embedding. As such, it is more informative to compare the zoomed insets of each image Fig. 3.5a-A-E that rescale each manifold onto approximately the same scale. These visualizations reveal that increases in λ result in increased averaging over contiguous molecular configurations and a concomitant mixing of configurations between the folded and unfolded states that smears out of the free energy minimum. At values of λ greater than 2 ns, this results in the a large free energy minimum that encapsulates both folded and unfolded states. In all cases, however, the relative locations of the folded and unfolded configurations are insensitive to the value of λ .

Again we hypothesize that configurations within the transition region and extended states outside of the metastable unfolded state are more likely to be poorly reconstructed than those within the metastable basins due to relatively poorer sampling at large λ . To test this hypothesis, we select from our validation trajectory five representative configurations A-E and project them onto each manifold $\mathcal{M}$ in Fig. 3.5a. Configurations A and B belong to the folded macrostate while C, D, and E are part of the unfolded ensemble. In Fig. 3.5b we present each reconstructed configuration $\hat{\mathbf{r}}$ at each value of λ superposed together with the true configuration $\mathbf{r}$. The folded configurations A and B possess RMSD reconstruction accuracies of 0.15-0.42 nm, the unfolded configurations C, D, and E possess accuracies of 0.33-0.55 nm. Although we see that STAR can adequately track the overall features of the protein conformational state over the five selected points, our analysis confirms that less frequently observed (i.e., higher free energy) states tend to have poorer reconstruction accuracies and that this effect is amplified with increasing λ as illustrated in Fig. 3.5c.

3.5 Conclusions

In this work, we have demonstrated the use of an approach based on Takens’ Delay Embedding Theorem termed Single-molecule TAKens Reconstruction (STAR) to predict the molecular structure of proteins from low-dimensional time series of intramolecular distances. The fundamental motivation of this approach is to provide a means to “upgrade” experimental measurements of intramolecular distances accessible to experimental techniques such as smFRET to a prediction for the atomistic coordinates of the molecule. In this manner, we can use a trained STAR model to furnish a time-resolved molecular trajectory directly from experimental data. The STAR models are trained over molecular simulation trajectories that provide the molecular configurations and intramolecular distances needed to learn the mapping from the latter to the former. The trained model may then be applied to novel smFRET data without the need to conduct any additional simulations. Provided the molecular simulation model employed is a good representation of the protein under the conditions of interest, the simulation trajectories are sufficiently long to sample the experimentally relevant configurational states, and the STAR model is properly trained, we anticipate that the model should be able to accurately predict molecular configurations from new (or hold-out) simulated smFRET trajectories and, ultimately, experimental smFRET data.

The primary contribution of the present work is to demonstrate that we can construct STAR models for two fast-folding mini-proteins, Chignolin ($\tau_{\text{fold}} = 0.6 \mu\text{s}$) and Villin ($\tau_{\text{fold}} = 2.8 \mu\text{s}$) under conditions of trajectory length, time resolution, and signal-to-noise ratio (i.e., dye intensity) that bridge computationally tractable simulations to experimentally realistic FRET conditions. The trained models achieve heavy atom RMSD reconstruction accuracies over a hold-out molecular dynamics test set of 0.12 nm and 0.42 nm, respectively. As a point of comparison, these accuracies are commensurate with the ~ 0.1 nm accuracies attainable by cryo-electron microscopy and X-ray crystallography Schüttelkopf and van Aalten [2004], Chang and Rosenthal [2011]. In each case we achieve these results by training over molecular simulation trajectories of $0.7\text{--}3.3\times$ the

characteristic protein folding time, with a temporal resolution of $1/120$ - $1/280\times$ the folding time, and signal-to-noise ratios commensurate with $\sim 10^5$ photons per time bin.

The present work demonstrates and validates STAR under against synthetic smFRET trajectories generated from hold-out molecular simulation trajectories. This is vital for validation of the method since the ground truth atomic coordinates of the testing trajectories are exactly known, but it would be desirable in future work to apply a trained STAR model to real experimental smFRET data. The mini-proteins studied herein are too fast folding to be accessible to existing smFRET technology, but our results lay the foundations and specify the experimental conditions necessary to perform extrapolative identification of putative target systems. State-of-the-art photon-by-photon single-molecule instruments can produce observations at $\lambda=1$ - $10\ \mu\text{s}$ with a fluorophore pair like Cy3/Cy5 Phelps et al. [2017], Oikawa et al. [2018], Pirchi et al. [2016]. The constraints on data volume, time resolution, and signal-to-noise ratio identified in this work, suggest that STAR could be deployed on proteins with characteristic folding times of $\tau_{\text{fold}} = 100$ - $1000\ \mu\text{s}$. Such a protein system would be simultaneously amenable to sufficiently high temporal resolution smFRET measurements using state-of-the-art probes and sufficiently fast-folding that it would require simulation trajectory training trajectories totaling 100 - $3,000\ \mu\text{s}$. STAR training only requires temporally-continuous blocks of molecular simulation trajectories of length $d\tau$, where d is the delay dimensionality and τ is the delay time, meaning that the training data can comprise a large number of short, discontinuous training trajectories efficiently generated by parallel computation. All-atom molecular dynamics simulations at these time scales are expensive but relatively accessible on high-performance supercomputing hardware Shaw et al. [2009], Lindorff-Larsen et al. [2011], Han and Schulten [2014], Yin et al. [2009]. This analysis suggests as one possible target system the 54-residue engrailed homeodomain protein (PDB: 1ENH), which possesses a characteristic unfolding time of $910\ \mu\text{s}$ at 25°C that can be modulated to $4.8\ \mu\text{s}$ at 63°C Mayor et al. [2000], has been extensively studied by molecular simulation Zhao et al. [2006], Koulgi et al. [2010] and is sufficiently large to accommodate FRET fluorophores operating within the preferred range of 2 - $8\ \text{nm}$ Mayor et al. [2000], Huang et al.

[2008].

In future work, we plan to increase the robustness of STAR to noise, integrate multichannel smFRET information, and explore the transferability of our models. Reduction in noise effects via kernel choice Takeda et al. [2006] or integration of hidden Markov models Liu et al. [2010] may help reduce photon count requirements. Multichannel smFRET signals simultaneously recording multiple intramolecular distances between multiple pairs of probes can be harnessed in conjunction with multivariate Takens' Theorem Cao et al. [1998], Gambin and Deniz [2010] to improve reconstruction quality by multiplexed dynamical observations. Study of optimal FRET fluorophore placement can help identify preferred grafting positions for the probes to maximize the dynamical information captured by smFRET observables and minimize reconstruction errors Mittal and Shukla [2018], Dimura et al. [2020]. Investigation of transferability of latent space manifolds and STAR mappings across temperature, pressure, molecular force field, coarse graining, and solvent viscosity would improve the versatility of trained STAR models while reducing training requirements. Furthermore, transitory and extended state reconstruction can be improved by adaptive sampling of rarely visited states Chen and Ferguson [2018] and facilitate applications of STAR to larger and slower folding molecules by judicious selection and generation of training data. Beyond protein reconstruction, we would also like to study other biomolecules such as DNA and RNA and also consider the incorporation of solvent-based observables Pietrucci [2020], Han et al. [2018], Stark [1999], Stark et al. [2003]. Lastly, we also see opportunities for applications of STAR to other fields where it is of interest to reconstruct the state of a high-dimensional dynamical system that is implicitly observed through an incomplete set of low-dimensional variables, including epidemiology, climatology, and econometrics.

Acknowledgments

This material is based on work supported by the National Science Foundation under Grant No. DMS-1841810. We are grateful to D.E. Shaw Research for sharing the simulation trajectories used as the basis of our model training. 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.

3.6 Appendix

3.6.1 *Synthetic smFRET noise model*

The distance between a smFRET donor and acceptor pair is computed from the measured FRET efficiency – the fraction of donor excitons that are transferred to the acceptor – by measuring the emission intensities (“brightnesses”) of the donor and acceptor fluorophores Roy et al. [2008], Hellenkamp et al. [2018]. Mathematically, the distance r is related to the emission intensities as Roy et al. [2008], Hellenkamp et al. [2018],

$$r = R_0 \left(\frac{\phi_A I_{DA}}{\phi_D (I_{AD} - I_A)} \right)^{1/6}, \quad (3.3)$$

where ϕ_D is the quantum yield of the donor, ϕ_A is the quantum yield of the acceptor, R_0 is the characteristic FRET radius for the donor-acceptor pair, I_A is the intensity of the acceptor channel under direct excitation by the laser without donor present in the system, I_{DA} is the intensity of the donor channel when acceptor is present, and I_{AD} is the intensity of the acceptor channel when donor is present. The quantum yields and FRET radius are determined by the particular choice of donor and acceptor fluorophores. Ideal quantum yields correspond to $\phi_D = \phi_A = 1$. Experiments typically use dyes with quantum yields are on the order of 0.1-1, with dyes like Rhodamine 6G having a

yield of 0.95 Roy et al. [2008], Magde et al. [1999]. Typical FRET radii are on the order of 1-10 nm Roy et al. [2008].

Assuming no correlations between independent variables, no detector noise, isotropic dye orientations, and no direct excitation of the acceptor fluorophore by the excitation laser ($I_A=0$), propagation of uncertainties yields,

$$\sigma_r^2 = \left(\frac{\partial r}{\partial I_{DA}} \right)^2 \sigma_{I_{DA}}^2 + \left(\frac{\partial r}{\partial I_{AD}} \right)^2 \sigma_{I_{AD}}^2, \quad (3.4)$$

where σ_r is the standard deviation in r , $\sigma_{I_{DA}}$ is the standard deviation in I_{DA} , and $\sigma_{I_{AD}}$ is the standard deviation in I_{AD} . The partial derivatives follow straightforwardly from Eqn. 3.3 as,

$$\frac{\partial r}{\partial I_{DA}} = \frac{r}{6I_{DA}}, \quad (3.5)$$

$$\frac{\partial r}{\partial I_{AD}} = -\frac{r}{6I_{AD}}. \quad (3.6)$$

Assuming Poisson statistics in photon counting, $\sigma_{I_{DA}}^2 = I_{DA}$ and $\sigma_{I_{AD}}^2 = I_{AD}$. It immediately follows that,

$$\sigma_r^2 = \frac{I_{AD}r^2 + I_{DA}r^2}{36I_{AD}I_{DA}} \quad (3.7)$$

$$\Rightarrow \left(\frac{\sigma_r}{r} \right)^2 = \frac{1}{36I_{DA}} \left(1 + \frac{I_{DA}}{I_{AD}} \right) \quad (3.8)$$

Since I_{DA} and I_{AD} are not directly available from our MD simulation trajectory, it is convenient to re-express the right hand side as a function of r and any fluorophore-specific constants. To do so, we first rearrange Eqn. 3.3 to eliminate I_{DA}/I_{AD} ,

$$\frac{I_{DA}}{I_{AD}} = \left(\frac{r}{R_0} \right)^6 \left(\frac{\phi_D}{\phi_A} \right), \quad (3.9)$$

and substitute into Eqn. 3.8 to yield,

$$\left(\frac{\sigma_r}{r}\right)^2 = \frac{1}{36I_{DA}} \left[1 + \left(\frac{r}{R_0}\right)^6 \left(\frac{\phi_D}{\phi_A}\right) \right]. \quad (3.10)$$

To eliminate I_{DA} we require an additional equation. Eqn. 3.3 computes r from experimental measurements of I_{AD} and I_{DA} that appeal to measurements of both the donor and acceptor channels. It is also possible to estimate r using the donor channel alone Schuler [2013], Hellenkamp et al. [2018],

$$r = R_0 \left(\frac{I_{DA}}{I_D - I_{DA}} \right)^{1/6}, \quad (3.11)$$

where I_D is the intensity of the donor channel under direct excitation by the laser without acceptor present in the system, which is dictated by the choice of donor fluorophore and laser power, and is not a function of r . Rearranging this expression for I_{DA} yields,

$$I_{DA} = \frac{I_D}{1 + \left(\frac{R_0}{r}\right)^6}. \quad (3.12)$$

Inserting Eqn. 3.12 into Eqn. 3.10 results in our noise model,

$$\left(\frac{\sigma_r}{r}\right)^2 = \frac{1 + \left(\frac{R_0}{r}\right)^6}{36I_D} \left[1 + \left(\frac{r}{R_0}\right)^6 \left(\frac{\phi_D}{\phi_A}\right) \right] \quad (3.13)$$

In Fig. 3.6, we present a plot of Eqn. 3.13 to illustrate the variation of σ_r/r as a function of r parameterized by I_D . We observe the lowest uncertainties in the vicinity of the FRET radius R_0 and an asymmetric increase as the donor-acceptor pair moves to smaller or larger separations. Physically, this is due to increased shot noise in I_{DA} at shorter distances as the FRET efficiency increases and the donor becomes less bright (i.e., fewer emitted donor photons), and increased shot noise in I_{AD} at longer distances as the FRET efficiency decreases and the acceptor becomes less bright. As anticipated, the relative error in r is mitigated by the use of brighter donors with larger values of I_D .

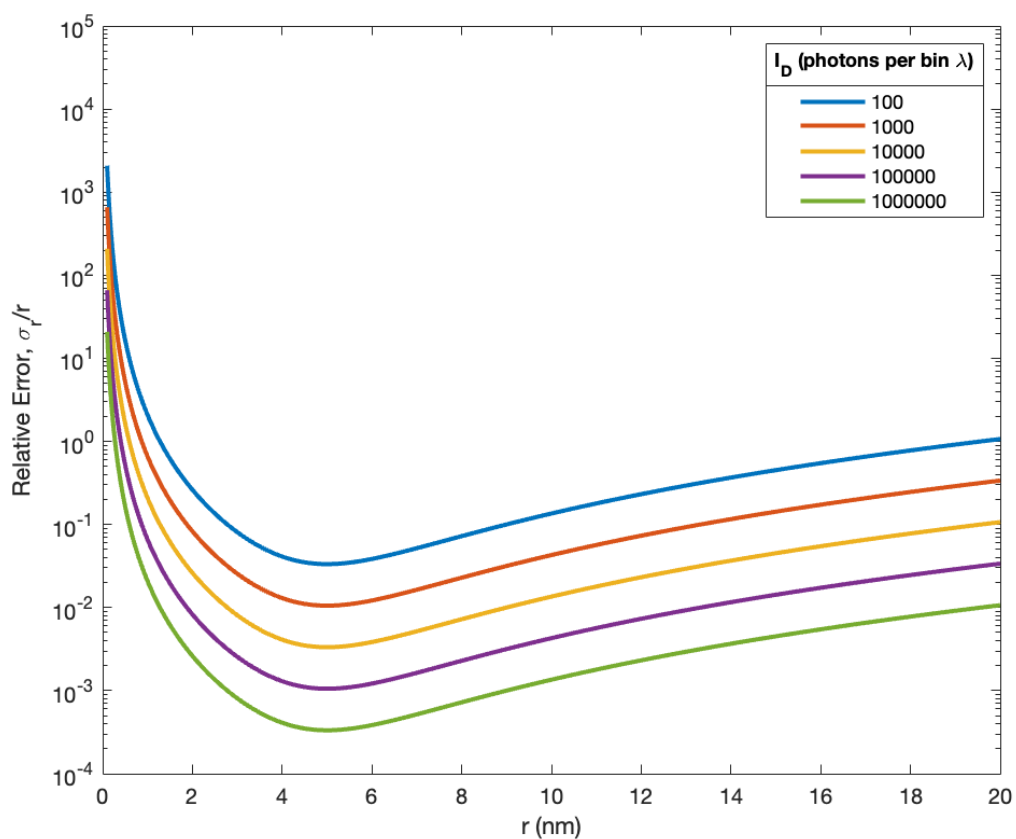


Figure 3.6: Variation of σ_r/r defining the relative uncertainty in r as a function of the donor-acceptor distance r at various values of donor intensity I_D (Eqn. 3.13). For the purposes of this calculation we adopt prototypical values of $\phi_D = \phi_A = 1$ for the donor and acceptor quantum yields and $R_0 = 5$ nm for the FRET radius of the donor-acceptor pair.

3.7 Supplementary Methods

Learning the atomistic manifold $\mathcal{M}$ from the MD trajectory $\mathbf{r}(t)$

Cooperative couplings between the configurational degrees of freedom in molecular systems mediated by the bonded and non-bonded interactions generically lead to emergent low-dimensional behaviors. Formally, the dynamical evolution of a molecular system is effectively contained within a $k \ll 3N$ -dimensional manifold $\mathcal{M}$ spanned by a small number of k collective variables that are much fewer in number than the $3N$ Cartesian degrees of freedom of the N constituent atoms Ferguson et al. [2010, 2011]. A number of unsupervised dimensionality reduction techniques have been applied to learn $\mathcal{M}$ from MD trajectories Ichiye and Karplus [1991], Chiavazzo et al. [2017], Ferguson et al. [2011, 2010], Rohrdanz et al. [2011], Das et al. [2006], Chen and Ferguson [2018], Sidky et al. [2020b], Chen et al. [2019]. In this work we employ diffusion maps Coifman et al. [2008a], Ferguson et al. [2010], Das et al. [2006], Ferguson et al. [2011], Coifman et al. [2005a], Coifman and Lafon [2006], Nadler et al. [2006b,a], Coifman et al. [2005b] as a form of nonlinear unsupervised dimensionality reduction to learn from the MD training trajectory $\mathbf{r}(t) \in \mathbb{R}^{3N}$ the dimensionality k of $\mathcal{M}$ and the k collective variables $\{\Psi_1, \Psi_2, \dots, \Psi_k\}$ with which to parameterize it. Diffusion maps construct a random walk over the data points constituting the MD trajectory and perform a spectral decomposition into the leading nonlinear collective variables parameterizing the extended directions of the data within the ambient $3N$ -dimensional space. A gap in the eigenvalue spectrum $\{\lambda_i\}$ informs the existence and effective dimensionality of a low-dimensional manifold $\mathcal{M}$ containing the data Ferguson et al. [2011], Nadler et al. [2006a]. The k eigenvectors below the spectral gap provide an embedding of the data into the collective variables $\{\Psi_i\}_{i=1}^k$ spanning the manifold.

Construction of the random walk requires specification of a distance metric $k(\mathbf{r}_i, \mathbf{r}_j)$ defining the dissimilarity of pairs of configurations $(\mathbf{r}_i, \mathbf{r}_j)$ within the MD trajectory and a bandwidth ϵ defining the characteristic step size of the random walk. The root mean squared deviation (RMSD)

between the atomic coordinates of aligned configurations is a natural measure of the configurational dissimilarity of single-chain protein molecules that is known to perform well as a diffusion map kernel Ferguson et al. [2010, 2011], Wang and Ferguson [2016, 2018a], Topel and Ferguson [2020]. In order for this metric to admit a diffeomorphism between the manifolds $\mathcal{M}$ and $\mathcal{M}'$ learned from the all-atom MD trajectory and synthetic smFRET data, Takens' Theorem requires that this metric be appropriately symmetrized to eliminate any spatial symmetries that cannot be distinguished by the choice of smFRET observable v Wang and Ferguson [2016, 2018a], Topel and Ferguson [2020]. Our choice of the head-to-tail molecular distance is invariant to translation, rotation, and mirror inversion. (For head-to-tail symmetric molecules such as simple polymer chains, there is an additional invariance to head-to-tail inversion, but none of the molecules considered in the present work possess this symmetry.) As such, we choose $k(\mathbf{r}_i, \mathbf{r}_j)$ to be the RMSD between pairs of configurations rotationally and translationally aligned using the Kabsch algorithm Kabsch [1976] that naturally eliminates the rotational and translational symmetries and minimized under a mirror inversion,

$$k(\mathbf{r}_i, \mathbf{r}_j) = \min [RMSD(\mathbf{r}_i, \mathbf{r}_j), RMSD(\mathbf{r}_i, \text{mirror}(\mathbf{r}_j))] . \quad (3.14)$$

After computing all $k(\mathbf{r}_i, \mathbf{r}_j)$ elements, the bandwidth ϵ can be automatically tuned based on the distribution of distances Ferguson et al. [2010], Coifman et al. [2008b].

The embedding of the MD trajectory into the leading k eigenvectors of the diffusion map defines the location $[\Psi_1(i), \Psi_2(i), \dots, \Psi_k(i)]$ of each configuration $\mathbf{r}(t_i)$ upon the intrinsic manifold $\mathcal{M}$. Any additional molecular configurations $\mathbf{r}_{\text{new}}$ not included within the diffusion map analysis can be projected onto the manifold using the Nyström extension Sunday et al. [2009], Laing et al. [2007], Long and Ferguson [2019a], Topel and Ferguson [2020]. The empirical distribution of configurations over the manifold $P(\Psi_1, \Psi_2, \dots, \Psi_k)$ permits estimation of the free energy surface supported by $\mathcal{M}$ as $\beta F(\Psi_1, \Psi_2, \dots, \Psi_k) = -\ln P(\Psi_1, \Psi_2, \dots, \Psi_k) + C$ where C is an arbitrary additive constant reflecting our ignorance of the absolute free energy scale and $\beta = 1/k_B T$.

Learning the Takens’ manifold $\mathcal{M}'$ from the smFRET time series $v(t)$

Takens’ Theorem asserts that the effectively k -dimensional MD trajectory $\mathbf{r}(t)$ evolving over $\mathcal{M} \subset \mathbb{R}^k \in \mathbb{R}^{3N}$ is uniquely specified by a $(d \geq (2k + 1))$ -dimensional delay embedding $\mathbf{y}(t) = [v(t), v(t - \tau), v(t - 2\tau), \dots, v(t - (d - 1)\tau)] \in \mathbb{R}^d$ of a generic scalar observable $v(t)$ constructed at a delay time τ and delay dimensionality d Wang and Ferguson [2018a], Takens [1981], Sauer et al. [1991], Packard et al. [1980], Broomhead and King [1986], Cao et al. [1998], Stark [1999], Stark et al. [2003], Vialar [2009], Kantz and Schreiber [2005], Zerze et al. [2014]. Importantly, Takens’ Theorem holds for both autonomous dynamical systems and those subject to deterministic or stochastic forcing Stark [1999], Stark et al. [2003]. This enables the approach to be applied to MD simulations of molecules whose dynamics may be coupled to a solvent bath, thermostats, and/or barostats that may be deterministic or stochastic in nature. Takens’ Theorem asserts that we should expect to recover a manifold $\mathcal{M}' \subset \mathbb{R}^k \in \mathbb{R}^d$ that is an image of $\mathcal{M}$ to which it is related by a diffeomorphism (i.e., smooth and invertible transformation) $\Theta : \mathcal{M}' \rightarrow \mathcal{M}$ Takens [1981], Sauer et al. [1991], Packard et al. [1980], Broomhead and King [1986], Vialar [2009], Kantz and Schreiber [2005], Topel and Ferguson [2020]. Formally, the manifold $\mathcal{M}'$ is a topologically, but not necessarily topographically, equivalent representation of $\mathcal{M}$ that preserves continuity, connectivity, and dimensionality. Colloquially, $\mathcal{M}'$ is a representation of $\mathcal{M}$ that is subject to an *a priori* unknown degree of position-dependent stretching and/or squashing, but which has not been torn apart or stitched together in new ways Takens [1981], Broomhead and King [1986], Topel and Ferguson [2020]. The crux of STAR rests upon this theoretical result: the bijective mapping between the two manifolds guaranteed by Takens’ Theorem provides a route to map each smFRET measurement to a corresponding all-atom molecular configuration.

We recover $\mathcal{M}'$ during the training phase of STAR by applying diffusion maps to the delay embedding $\mathbf{y}(t) = [v(t), v(t - \tau), v(t - 2\tau), \dots, v(t - (d - 1)\tau)]$ of the synthetic smFRET scalar time series $v(t)$. Although Takens’ Theorem holds for any choice of τ and $d \geq (2k + 1)$, in practice we determine a suitable value of τ as the first minimum of the autocorrelation of $v(t)$ Fraser and

Swinney [1986a] and d using the E_1 method of Cao Cao [1997], Villani and Zaldivar Comenges [2000], Kennel et al. [1992]. Values of τ are constrained to be an integer multiple of the bin size λ , so its minimum possible value is $\tau=\lambda$.

As was the case for our application of diffusion maps to the all-atom MD trajectories, we need to define a dissimilarity measure $k'(\mathbf{y}_i, \mathbf{y}_j)$ between delay vectors. In the same manner that Takens' Theorem required us to design our all-atom diffusion maps kernel to eliminate translational, rotational, and mirror symmetries for which the smFRET observable v is invariant, in this instance we must eliminate a temporal symmetries within the delay vectors for which the all-atom molecular configurations $\mathbf{r}$ are invariant. This symmetry originates from the imposition of a temporal ordering induced by the delay embedding vector Wang and Ferguson [2016, 2018a], Topel and Ferguson [2020]. Systems at equilibrium must obey the condition of detailed balance, which asserts that the forward and backward transitions between any pair of microstates have equal flux Zwanzig [2001], Tolman [1938]. For any microstate observed at a particular instant, it should not be possible to determine whether the microstate was an intermediate in a forward or reverse transition Wang and Ferguson [2016, 2018a], Topel and Ferguson [2020]. Assuming that we constrain d to be odd we may adopt a convention that defines a mapping between configurations r and delay vectors $\mathbf{y}$ based on the central element of the delay vector. This means that the same configuration maps to both $\mathbf{y}$ and $flip(\mathbf{y})$ – where the *flip* operation inverts the order of vector elements – introducing a spurious temporal symmetry arising from our artificial imposition of a temporal ordering within our delay vectors that detailed balance insists should be indistinguishable under time-reversal symmetry. We have previously demonstrated that it is vital to eliminate this symmetry in applications of Takens' Theorem to equilibrium molecular systems Wang and Ferguson [2016, 2018a], Topel and Ferguson [2020]. We design a diffusion map kernel to eliminate this symmetry by adopting a Euclidian distance that is minimized under flipping of the order of elements in the delay vectors,

$$k'(\mathbf{y}_i, \mathbf{y}_j) = \min [\|\mathbf{y}_i - \mathbf{y}_j\|_2, \|\mathbf{y}_i - flip(\mathbf{y}_j)\|_2]. \quad (3.15)$$

Again, the diffusion map bandwidth is automatically adapted to the distribution of distances $k'(\mathbf{y}_i, \mathbf{y}_j)$ Ferguson et al. [2010], Coifman et al. [2008b].

The leading k diffusion map eigenvectors $[\Psi'_1(i), \Psi'_2(i), \dots, \Psi'_k(i)]$ define the location of each delay vector $\mathbf{y}(t_i)$ upon the intrinsic manifold $\mathcal{M}'$. The empirical distribution of points defines the free energy surface supported by $\mathcal{M}'$ as $\beta F(\Psi'_1, \Psi'_2, \dots, \Psi'_k) = -\ln P(\Psi'_1, \Psi'_2, \dots, \Psi'_k) + C'$ where C' is an arbitrary additive constant. New delay vectors generated by novel (synthetic) smFRET trajectories can be projected onto $\mathcal{M}'$ using the Nyström extension Sunday et al. [2009], Laing et al. [2007], Long and Ferguson [2019a].

Learning the diffeomorphism from $\mathcal{M}'$ to $\mathcal{M}$

Takens' Theorem guarantees the existence of a diffeomorphism $\Theta : \mathcal{M}' \subset \mathbb{R}^k \rightarrow \mathcal{M} \subset \mathbb{R}^k$ defining a smooth and invertible mapping between the two manifolds, but an expression for Θ itself is *a priori* unknown. A number of numerical approaches exist to learn this mapping including kernel methods Schölkopf and Smola [2002], position dependent Jacobians Wang and Ferguson [2016, 2018a], Aalto and Verma [2019], and k-nearest neighbors Cover and Hart [1967], as well as more sophisticated techniques based on empirical isometrization of the manifolds to furnish intrinsic coordinate systems that are invariant to diffeomorphic transformations Dsilva et al. [2013], Peterfreund et al. [2020]. In this work, we approximate Θ by training a simple fully-connected feedforward artificial neural network (ANN) over $(\Psi(i), \Psi'(i))$ pairs, where $\Psi(i) = [\Psi_1(i), \Psi_2(i), \dots, \Psi_k(i)]$ is the embedding of configuration $\mathbf{r}_i$ from the MD trajectory onto manifold $\mathcal{M}$, and $\Psi'(i) = [\Psi'_1(i), \Psi'_2(i), \dots, \Psi'_k(i)]$ is the embedding of delay vector $\mathbf{y}_i$ from the smFRET time series onto manifold $\mathcal{M}'$. We adopt a convention wherein we associate each delay vector $\mathbf{y}(t)$ with the configuration $\mathbf{r}(t - (d - 1)\tau/2)$ corresponding to its middle element. To avoid ambiguity in this mapping we constrain d to be odd and we eliminate MD configurations within the first and last $(d - 1)\tau/2$ periods of our MD simulation trajectories from our analysis since they cannot be associated with any complete delay vector. We have found that simple fully-connected feedforward networks containing 4-8 hidden

layers with 10^3 - 10^4 neurons per layer are sufficient to learn accurate approximations to Θ . Networks were trained using the Adam algorithm with a learning rate of 10^{-4} to minimize the loss function $\mathcal{L} = \|\hat{\Psi}'(i) - \Psi'(i)\|_2$, where $\hat{\Psi}'(i) \in \mathbb{R}^k$ is the output of the ANN Kingma and Ba [2014]. We guard against overfitting by employing early stopping based on the validation error on a 20% randomly sampled hold-out validation partition.

Learning the reconstruction $\hat{\mathbf{r}}(t)$ from the manifold $\mathcal{M}$

Finally, we must learn reconstructions $\hat{\mathbf{r}}_i \in \mathbb{R}^{3N}$ of the molecular configurations $\mathbf{r}_i \in \mathbb{R}^{3N}$ from their projections $\Psi(i) = [\Psi_1(i), \Psi_2(i), \dots, \Psi_k(i)]$ onto manifold $\mathcal{M} \subset \mathbb{R}^k$. By construction, the manifold $\mathcal{M}$ contains the important high-variance collective variables corresponding to the most significant dynamical motions of the molecule. Encodings of $\mathbf{r}$ into the latent manifold $\mathcal{M}$ are therefore expected to preserve the important information on the salient molecular configurations and can be approximately recovered by learning the inverse decoding. If the diffusion map has done a satisfactory job in properly learning the collective variables spanning $\mathcal{M}$, then we anticipate that the information loss of the encoding should be restricted to the relatively unimportant low-variance collective degrees of freedom Wang and Ferguson [2016, 2018a], Topel and Ferguson [2020].

We learn the decoding from $\Gamma : \Psi \rightarrow \mathbf{r}$ over $(\Psi(i), \mathbf{r}_i)$ pairs from the training data using simple fully-connected feedforward ANNs containing 3-5 hidden layers each containing 10^3 - 10^4 neurons. These very simple and modestly sized networks were found to have more than sufficient capacity to learn accurate mappings. Networks were trained using the Adam algorithm with a learning rate of 10^{-5} to minimize the loss function $\mathcal{L} = \text{RMSE}(\hat{\mathbf{r}}_i, \mathbf{r}_i)$ Kingma and Ba [2014]. We guard against overfitting by employing early stopping based on the validation error on a 20% randomly sampled hold-out validation partition.

Deploying the trained STAR model

A separate STAR model is trained for each $\{\lambda, D_v, I_D\}$ combination, where training data is extracted from the first 80% of a single long MD simulation trajectory and the remaining 20% is held out as a test partition. Application of the trained STAR models have no further requirement for MD trajectory data and operate only on the synthetic smFRET time series $v(t)$ extracted from the 20% test partition. The synthetic trajectory $v(t)$ is processed as described in the main text to possess a particular time resolution λ and donor intensity I_D , then passed through the trained STAR pipeline to produce a reconstructed atomistic molecular trajectory $\hat{\mathbf{r}}(t)$. Commencing from the synthetic smFRET time series, the reconstruction process constructs a delay vector $\mathbf{y}(t)$, projects this onto the manifold $\mathcal{M}'$ using the Nyström extension Reinhardt [1985], maps over to the manifold $\mathcal{M}$ using the first trained ANN, and finally decodes to the all-atom reconstruction $\hat{\mathbf{r}}$ using the second trained ANN (Fig. 1*b* $\rightarrow$ *d* $\rightarrow$ *c* $\rightarrow$ *e*).

3.8 Supplementary Figures

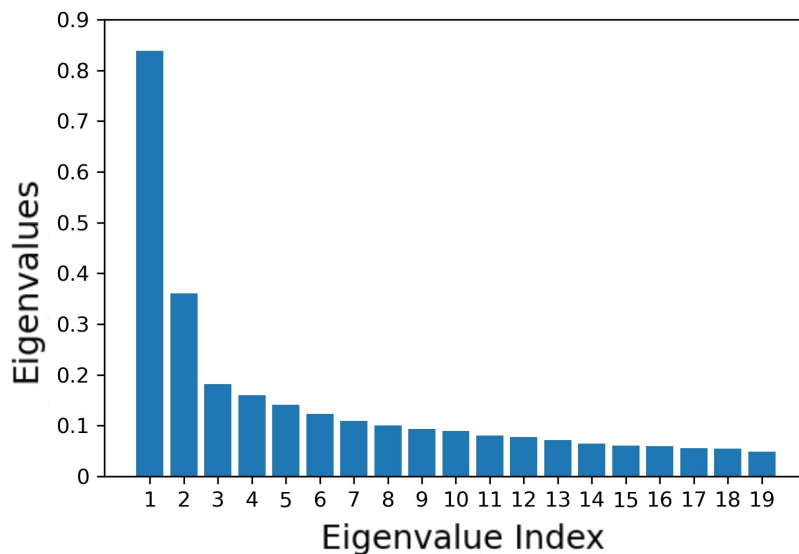


Figure 3.7: Eigenvalue spectrum resulting from the application of diffusion maps to the MD trajectory $\mathbf{r}(t) \in \mathbb{R}^{3N}$ for Chignolin. A gap in the eigenvalue spectrum after the second non-trivial eigenvalue λ_2 is identified using the L-method of Salvador and Chan Salvador and Chan [2004] and informs extraction of a 2D intrinsic manifold $\mathcal{M}$ spanned by $\{\Psi_1, \Psi_2\}$.

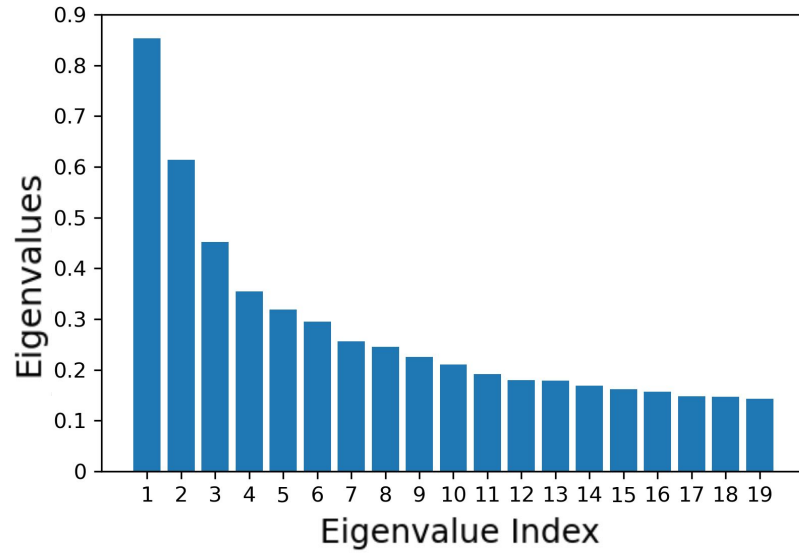


Figure 3.8: Eigenvalue spectrum resulting from the application of diffusion maps to the MD trajectory $\mathbf{r}(t) \in \mathbb{R}^{3N}$ for Villin. A gap in the eigenvalue spectrum after the third non-trivial eigenvalue λ_3 is identified using the L-method of Salvador and Chan Salvador and Chan [2004] and informs extraction of a 3D intrinsic manifold $\mathcal{M}$ spanned by $\{\Psi_1, \Psi_2, \Psi_3\}$.

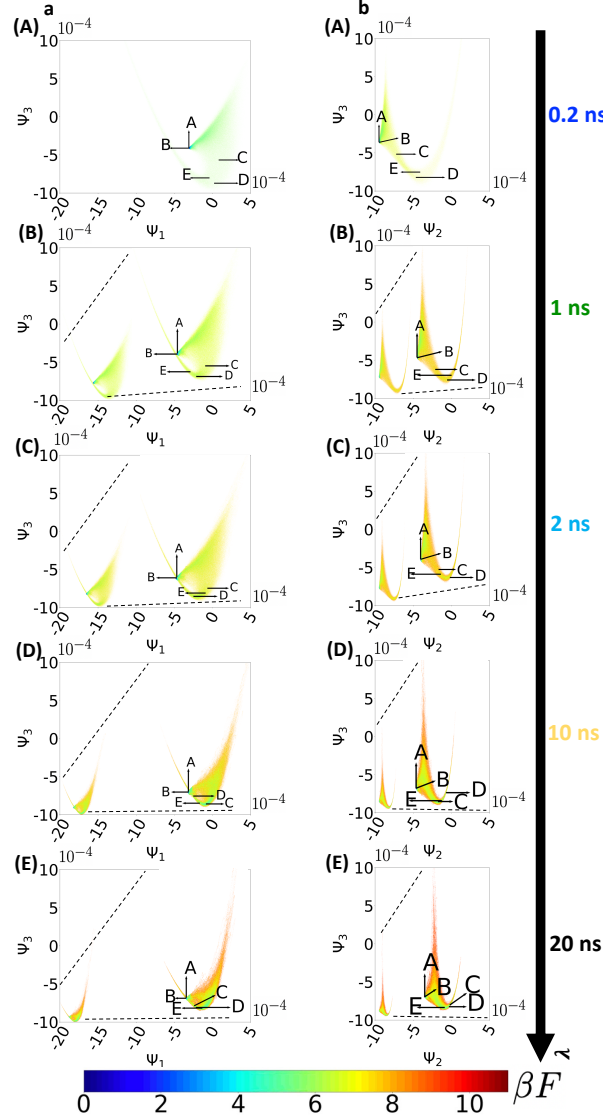


Figure 3.9: All-atom manifolds $\mathcal{M}$ for Villin within STAR pipelines trained at $D\nu = 4 \times 10^4$ samples and $I_D = 10^5$ for each value of $\lambda = \{0.2, 1, 2, 10, 20\}$ ns. The 3D manifolds in (Ψ_1, Ψ_2, Ψ_3) are represented as scatterplots showing the embedding of each of the $D\nu = 4 \times 10^4$ MD configurations $\mathbf{r}$ into the (a) $\{\Psi_1, \Psi_3\}$ and (b) $\{\Psi_2, \Psi_3\}$ projections. Color distributions characterize associated free energy $\beta F(\Psi_i, \Psi_j)$ of each point computed by collecting histograms over the empirical probability distribution at a bin size of $\{\Delta\Psi_1 = 5 \times 10^{-7}, \Delta\Psi_3 = 5 \times 10^{-7}\}$ and $\{\Delta\Psi_2 = 5 \times 10^{-7}, \Delta\Psi_3 = 5 \times 10^{-7}\}$. The arbitrary zero of free energy in each panel is specified by subtracting computed energy of the minimum free energy point on the landscape from all other points. The markers A-E indicate five representative configurations selected from the MD test trajectory and for which molecular visualizations are presented in Fig. 5b.

CHAPTER 4

OPTIMIZING DATA STRUCTURE TO LEARN DYNAMICS FROM TIME SERIES DATA

4.1 Abstract

Embedding theorems can be used to provide theoretical guarantees about the relation between low-dimensional observations of a system and the full-dimensional dynamics. Such theorems do not, however, provide guidance on optimal observable choice or methodologies to learn the mapping between the embedded and full-dimensional dynamics. In previous work, we have shown that it is possible to learn these relations through a combination of Takens' Theorem, manifold learning techniques and universal function approximators. Here we provide an algorithmic framework and software package that performs reconstructions of arbitrary dynamical systems from time series observables of those systems, TAKens Reconstruction (TAR). We use TAR to investigate how observable choice and delay vector structure impact reconstruction quality across a variety of systems. With an ecological system, we show that simple periodic dynamics can be reconstructed with observations taken over a wide variety of embedding parameters. With the protein Villin, we introduce the multi-temporal Takens' approach, demonstrating that dynamics from multiple time delays can be used to improve reconstruction results in systems with characteristic times of dynamics spanning many timescales. In a straight chain hydrocarbon, $C_{24}H_{50}$ we show that Kolomogorov-Sinai Entropies can predictively select optimal time series observables to learn system dynamics from. Finally, using trade histories from the order book of Vanguard S&P 500 ETF we demonstrate how TAR can be used to learn the rule sets behind systems for which we have no training data. This algorithm and these guidelines for maximizing time series information embedded in observable streams can be applied to arbitrary dynamical systems, aiding in the reconstruction of full-dimensional dynamics from arbitrary time series.

4.2 Introduction

Dynamical systems, be they collections of molecules, colonies of bacteria or groups of animals, are often modeled by breaking down each system in to its constituent elements and propagating them according to the appropriate differential equations relating each constituent of the system with all others. The state of each system can be given by the unique combination of coordinates serving as inputs to these equations. We can then define the accessible phase space corresponding to each dynamical system as those areas of coordinate space that any given dynamical system can visit. The full-dimensional representation of a given dynamical system can be described by a vector containing the coordinates of every input to the propagation equations.

In the case of computer simulations of an N atom molecule, for example, we can completely describe the molecular structure by recording the $\mathbb{R}^{3N}$ positions of every atom in that molecule and its solvent as we propagate them with molecular dynamics (MD) algorithms Hollingsworth and Dror [2018]. In experiments observing molecules evolving over time, however, such atomistic detail is not available Wu and Brand [1994], Roy et al. [2008]. The problem becomes even more pronounced when moving beyond MD to biological systems, where such calculations are intractable, or to social systems, where the underlying rules are unknown. We can, however, collect time series observations of most dynamical systems. A central objective of dynamical systems theory is to learn an injective and surjective mapping between experimental observations and the full-dimensional representation of a system’s dynamics Deyle and Sugihara [2011], Lu et al. [2022], Berry and Das [2023], Yair et al. [2017].

Limits on the nature of the relationship between time series and full-dimensional dynamics have been thoroughly explored analytically Takens [1981], Cao et al. [1998], Stark [1999], Broomhead and King [1986], Stark et al. [2003], Barański et al. [2020]. Takens’ Delay Embedding Theorem proved that, under mild technical conditions, there exists a bijective map between vectors constructed from time series observables and the full-dimensional representation of a system Takens [1981].

Stark and Broomhead elaborated upon this work to demonstrate that this map exists even for systems under stochastic or deterministic forcing Stark [1999], Stark et al. [2003], Broomhead and King [1986]. Cao furthered Takens’ univariate work by extending the mathematical principle to integrate information from multiple observables Cao et al. [1998]. Barański et al. showed that Takens’ Theorem can be extended to a non-dynamical probabilistic embedding theorem Barański et al. [2020]. While theoretical guarantees exist regarding the relation between experimental observations and full-dimensional system representations, Takens’ Theorem provides no insight on the form of this map. Moreover, the theorem pertains to systems for which we have a complete sampling of a phase space, giving little guidance on its utility in empirical studies where it is unknown if accessible phase space is fully sampled. The accessible observables and time series may not be arbitrary, but rather imposed by practical limitations. These considerations motivate empirical study of optimal data structuring to best represent system dynamics and backmapping of individual time series observations to full-dimensional phase space diagrams.

With the advent of artificial neural networks as universal functional approximators Hassoun et al. [1996], it has become possible to readily learn these *a priori* unknown mappings in systems where it is not possible to solve for these relations analytically. We combine the embedding theorem mathematical framework that guarantees diffeomorphic mappings between vectors constructed from arbitrary experimental observations and full-dimensional representations with universal functional approximators capable of estimating their forms. The central challenge we seek to address - how can we reconstruct full-dimensional dynamics from experimental observations - can be divided in two parts. Firstly, there is a methodological question regarding how one can perform these reconstructions. This has been the focus of some of our previous work Wang and Ferguson [2018a], Topel and Ferguson [2020], Topel et al. [2023]. Secondly, there is a practical question of how one should choose and structure the experimental observables used as inputs to these reconstruction pipelines. In this work, we describe several novel methods for selecting and structuring these time series. With a toy Lotka-Volterra predator prey model Wangersky [1978], we demonstrate the

effect of delay choice in reconstruction quality. With the straight chain hydrocarbon, $C_{24}H_{50}$, we explore observable selection based on minimization of chaos in our input data streams to reduce reconstruction errors. Using the example of the Villin protein, we leverage multiple timescale dynamics to implement Takens' Theorem over multiple time delays, allowing us to capture dynamics occurring on many time scales. We also extend the theoretical innovation of Cao Cao et al. [1998] in multiplexing Takens' Theorem to empirically perform reconstructions that engage multiple simultaneous observable streams. Lastly, we demonstrate how TAR can be used to learn the phase space of price changes in the Vanguard S&P 500 ETF in the absence of full-dimensional training data from historical time series.

We also provide a software package, Takens Reconstruction (TAR), to perform each of these phase space reconstruction calculations that can be downloaded at <https://github.com/maxtopel/TAR>.

The structure of the remainder of the paper is as follows. In Sect. 4.3, we discuss how arbitrary time series observations can be reduced to latent space representations of their dynamics. In Sect. 4.4, we introduce various ways of constructing delay vectors to capture information available from multiple observables or dynamics evolving across multiple time scales. In Sect. 4.5, we present the combined Takens Reconstruction (TAR) pipeline. In Sect. 4.6 we define dynamical systems metrics used to quantify the qualities of input data to assess which will optimally reconstruct dynamics via TAR. In Sect. 4.7, we apply TAR to a toy predator-prey ecological dynamics model to demonstrate the approach and test the influence of time delay choice on reconstruction quality. In Sect. 4.8, we study how observable choice effects reconstruction quality for a simple molecular system, the linear hydrocarbon $C_{24}H_{50}$. In Sect. 4.9, we demonstrate how multi-temporal Takens delay vectors constructed over the same observable can embed information from different timescales using the Villin protein. In Sect. 4.10, we demonstrate how TAR can perform phase space reconstruction of securities returns and estimate returns probability distributions at levels superior to standard models

in the field over short time scales. Finally, in Sect. 4.11, we discuss limitations of TAR and future avenues for its development and application.

4.3 Dynamics from Time Series

The propagation of a dynamical system may be represented by a set of coupled differential equations or a matrix rotation acting upon a state space vector that defines the system's position in phase space Brin and Stuck [2002], Strogatz [2015]. The state space of an N dimensional dynamical system evolving in time, can be written as $\mathcal{S} \subseteq \mathbb{R}^N$ with state $\mathbf{r}(t) \in \mathcal{S}$ at time, t . Typically, individual degrees of freedom of a dynamical system will interact with one another and these couplings lead $\mathcal{S}$ to admit a lower dimensional representation, $\mathcal{M} \subseteq \mathbb{R}^k$, where k is the intrinsic dimensionality of the system Eckmann and Ruelle [1985], Sauer et al. [1991], Tenenbaum et al. [2000], Roweis and Saul [2000a], Coifman and Lafon [2006], Coifman et al. [2005a]. This intrinsic dimensionality of the system, the minimum dimensionality required to accurately describe the dynamics of our system, is generally much smaller than N Eckmann and Ruelle [1985], Sauer et al. [1991]. For example, we can describe the evolution of a molecular system composed of n atoms, $\mathcal{S} \subseteq \mathbb{R}^N$, at discrete times, t , in full dimension with $N = 6n$ dimensions (3 spatial and 3 velocity for each atom), and plot its evolution in that high dimensional space. However, numerous works have demonstrated that the long time dynamical evolution can be adequately described in a low dimensional space of leading collective variables where $\mathcal{S}$ admits a low dimensional representation $\mathcal{M} \subseteq \mathbb{R}^k$ where $k \ll 6n$ Takens [1981], Stark [1999], Stark et al. [2003], Broomhead and King [1986] For instance, in previous work, we have shown that the proteins chignolin (PDB:2RVD) and Villin (PDB:2F4K) can be described in 2 and 3 dimensions respectively Topel et al. [2023].

4.3.1 Embedding theorems

The (weak) Whitney's embedding theorem guarantees that any continuous function $\mathcal{F}(t)$ evolving on an N dimensional manifold $\mathcal{S} \subseteq \mathbb{R}^N$ can be approximated by a smooth embedding on a

$2N + 1$ dimensional manifold spanned by at least k independent variables where k is the intrinsic dimensionality of the system Hirsch [1994]. However, for dynamical systems where some low dimensional attractor exists, we can extend this embedding principle to restrict the dynamics on $\mathcal{S}$ to some low dimensional phase space $\mathcal{M} \subseteq \mathbb{R}^k$ where $k \ll N$. Interactions between individual components of $\mathcal{S}$ mean that time series observations of each degree of freedom in a dynamical system may contain information about other degrees of freedom and consequently, the intrinsic dimensionality of the dynamics on $\mathcal{S}$ can be described in reduced dimension $\mathbb{R}^k$ where $k \ll 6n$ Eckmann and Ruelle [1985], Sauer et al. [1991]. There exists some manifold $\mathcal{M} \subseteq \mathbb{R}^k$ that then embeds the dynamics on $\mathcal{S}$ in a latent space. The dimensionality, k , is bounded by $k \leq d_S$, the box counting dimension of $\mathcal{S}$ Hirsch [1994].

High dimensional dynamics can, therefore, generally be embedded in reduced dimension, but how can these dynamics evolving in $\mathbb{R}^k$ with $k \geq 1$ be recovered from univariate time series? Time series offer a window into both the evolution of a system in phase space and the propagators driving these systems. Although a univariate time series $\mathcal{O}(t) \in \mathbb{R}^1$ evolves in a space of equal or lesser dimensionality to the intrinsic dimensionality of our system, the trajectories that system takes through $\mathbb{R}^1$ are generated by whatever dynamics propagates the system. Consequently, vectors constructed from time delayed observations of a system reflect those mathematically predetermined trajectories in deterministic systems. This approach is extend to cover deterministically evolving systems under both deterministic and stochastic forcing by Stark and Broomhead Stark et al. [2003], Stark [1999], Broomhead and King [1986] .

Takens' Delay Embedding Theorem can be viewed as an elaboration on the Whitney Embedding Theorem on time series of a single system observable. The theorem states that, if properly arranged, scalar time series data observed over a sufficiently long period of time to well-sampled phase space embed within them the same dynamics, up to a smooth and continuous mapping, as the full-dimensional representation Takens [1981]. More precisely, any generic observable, $\mathcal{O}(t)$, of a

dynamical system can be used to construct a time delay vector $y(t) = [\mathcal{O}(t), \mathcal{O}(t - \tau) \dots \mathcal{O}(t - 2m\tau)] \in \mathbb{R}^{2m+1}$ with time delay τ and intrinsic dimensionality m which enjoy a unique relation with the full-dimensional representation of that system's state space $\mathbf{r}(t) \in \mathcal{S} \subseteq \mathbb{R}^N$. This relation has three key components.

1. The dynamics of $y(t)$ and $\mathbf{r}(t)$ are C^1 -equivalent up to observable symmetry (i.e. identical up to a smooth and continuous mapping up to observable invariance or degeneracy)
2. $y(t)$ uniquely specifies the state of the system (i.e. $y(t)$ is defined by at least k independent variables that unambiguously define system configuration over all t)
3. $y(t)$ lies on manifold $\mathcal{M}'$ which is related by a smooth, invertable and bijective mapping (i.e. diffeomorphism) to the manifold $\mathcal{M}$, $\Theta : \mathcal{M}' \rightarrow \mathcal{M}$

Given the C^1 -equivalence of the dynamics of $y(t)$ and $\mathbf{r}(t)$ and the diffeomorphic relation between $\mathcal{M}'$ and $\mathcal{M}$, these manifolds are topologically identical. However, squashing or stretching of $\mathcal{M}'$ means that while continuity and connectivity are preserved, topography is not guaranteed to be preserved Takens [1981], Stark [1999], Stark et al. [2003], Sauer et al. [1991], Cao et al. [1998], Packard et al. [1980]. To ensure that $y(t)$ uniquely specifies the state of the system, we must collect sufficient observations such that we well sample the phase space accessible to our dynamical system Takens [1981]. Moreover, delay vectors constructed from observables that have some type of symmetry (e.g. mirror symmetry) will not allow for the construction of delay vectors able to resolve the difference between symmetric states.

Takens' Theorem can be used to construct an appropriate basis that spans the phase space accessible to the parts of the dynamical system's history sampled by our time series data Takens [1981]. Moreover, the theorem has been shown to hold for systems subject to both stochastic and deterministic forcing Broomhead and King [1986], Stark [1999], Topel et al. [2023]. Consequently, it provides a robust theoretical basis upon which we can build mathematically rigorous representations of

full-dimensional dynamics with data-driven models.

4.3.2 *Dimensionality Reduction and Manifold Learning*

Empirically based manifold learning techniques can also be used to learn a map from high dimensional representations of dynamics in $\mathcal{S} \subseteq \mathbb{R}^N$, to a low dimensional manifold $\mathcal{M} \subseteq \mathbb{R}^k, k \ll N$ Coifman and Lafon [2006], Coifman et al. [2005a], Roweis and Saul [2000a], Tenenbaum et al. [2000], Scholz et al. [2008]. Many of these techniques rely on a chosen distance metric (e.g. Euclidean, L1, cosine similarity, Mahalanobis...) between points in the high dimensional space to construct a lower-dimensional vector space, spanned by collective variables, that approximately preserve distances between points Coifman and Lafon [2006], Coifman et al. [2005a], Tenenbaum et al. [2000], Roweis and Saul [2000a], Scholz et al. [2008]. These choices will have an effect on the shape and structure of these latent space, embedding the same information differently. We can recover the intrinsic system dimensionality of time series data, for example, via Cao’s E1(d) method Cao [1997] and construct low-dimensional embeddings using manifold learning techniques Coifman and Lafon [2006], Coifman et al. [2005a], Tenenbaum et al. [2000], Roweis and Saul [2000a], Scholz et al. [2008]. Combining this manifold reduction technique with embedding theorems allows us to harness the information gains from ordering time series data via Takens’ Theorem while producing the lowest dimensional representation that still embeds the full-dimensional dynamics with manifold learning.

In this work, we apply the diffusion map nonlinear manifold learning technique to reduce full-dimensional training trajectories and time delayed vectors constructed from experimental observables to low dimension Coifman and Lafon [2006], Coifman et al. [2005a]. Diffusion maps are a technique that performs a spectral decomposition of a diffusion operator on a high dimensional data set to produce an embedding of that data in a lower dimensional space Coifman and Lafon [2006], Coifman et al. [2005a]. In doing so, diffusion maps allow for the embedding of a high dimensional data set on a lower dimensional manifold where the Euclidean distance between points

is approximately equal to the diffusion distance on the high dimensional manifold Coifman and Lafon [2006], Coifman et al. [2005a]. This favorable property of the latent space embeddings means that states that have high transition probabilities are embedded next to each other and those that have low transition probabilities are embedded further apart, such that the latent space tends to preserve the topology (more specifically, the connectivity) of the high dimensional phase space. Furthermore, diffusion maps also provide a straightforward means to extract intrinsic dimensionality through the spectral gap in the harmonic decomposition Nadler et al. [2006a,b] and also are very robust to noise Coifman and Lafon [2006].

4.4 Construction of Delay Vectors

In many experiments, we are able to take time series, $\mathcal{O}_t^j \in \mathbb{R}^1$, of observable j recorded at time t . Such observations may measure quantities related to individual elements of a system such as the price of a stock, multiple elements of a system such as the distance between two atoms in that molecule or measures of continuum dynamics such as the velocity of a fluid flow. In this section, we discuss methods of time delaying data to recover higher dimensionality dynamics from time series.

Univariate data streams: We can supplement a single time series observation $\mathcal{O}_t^j$ with observations from previous times to collect histories of these observables into delay vectors as proposed by Floris Takens [1981],

$$\mathcal{Y}_t^j = [\mathcal{O}_t^j, \mathcal{O}_{t-\tau}^j, \dots, \mathcal{O}_{t-2m\tau}^j]$$

where τ is the time delay between observations and m is the dimensionality of the delay vector. Conceptually, we can think of a system history as a way of specifying the path our system takes through phase space, thereby uniquely defining where on the high dimensional manifold $\mathcal{S} \subseteq \mathbb{R}^N$ the observation lies at time t and enabling us to discriminate between identical observations of $\mathcal{O}_t^j$.

Multivariate Data Streams: In many experiments, it is possible to collect multiple sources of

time series data. We can construct a multivariate time delay vector from multiple data streams as proposed by Liangyue Cao Cao et al. [1998], $\mathcal{Y}_t^{i,j}$, by concatenating $\mathcal{Y}_t^i$ and $\mathcal{Y}_t^j$

$$\mathcal{Y}_t^{i,j} = [\mathcal{O}_t^i, \mathcal{O}_{t-\tau^i}^i, \dots, \mathcal{O}_{t-2m\tau^i}^i, \mathcal{O}_t^j, \mathcal{O}_{t-\tau^j}^j, \dots, \mathcal{O}_{t-2m\tau^j}^j] \quad (4.1)$$

where τ^i and τ^j are time delays chosen for observable $\mathcal{O}^i$ and $\mathcal{O}^j$ respectively. Multivariate data streams allow for information from various sources with potentially different symmetries and noise susceptibilities to be combined to provide a more robust description of $\mathcal{M}'$. While we do not apply this method to any system in this paper, we have included it as an option for delay vector construction in the TAR software package.

Multi-temporal dynamics: Many real world dynamical systems may evolve on multiple time scales. Separability of dynamics occurring on varying time scales implies that we cannot learn information about separate dynamics occurring at a time scales $\mathcal{T}_\alpha \ll \mathcal{T}_\beta$ from the same time delayed vectors

$$\mathcal{Y}_t^j = [\mathcal{O}_t^j, \mathcal{O}_{t-\tau}^j, \dots, \mathcal{O}_{t-2m\tau}^j]$$

unless τ is sufficiently small and m is sufficiently large to span multiple time scales, resulting in excessively high dimensional time delay vectors. However, we can construct separate time delay vectors corresponding to the same discrete time, t , from the same time series data stream with distinct time delays τ_α and τ_β . We can then either concatenate these to construct a single time delay vector $\mathcal{Y}_t^{i;\alpha,\beta}$ from the same time series data stream, i ,

$$\mathcal{Y}_t^{i;\alpha,\beta} = [\mathcal{O}_t^i, \mathcal{O}_{t-\tau^i;\alpha}^i, \dots, \mathcal{O}_{t-2m_\alpha\tau^i;\alpha}^i, \mathcal{O}_t^i, \mathcal{O}_{t-\tau^i;\beta}^i, \dots, \mathcal{O}_{t-2m_\beta\tau^i;\beta}^i] \quad (4.2)$$

or from multiple time series data streams, i and j ,

$$\mathcal{Y}_t^{i;\alpha} = [\mathcal{O}_t^i, \mathcal{O}_{t-\tau^{i;\alpha}}^i, \dots, \mathcal{O}_{t-2m_\alpha\tau^{i;\alpha}}^i]$$

$$\mathcal{Y}_t^{i;\beta} = [\mathcal{O}_t^i, \mathcal{O}_{t-\tau^{i;\beta}}^i, \dots, \mathcal{O}_{t-2m_\beta\tau^{i;\beta}}^i]$$

where $m_{\alpha,\beta}$ are the appropriate delay dimensionalities for each time delay. Consideration of dynamics occurring over multiple characteristic times allows for access to information about the evolution of a given dynamical system from a single time series. In Sect. 4.9, we demonstrate that this innovation provides nontrivial information and performance gains for complex systems with lower experimental overhead than multivariate approaches.

Choosing delay times and dimensionalities: While Takens' Delay Embedding Theorem does not explicitly restrict our choice of time delay, τ , in practice, this value may be tuned to match characteristic time scales of the dynamics we wish to reconstruct. In the univariate case, delay times are chosen by taking the first minimum of each observable's autocorrelation or mutual information function or by choosing a cutoff value of the autocorrelation function to minimize redundancy (i.e. select τ such that $AC=0.2$) Fraser and Swinney [1986b]. For multi-temporal cases, we choose time delays that access different timescales. For instance, we may use time delays of 1s, 10s and 100s to get a full covering of delay times for dynamics spanning from 1-1000 seconds. We may also use physical knowledge of system dynamics, for instance, characteristic times of dynamics of particular phenomena known to influence a given system. We estimate the intrinsic dimensionality, k , of the system from time series using the E1(d) method of Cao Cao [1997], and select $m = k$ as dictated by Takens' Theorem Takens [1981].

4.5 Computational Application: Architecture of TAKens Reconstruction (TAR)

In this section, we assemble the TAR pipeline. We first structure our data inputs into univariate, multivariate or multi-temporal Takens vectors. We then reduce vector dimensionality with diffusion maps on to the intrinsic manifold. Finally, we learn approximate maps back to the full-dimensional phase space using artificial neural networks (ANNs). This pipeline serves as a tool set for optimization of data structuring to maximize relevant information in time series measurements, providing a learned low-dimensional embedding representation of the data to understand and predict system dynamics. Furthermore, if high dimensional training data is available, TAR serves as a means to approximately reconstruct the high dimensional system state from time series measurements. We have previously developed and relayed components of this protocol in the context of molecular systems in our STAR (Single-molecule TAKens Reconstruction) pipeline Topel and Ferguson [2020], Topel et al. [2023]. The present work extends this approach to arbitrary dynamical systems.

4.5.1 Data Structuring

We begin by assembling our time series inputs in time delay vectors (Fig. 4.1(a)-(b)) and splitting them up in to training and test set. For a time series with K observations with time delay τ and delay embedding dimensionality $2m + 1$, we can construct $K - 2m\tau$ Takens delay vectors from the entirety of our time series. In all cases we use the first half of the time series trajectory for training and the second half for testing.

4.5.2 Manifold Learning

We employ diffusion maps to learn representations of the intrinsic manifold $\mathcal{M}'$ from Takens delay vectors and, if available, $\mathcal{M}$ from full-dimensional system observations (Fig. 4.1(b)-(c) and Fig. 4.1(f)-(d) respectively). For some high dimensional data $V \in \mathbb{R}^N$, we compute the pairwise distances

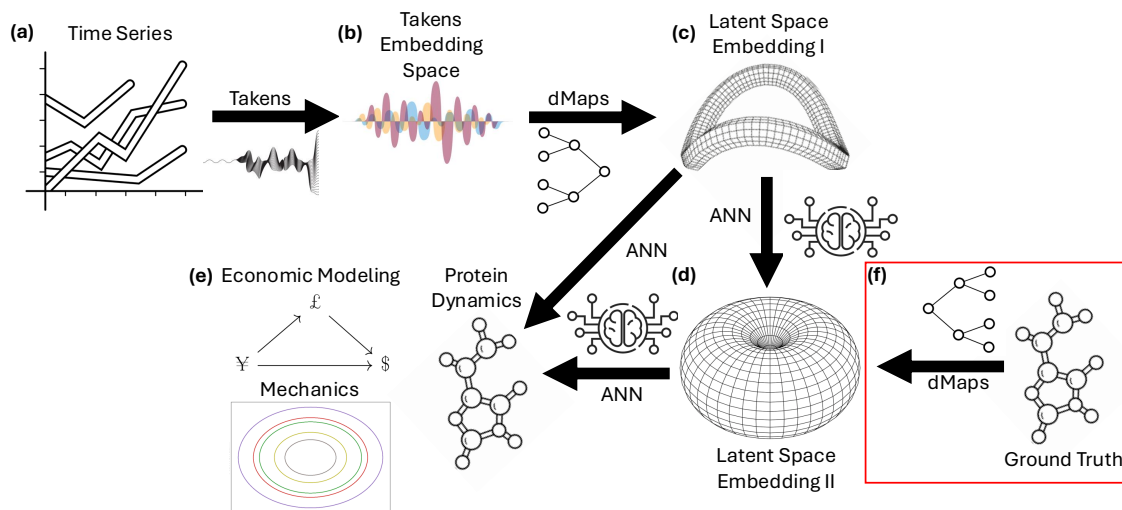


Figure 4.1: Schematic of the TAR pipeline. (a) TAR is given an input time series. (b) The time series is structured into either univariate, multivariate or multi-temporal Takens delay vectors generating our Takens embedding manifold. (c) Latent space extraction using diffusion maps manifold learning to learn latent space embedding I, $\mathcal{M}'$. (e) Extraction of a latent space embedding, $\mathcal{M}$, from a full-dimensional training trajectory. This step is only required if a training trajectory is available. (d) Reconstruction by backmapping to the latent space representation of the full-dimensional system learned from the full-dimensional system observations in (e). If no full-dimensional training trajectory is available, we learn the map from (c) to (e) directly via an artificial neural network.

between each point $v \in V$, construct a kernel matrix by applying a Gaussian kernel to the pairwise distance matrix and normalize that matrix to guarantee that we are accurately representing the transition matrix. We then transform the similarity matrix to a Laplacian matrix and finally perform eigenvalue decomposition of this matrix to get the eigenvectors that span our new latent space Coifman and Lafon [2006], Coifman et al. [2008a]. We choose an appropriate Gaussian kernel width by recursively performing diffusion maps over a small training set, normally 5000 points, estimating intrinsic dimensionality using the L method of Salvador and Chan, and measuring the implied dimensionality against the results of our E1(d) analysis Salvador and Chan [2004], Cao [1997], Coifman and Lafon [2006]. We repeat this process until we have agreement on the implied dimensionality and then perform diffusion maps on a larger training set. The remainder of the delay vectors not used in training are Nyström extended to the latent space manifold Williams and Seeger [2000]. This use of diffusion maps allows us to reduce the dimensionality of the system we are observing while allowing for an appropriate representation of nonlinearity in system dynamics. The values of each latent space coordinate extracted by diffusion maps do not necessarily reflect an easily interpretable collective variable. As each eigenvector produced by diffusion maps seeks to create a phase space where distance reflects transition probability, the exact values of each coordinate are not relevant but rather their positions relative to one another.

4.5.3 *Reconstruction*

We use simple feed forward fully-connected artificial neural networks (ANNs) Hassoun et al. [1996] to learn the mappings between latent spaces $\mathcal{M}'$ extracted from time series and $\mathcal{M}$ extracted from full-dimensional trajectories (Fig. 4.1(c)-(d), Fig. 4.1(c)-(e) and 4.1(d)-(f)). In cases where the full-dimensional trajectory is already low dimensional we map directly to the trajectory. In cases where the full-dimensional trajectory is high dimensional, we map first from the latent space $\mathcal{M}'$ extracted from a given time series to that $\mathcal{M}$ constructed from full-dimensional trajectories and then finally to the full-dimensional trajectory.

4.5.4 *Synthesis*

The TAR pipeline has three variants depending on the nature of the data: (1) explicitly high-dimensional dynamical systems with intrinsically low-dimensional dynamics for which we have both time series and corresponding full-dimensional trajectories are available, (2) explicitly low-dimensional dynamical systems for which we have both time series and corresponding full-dimensional trajectories are available and (3) intrinsically low-dimensional dynamical systems for which we only have time series available.

For high dimensional dynamical systems for which single or multivariate time series and corresponding full-dimensional trajectories are available, we construct Takens delay vectors as in Fig. 4.1(a)-(b) either using univariate, multivariate or multi-temporal Takens delay vectors and then construct a latent space $\mathcal{M}'$ manifold by running diffusion maps on the training portion of the Takens delay vectors as in Fig. 4.1(b)-(c). This latent space $\mathcal{M}'$ is diffeomorphically related to the embedding of the full-dimensional dynamics $\mathcal{M}$ provided we have sufficiently adhered to the dictates of Takens' Theorem. This manifold is the final product in case (3). In cases (1) and (2) we can use an ANN to map back to either the latent space embedding $\mathcal{M}$ of the full-dimensional system learned by applying diffusion maps to the full-dimensional system observations or the full-dimensional system itself corresponding to Fig 4.1(c-d) then (d-e) and (f-d) and (d-e) respectively. We opt to do the former for case (1) as an intermediary to mapping back to the full-dimensional system, as this improves results empirically, and map directly to the full-dimensional system in case (2).

4.6 Dynamical systems metrics

There are a number of information theoretic and dynamical systems measures available to interpret time series data. These include the degree to which a system exhibits chaotic behavior (Lyapunov exponents Wilkinson [2016]) and the rate of information loss as a result of chaotic behavior (Kolmogorov-Sinai Entropy Dorfman [1999]). Further measurements, the amount of 'information

surprise’ in a data set (Shannon Entropy Shannon [1948]) and the information shared by two univariate time series (mutual information Fraser and Swinney [1986b]) are discussed in supporting materials section 4.12.1. We combine these different metrics to optimize time delay and observable choice in constructing our Takens delay vectors.

Deterministically evolving dynamical systems that exhibit chaos may not be predictable after a certain characteristic time if we begin our observations with limited information. Chaotic systems display properties such as strong sensitivity to choice of initial conditions, topological transitivity and dense periodic orbits Hasselblatt and Katok [2010], that make prediction difficult in spite of this deterministic evolution. We can measure the degree to which those initial conditions matter to the evolution of a time series using Lyapunov exponents and estimate the information loss over time using the Kolmogorov-Sinai Entropy Wilkinson [2016], Dorfman [1999].

Lyapunov exponents: We can measure the rate of divergence of two trajectories with infinitesimally close initial conditions in a phase space as a function of the distance between those trajectories in that space. This measure serves to quantify the degree to which initial conditions alter trajectories in phase space and thus how ‘chaotic’ a system is. The distance, $D(t)$, is then related to the Lyapunov exponent, λ and the initial separation δd by the relation Rosenstein et al. [1993]:

$$D(t) = e^{\lambda t} \delta d$$

There will be a number of Lyapunov exponents corresponding to the dimensionality of the phase space as rates of separation may vary by dimension. We hypothesize that time series that display lower levels of chaos as measured by their Lyapunov exponents should be more easily predicted than those with higher levels of measured chaos. Takens’ Theorem applies to generic observables and, in the limit of infinite data, we expect that all time delayed representations of our observable data will be related in a bijective fashion. The stretching or squeezing of the true phase space

manifolds constructed from Takens' Theorem can result in local differences in rates of divergence of the same trajectory across time series of different observables of the same system. Even in the large information limit, different observables will likely produce manifolds that while topologically equivalent are not topographically equivalent by Takens' Theorem. In this limit, we expect convergence of the Lyapunov exponents on each manifold by Oseledets Theorem Walters [1993]. However, given topographical differences on the manifolds, Lyapunov exponents across all manifolds should not converge to a unique set of values. We hypothesize that the measure of chaotic character in a time series of an observable may be used as a principle to guide selection of an observable that in the finite data limit, can promote higher predictive accuracies of the system state within TAR. Specifically, we conjecture that the time series of observables with smaller Lyapunov exponents, and therefore lower chaotic character, should produce higher accuracy TAR predictions. We compute Lyapunov exponents from time series using the nolds package Schölzel [2020], Eckmann et al. [1986].

Kolmogorov-Sinai Entropy: Related to this sensitivity to initial conditions is the rate of information creation and loss in our observable data stream, quantified by the Kolmogorov-Sinai Entropy (KSE) Dorfman [1999]. For (deterministic) chaotic systems where we do not have full knowledge of our system, our prediction capabilities are limited to a certain characteristic time bound by this rate of information loss Dorfman [1999]. We can use the sum of the positive Lyapunov exponents to compute an upper bound quantity for n dimensional dynamics of Takens delay vectors constructed from a time series, $\mathcal{X}$ following Ruelle's Inequality Ruelle [1978]:

$$KSE(\mathcal{X}) \leq \sum_{i=1}^n \lambda_i$$

We can use this quantity to rank order time series before performing reconstructions, making use of those with the smallest KSEs as, in principle, they correspond to greater long term predictive power.

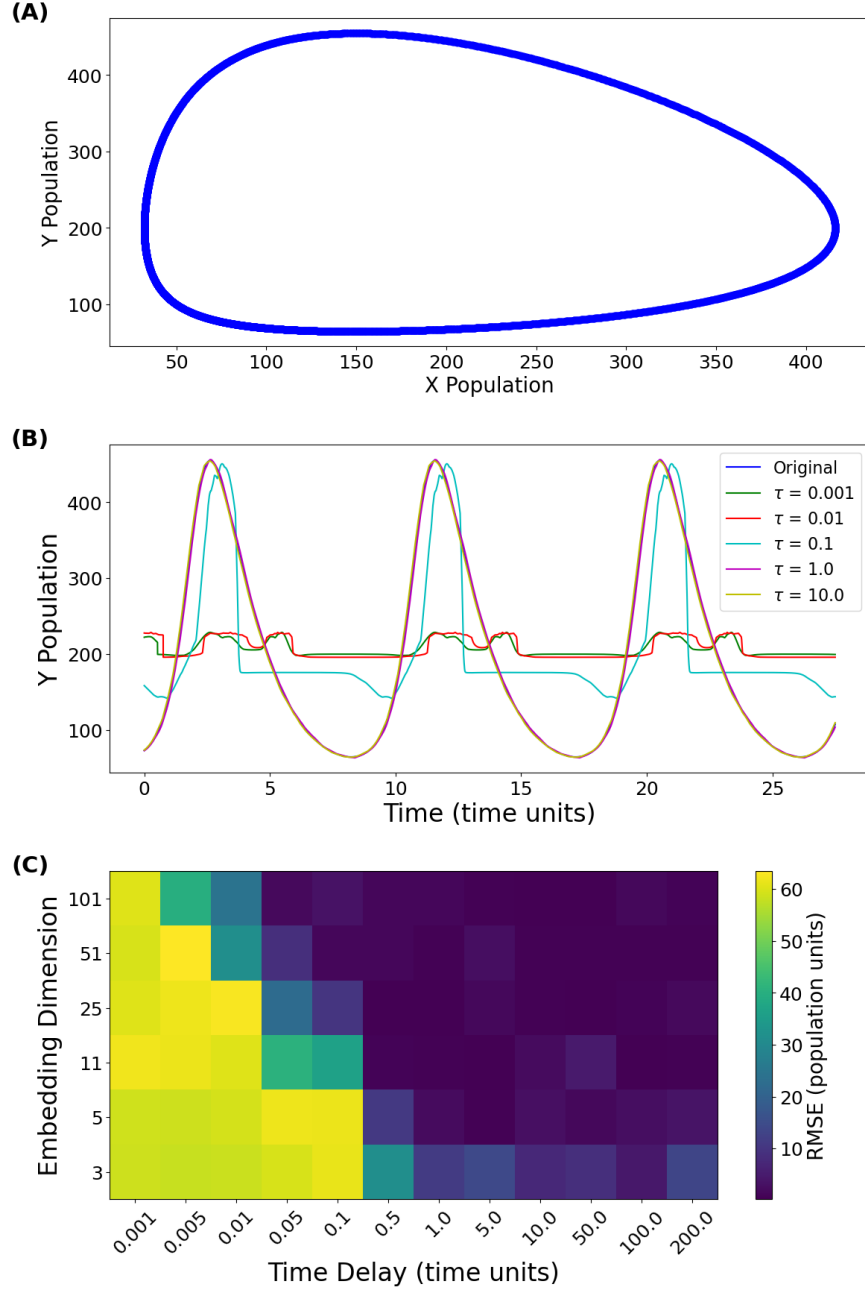


Figure 4.2: Lotka Volterra model with TAR reconstructions using a variety of time delays. **(A)** Phase space diagram of X, Y populations. **(B)** Reconstruction and true populations of species Y over a 25 time unit period from the test set with $\tau = [0.001, 0.01, 0.1, 1, 10]$ with $(2m\tau) = 11$. **(C)** Reconstruction error in phase space by time delay, τ and embedding dimension, $2m + 1$, used to construct delay embedding vectors. We compare these errors to a benchmark of the minimal root mean squared distance point in the training population trajectory compared with all other frames ≈ 68 population units.

4.7 Introducing Takens Informed Reconstruction: the Lotka Volterra Model

As a proof of concept, we introduce TAR with a simple toy example, the Lotka Volterra model Wangersky [1978] modeling the population dynamics of a predator-prey system. Here we show that TAR can reconstruct the full-dimensional system history (both predator and prey populations) from univariate time series using various time delays. We find that reconstruction qualities are dependent on time delay choice, but, at least for simple periodic systems like this one, we can use time delays significantly longer than the characteristic time series of dynamics. Excessively short time delay models are, however, unsuccessfully reconstructed.

The Lotka Volterra equations describe the interactions of a two-species ecosystem of predator and prey through a set of coupled nonlinear first order differential equations;

$$\frac{dx}{dt} = \alpha x - \beta xy; \quad \frac{dy}{dt} = \delta xy - \gamma y \quad (4.3)$$

where x and y are the instantaneous predator and prey populations, and $\alpha = 1$, $\beta = 0.005$, $\delta = 0.8$ and $\gamma = 0.6$. We propagate the Lotka Volterra coupled population dynamics equations for 100 000 time units using a Runge-Kutta 4 Butcher [1964] ordinary differential equation solver with a time step of 0.001 time units yielding 100 000 000 points available for reconstruction. In practice we only use enough points to compute 100 000 delay vectors. The required number of points is given by $100000 + 2m\tau$. The simulation was adapted from Charleux et al. [2024]. All training was performed on a 5 000 observation subset of the first half of the constructed delay vectors while testing was performed on the last half of delay vectors. Further simulation and TAR parameter details are available in Appendix 4.12.2.

We demonstrate the application of the trained Lotka Volterra TAR models constructed with a variety of inputs in Fig. 4.2. A phase space diagram representation of the X, Y populations over the course

of our simulation of this model is described in Fig. 4.2(A). We perform univariate TAR on the time series of the X populations with the goal of predicting Y populations.

In Fig. 4.2(B), we present the resulting predicted Y populations over the course of a sample 25 time unit period from the test set for $\tau = [0.001, 0.01, 0.1, 1, 10]$ with $(2m\tau) = 11$. We choose these time delays to sample many different orders of magnitude of dynamics while $(2m\tau) = 11$ is chosen so that $m > k = 1$, the intrinsic dimensionality of the attractor. Reconstruction errors are determined by minimum root mean squared error (RMSE) between TAR predicted X, Y locations of the test set on the phase space diagram and the correct positions. We get errors on the order of 0.25 – 31.02% with 0.25% error constituting an RMSE of 0.50 ± 0.07 population units for $\tau = 1$ and 31.02% corresponding to the shortest time delay $\tau = 0.001$ with an RMSE of 62.04 ± 1.69 population units. As is evidenced by the red and green lines, this is only slightly better than estimating the mean value of Y across all time points (68.45 units). Models with $\tau = 0.001, 0.01$ are not able to reconstruct dynamics and instead correspond to the average Y population (~ 200 units). $\tau = 0.1$ is nearly sufficient to reconstruct peaks of Y but not the complete oscillation. All tested $\tau > 1$ are capable of reconstructing Y with $\sim 0.5 \pm 0.4\%$ error. The blue, magenta and olive lines are plotted nearly exactly on top of one another as reconstruction errors are quite small. All uncertainties are computed by bootstrapping; we split the test trajectory into thirds, reconstruct the system via TAR and compute the root mean squared errors of reconstruction for each of the test sets.

Fig. 4.2(C) displays reconstruction errors by input time series. We use time delays (τ) of $[0.001, 0.005, 0.01, 0.05, 0.5, 1, 5, 10, 50, 100, 200]$ time units with embedding dimensions $(2m + 1)$ of $[3, 5, 11, 25, 51, 101]$ dimensions. This range allows us to probe histories ranging from far shorter to far longer than the period of oscillation exhibited by the Y population (~ 9.8 time units). The shortest time history is constructed over just $(2m + 1)\tau = 0.003$ time units while the longest spans $(2m + 1)\tau = 20200$ time units. We observe a diagonal pattern demarcating models that can predict values of Y from X and those that cannot. For histories of $2m\tau \sim 2$ time units, TAR models are able

to perform significantly better than guessing the mean value of Y (error ~ 30 population units). For $2m\tau < 1$ time units, TAR models predict Y with accuracies similar to guessing the mean value of Y whereas for $2m\tau \geq 5$ time units, errors fall to ~ 1 population unit or less.

This diagonal trend in the heatmap occurs as longer histories allow us to more assuredly identify where any given time delay vector is on the attractor describing Lotka Volterra dynamics. Once sufficiently long, a time delay vector can be uniquely mapped to its corresponding trajectory in phase space. Consequently, even for incredibly long time delays, i.e. $\tau = 200$ time units, reconstruction errors remain ~ 1 population unit. On the other hand, for very short short time delays, i.e. $\tau = 0.1$ time units, a sufficiently large embedding dimension is enough to overcome the limited view provided by very close observations used to construct the delay vector.

This toy model depicts several notable behaviors transferable to the study of more scientifically relevant systems. (1) A sufficiently long history (e.g. $2m\tau$) is required to uniquely identify where in phase space a given observation is. (2) For periodic non-chaotic systems like this one, reconstruction qualities may not require very high time resolution to reconstruct dynamics. For instance, the best $\tau = 1$ model, $(2m + 1) = 25$, RMSE 33% greater than the best $\tau = 50$ model, $(2m + 1) = 25$, even though a single period is ~ 9.8 time units. (3) Using measurements like the first minimum of the autocorrelation or mutual information functions Fraser and Swinney [1986b], often used to select appropriate time delays, give us suggested time delays on the order of 1-3 time units. In Fig. 4.2(C), we see that these values are not uniquely successful choices. In fact, the best performing model, $\tau = 50$, $(2m + 1) = 25$ has RMSE 0.23 ± 0.02 population units. (4) There is a lower bound limit for the optimal embedding dimensionality that may exceed $(2k + 1)$. The worst performing models for all $\tau \geq 0.5$ were $(2m + 1) = 3$ models even though the intrinsic dimensionality of these systems $k = 1$. For instance, for the $\tau = 50$ models which had the lowest average error across all tested embedding dimensionalities, the $(2m + 1) = 3$ model had an RMSE of 8.78 ± 1.59 , significantly under-performing other choices of m . (5) While longer histories (e.g.

greater embedding dimensionalities) tend to be an asset for limit cycle attractors, this relationship is non-monotonic. For instance, for the $\tau = 5$ model whose time delay is on the order of the period of oscillation of this system, the $(2m + 1) = 11$ model performs with an RMSE nearly half of that of the $(2m + 1) = 101$ model (0.26 ± 0.06 vs 0.50 ± 0.23 population units). We have observed a more dramatic version of this effect in chaotic systems, wherein information creation becomes relevant.

4.8 Experimental Observable Optimization: Molecular Dynamics of a Straight Chain Hydrocarbon

Although Takens' Theorem can be applied to arbitrary observables of system dynamics, they do not produce the same Takens delay vector spaces and consequently latent space embeddings. In fact, even in the limit of a full sampling of accessible phase space, Takens' theorem only guarantees that there exist a diffeomorphism between these spaces. Moreover, in experimental regimes, it can be difficult to make guarantees that a phase space has been "well sampled". It is then unsurprising that empirically, we do not observe the same reconstruction quality for any arbitrary choice of observable. Given many options for observable choice in modeling, we ask, can we predictively identify good observables in advance of modeling efforts?

We hypothesize that for TAR models of chaotic systems, the model accuracy is a function of the degree of chaos exhibited by a given time series. For any deterministic reconstruction problem, TAR seeks to learn a bijective map between each individual trajectory snippet constructed from observations in $\mathbb{R}^1$ to a position in the high dimensional $\mathbb{R}^K$, where K is the full dimension of the system. We rely on the system being sufficiently well sampled that any new delay vector passed to TAR can be mapped to a precise location on the attractor of the dynamics being modeled. Consequently, unique identification of any given trajectory snippet in phase space is paramount. If a system is more chaotic, trajectories will be more sensitive to infinitesimal differences in phase space coordinates. We measure this sensitivity formally with Lyapunov exponents Wilkinson

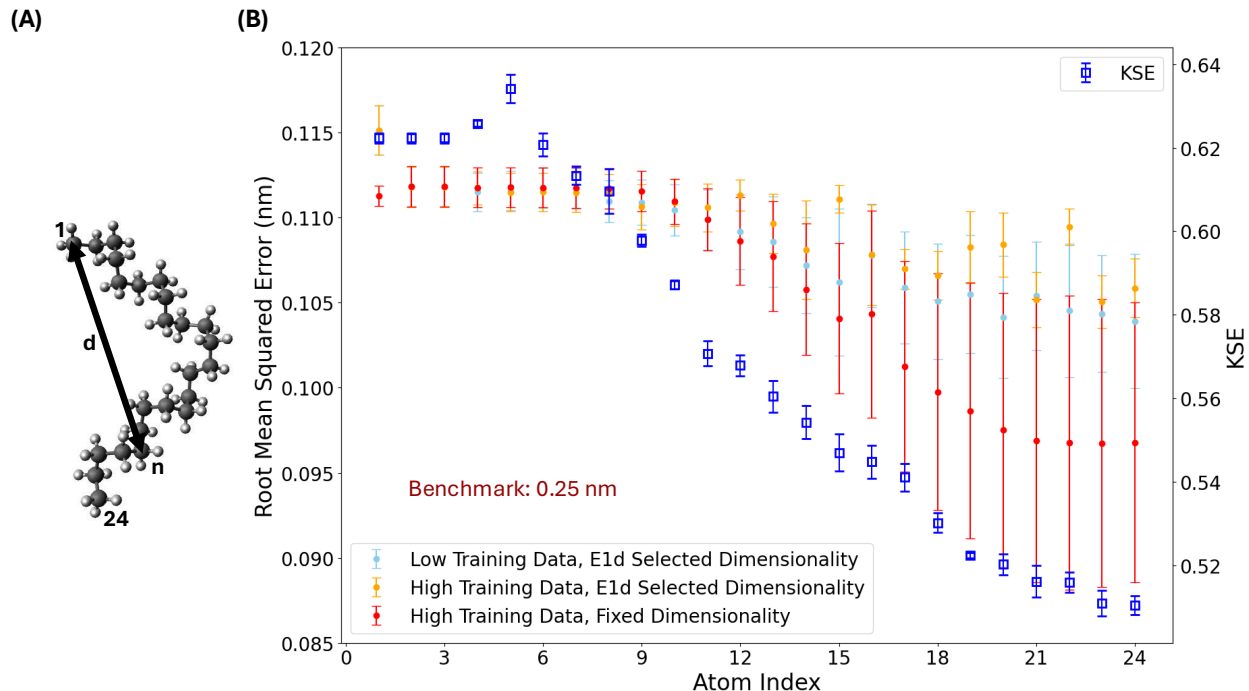


Figure 4.3: Reconstructions of the latent space representation of the straight chain hydrocarbon $C_{24}H_{50}$ as a function of atom index, n . **(A)** Visualization of the $C_{24}H_{50}$ including a label of the 1^{st} , n^{th} and 24^{th} atom in the molecule. We take as our set of available observables the distance between the 1^{st} and n^{th} atom. The rendering was performed using Avogadro Hanwell et al. [2012b] **(B)** Mean reconstruction errors for TAR models trained on time series of 1^{st} and n^{th} atom distance plotted against Kolmogorov-Sinai Entropies of each time series. Reconstruction errors correspond to the test set and can be compared to a benchmark of 0.25 nm corresponding to the minimum RMSD frame against all other frames. We plot 3 types of TAR models; In cyan we have TAR models trained over 1 000 points with delay dimensionality $(2m + 1)$ chosen using the E1(d) method of Cao Cao et al. [1998]. In yellow, we present TAR models trained over 20 000 points with delay dimensionality $(2m + 1)$ chosen using the E1(d) method of Cao Cao et al. [1998]. In red, we present TAR models trained over 20 000 points with delay dimensionality $(2m + 1 = 5)$ corresponding to $m = k = 2$, the intrinsic dimensionality of the attractor.

[2016], which correspond to the time constant corresponding to the exponential divergence in two trajectories started infinitesimally close in phase space. A phase space may be high dimensional and exhibit multiple subspaces whose rate of growth is exponential in time, meaning that there may exist many positive Lyapunov exponents. Under Ruelle’s Inequality Ruelle [1978], we have that the upper bound limit of the Kolmogorov-Sinai Entropy of a chaotic system, or rate of information creation, is given by the sum of positive Lyapunov exponents of a system. We then propose to use this Kolmogorov-Sinai Entropy (KSE) as a means of selecting minimally chaotic observations from time series. A more mathematical treatment of this hypothesis including a proof that topographical deformation of manifolds grows as a function of KSE is provided in Chapter 5.

In this section, we test this hypothesis with the example of the straight chain hydrocarbon $C_{24}H_{50}$ in water which, in principle, can be observed experimentally using the single molecule fluorescence resonance energy transfer (smFRET) technique and whose dynamics can be simulated using molecular dynamics Roy et al. [2008], van der Spoel et al. [2005]. In smFRET, a target molecule is tagged with two or more dyes whose fluorescence is a deterministic function of the distance between each dye. As such, an intramolecular distance between two or more chosen atoms on a molecule can be estimated up to an experimentally determined time resolution. Using a ground truth molecular dynamics simulation of $C_{24}H_{50}$, for which we know the positions of every atom in the molecule, we extract a synthetic smFRET signal.

By training TAR models on time series from various dye placement points, we learn a latent space phase space representation with TAR and backmapping to the all atom representation of the molecule and we empirically investigate the optimal point of dye placement. We hypothesize that as the distance between atoms on the hydrocarbon backbone increase, time series should better reflect folding dynamics. This connects to the Kolmogorov-Sinai entropies as follows; both time series of 1st to 24th atom and 1st to 4th atom distances embed information sufficient to reconstruct an attractor describing the dynamics of this molecule up to a smooth mapping according to Takens’

Theorem. However, the former has a mean value of 1.647 nm with a standard deviation of 0.54 nm and the latter has a mean value of 0.368 nm with a standard deviation of 0.038 nm. This means that the same dynamics information must be embedded in the 1st to 4th atom time series using changes of far smaller magnitudes.

We use a 100 ns molecular dynamics simulation of C₂₄H₅₀ with system configurations saved every 0.2 ps for a total of 500001 frames. We compute intramolecular distances between the head carbon (atom 1) and all other atoms in the molecule for a total of 23 time series. All TAR models are trained over some subset of the first half of the simulation trajectories and reconstruction accuracy is tested over the second half. Results are benchmarked against the lowest RMSE frame in the training trajectory relative to all other frames in the trajectory. The details of this simulation are available in the supplemental information of Topel and Ferguson [2020]. Further details of the TAR parameters can be found in supplemental materials Section 4.12.3.

We use TAR models to reconstruct phase space representations of the univariate time series describing the distances between atoms 1 and all other atoms in C₂₄H₅₀. Fig. 4.3(A) displays the C₂₄H₅₀ labeled with the head atom (1), tail atom (24) and n^{th} atom where the distance d used to construct Takens delay vectors is taken as the distance between atoms 1 and n . Note that this molecule is symmetric so our choice of head atom and tail atom makes no difference. In Fig. 4.3(B) We measure TAR model performance by the root mean squared distance or reconstruction error (RMSE) between each atom in each reconstructed test trajectory with the corresponding frame in the MD simulation over every frame. In cyan we have TAR models trained over 1 000 points with delay dimensionality $(2m + 1)$ chosen using the E1(d) method of Cao Cao et al. [1998]. In yellow, we present TAR models trained over 20 000 points with delay dimensionality $(2m + 1)$ chosen using the E1(d) method of Cao. In red, we present TAR models trained over 20 000 points with delay dimensionality $(2m + 1 = 5)$ corresponding to $m = k = 2$, the intrinsic dimensionality of the attractor. We compare reconstruction qualities to a benchmark error of 0.25 nm corresponding to

the RMSE of the lowest RMSE configuration against all other configurations in the trajectory. On the same plot we include KSEs measured for each time series of the 1st to n^{th} atom estimated using Ruelle’s inequality with Lyapunov exponents computed via the nolds package Schölzel [2020] using Eckmann’s approach Eckmann et al. [1986].

This figure demonstrates meaningful differences in reconstruction quality as a function of observable choice and predictability of those differences via dynamical systems measurements. The high sensitivity of short time series dynamics leading to higher chaos is reflected by the KSEs with a practical consequence of increased RMSEs. The effect of that sensitivity on reconstruction quality is, however, also a function of how we construct TAR. In 4.3(B), we observe a largely monotonic relationship between increased atom index and RMSE, however, there is variation with model choice. In red, we have TAR models trained over 20 000 points with delay dimensionality ($2m + 1 = 5$) corresponding to $m = k = 2$, the intrinsic dimensionality of the attractor. For these observations we observe the lowest RMSEs but highest uncertainties of the tested models. This may be attributed to the fact that we selected $m = k$, so the embedding dimensionalities accurately reflect the intrinsic dimensionality of the attractor. While larger embedding dimensionalities were perhaps not harmful for the previous example of a limit cycle attractor, for chaotic system this is tantamount to adding “deterministic noise” Eckmann and Ruelle [1985]. These have the most monotonic relationship between RMSE and KSE with a Spearman rank correlation coefficient of Spearman: 0.97 ($p=3.31e-15$) indicating a strongly monotonic relationship. Larger uncertainties are, however, likely attributable to these shorter time delay vectors than those chosen with the E1(d) method of Cao.

The RMSEs plotted in yellow corresponding to TAR models trained over 20 000 points with delay dimensionalities chosen by the E1(d) method have a reduced Spearman rank correlation coefficient of 0.91 ($p=1.02e-09$). This reduction correlation coefficient likely stems from much longer delay vectors for smaller values of atom index n . For instance, value of $(2m + 1)$ chosen for $n = 1$ is 43

compared to 19 for $n = 24$. It follows that more chaotic systems require longer histories (i.e. larger m) in order to uniquely specify a path in phase space than less chaotic ones. These larger m values are, however, insufficient to entirely eliminate the advantage of lower chaos in this case.

Another way of limiting the availability of histories is to reduce training data volume. In cyan, we present models with 1 000 points with delay dimensionality $(2m + 1)$ chosen using the E1(d) method of Cao. We measure a Spearman coefficient of 0.96 ($p=2.40e-13$), indicating a strongly monotonic trend. We understand this larger Spearman coefficient as stemming from limited training data volume. Less chaotic systems ought to be more predictable and consequently, while all of these systems are deterministic, predictions using minimally chaotic time series correspond to lower RMSEs.

This computational experiment provides practical insights for both experimentalists and computationalists analyzing time series data. Suboptimal choice of time series increases measured chaos, reducing system predictability. Here we have shown that for real systems like $C_{24}H_{50}$, KSEs and leading Lyapunov exponents optimize observable selection for TAR models, maximizing the relevant information captured by a univariate time series.

4.9 Multi-Temporal Takens' Theorem: Protein Dynamics and Folding

Dynamical systems often admit the simultaneous observations of multiple observables, allowing for the multiplexing of information flows. These extra information flows may provide new information, however, they may also come at a cost. In the example of smFRET, gathering more intramolecular distances requires tagging a molecule with more fluorophores, thus further altering molecular dynamics Reinartz et al. [2020]. Given that Takens' Theorem tells us that a univariate observable stream is sufficient to reconstruct dynamics under modest technical conditions, the utility of a new and potentially experimentally costly observable may not improve reconstruction quality.

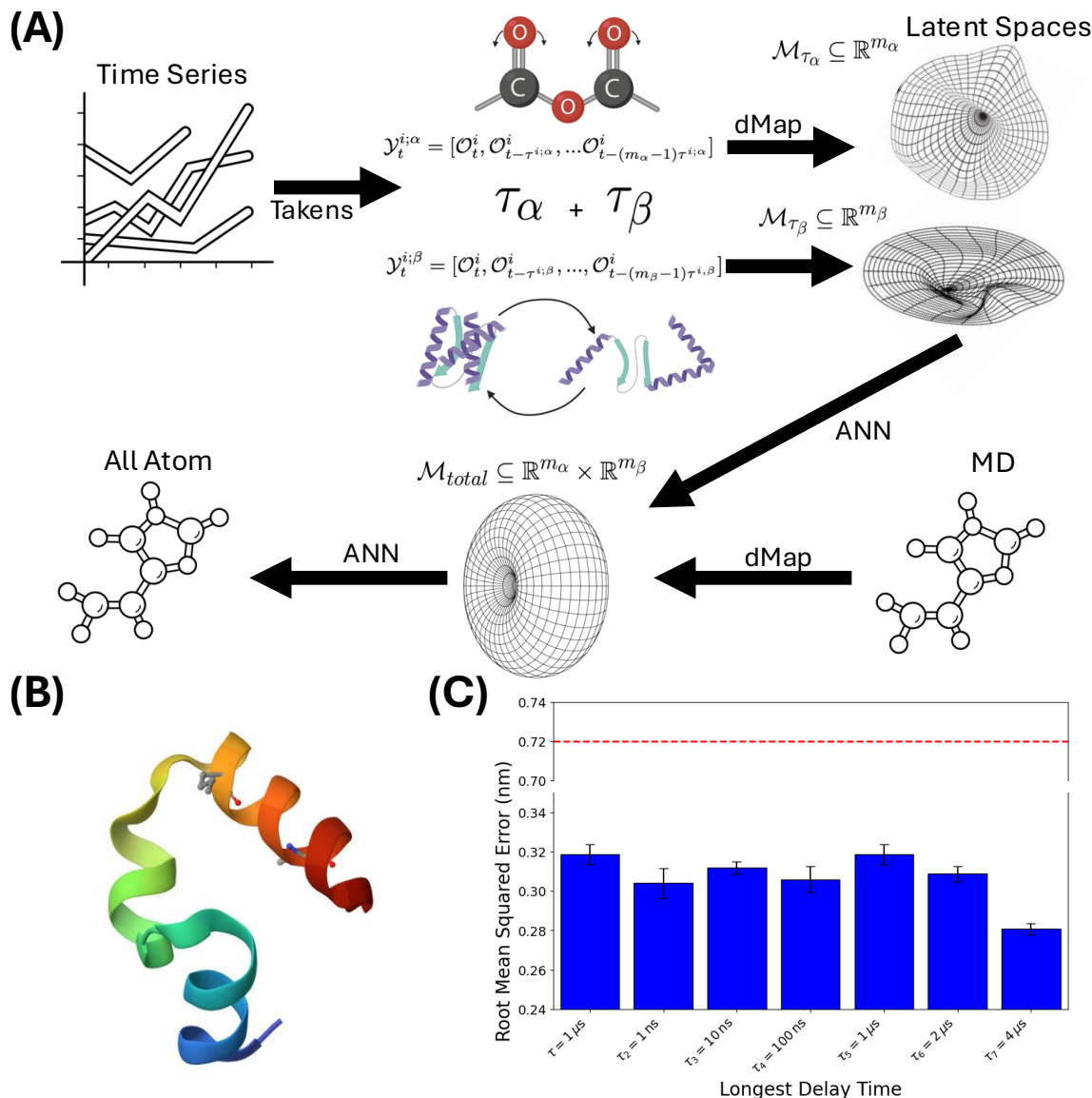


Figure 4.4: Schematic of multi-temporal TAR applied to the protein Villin with reconstruction errors corresponding to delay time choice. **(A)** The multi-temporal TAR pipeline. Beginning with a time series, we construct p sets of time delay vectors with $\tau_1, \tau_2 \dots \tau_p$ as the time delays and construct p latent space manifolds. We learn the map from $\mathcal{M}_1, \mathcal{M}_2 \dots \mathcal{M}_p$ to $\mathcal{M}_{total}$ via an ANN and reconstruct the full-dimensional molecule as in univariate TAR. **(B)** is a VMD rendering of the protein Villin (PDB:2F4K) Humphrey et al. [1996]. **(C)** describes mean reconstruction errors of the full-dimensional atomic trajectory over the course of the test set for each of the tested observables with benchmark line in red corresponding to the minimum RMSD frame against all other frames. Uncertainties are bootstrapped by dividing the delay vectors and trajectory in to 3 equal length portions, training TAR on a subset of each and testing on the others. On the far left we have a univariate model with $\tau = 1\mu s$ chosen from the first minimum of autocorrelation. All other models are multi-temporal including $\tau_1 = 200ps$ and all τ up to τ_p listed (i.e. $\tau = 3$ corresponds to a multi-temporal model using τ_1, τ_2 and τ_3).

In this section, we introduce a multi-temporal approach to Takens’ Theorem in which manifolds constructed using multiple time delays can offer a more complete description of dynamics than single or multivariate approaches alone. The reason for this multi-temporal approach stems from the separability of dynamics over time. We take as an example the molecular dynamics of the Villin protein. The folding of this protein occurs in on $\mathcal{O}(\mu\text{s})$ Kubelka et al. [2003], Henzler-Wildman and Kern [2007], rotameric dynamics occur on $\mathcal{O}(\text{ps-ns})$ Haddad et al. [2019], Henzler-Wildman and Kern [2007], local thermal fluctuations occur on $\mathcal{O}(\text{ps-ns})$ Kurplus and McCammon [1983], Henzler-Wildman and Kern [2007] and long range interactions may occur on $\mathcal{O}(\mu\text{s-ms})$ Henzler-Wildman and Kern [2007]. Single time delays offer insight on any one of these given phenomena, however, it would take extraordinarily high dimensional Takens Delay vectors to span dynamics from ps to ms ($\mathcal{O}(10^9)$ dimensions!). We recognize that practically, information cascades allow for the reconstruction of dynamics from across timescales. Weak nonlinear coupling, feedback loops, multiscale interactions, stochastic resonance and noise-driven cross-scale interactions can result in these multiscale information cascades, in some cases allowing for reconstruction without such a multi-temporal approach Haken [1983], Brummitt et al. [2012], Easley and Kleinberg [2010], Scheffer et al. [2009].

We hypothesize that combining time delay vectors from the same time series, but with different τ corresponding to distinct characteristic times, can embed multi-scale dynamics without causing an explosion in dimensionality. To test this, we merge latent space manifolds extracted from TAR across multiple timescales and evaluate reconstruction quality. As dynamics are separable by timescale, the manifolds constructed at different τ should capture independent components of the system’s behavior Kevrekidis et al. [2005]. We first construct a latent space for each timescale using univariate TAR, then learn a mapping from the set of all leading eigenvectors of these manifolds to the latent space of the ground truth molecular dynamics simulation. This approach enables recovery of the correct manifold dimensionality before backmapping to an all-atom representation.

In this section, we make use of a 125.6 μs MD simulation of the protein Villin produced by D. E. Shaw research Lindorff-Larsen et al. [2011] as our ground truth model. Frames were saved every 2 ps, producing 627 907 frames. All Takens delay vectors produced are created from synthetic time series produced by measuring intramolecular distances between two heavy atoms in the Villin protein. Simulation details are available in Lindorff-Larsen et al. [2011]. Model parameters for TAR are available in supplemental materials section 4.12.4.

Figure 4.4 provides an overview of the multi-temporal TAR reconstruction process and its application to the protein Villin to reconstruct its full-dimensional configurations from a synthetic smFRET time series generated from an MD simulation. In Fig. 4.4(A) we provide a schematic of the multi-temporal TAR setup. Beginning with a time series, we construct p sets of time delay vectors with $\tau_1, \tau_2 \dots \tau_p$ as the time delays and construct p latent space manifolds. We also learn a latent space representation of the molecular dynamics trajectory of Villin, $\mathcal{M}_{total}$. Next, we learn the map from $\mathcal{M}_1, \mathcal{M}_2 \dots \mathcal{M}_p$ to $\mathcal{M}_{total}$ via an ANN. Finally, we map from the reconstructed $\Theta(\mathcal{M}_{total})$ to the full-dimensional molecule. Fig. 4.4(B) is a VMD rendering of the protein Villin (PDB:2F4K) Humphrey et al. [1996]. Fig. 4.4(C) describes mean reconstruction errors of the full-dimensional atomic trajectory over the course of the test set for each of the tested observables with benchmark line in red corresponding to the minimum RMSD frame against all other frames. Uncertainties are bootstrapped by dividing the delay vectors and trajectory in to 3 equal length portions, training TAR on a subset of each and testing on the others. On the far left we have a univariate model with $\tau = 1\mu\text{s}$ chosen from the first minimum of autocorrelation. All other models are multi-temporal including $\tau_1 = 200\text{ps}$ and all τ up to τ_p listed (i.e. $\tau = 3$ corresponds to a multi-temporal model using τ_1, τ_2 and τ_3).

Fig. 4.4(C) demonstrates that all multi-temporal and univariate Takens’ approach perform with RMSEs that are a fraction of the benchmark RMSE ($\sim 0.3\text{ nm}$ vs 0.72 nm). The univariate reconstruction performed with $\tau = 1\mu\text{s}$, a time delay approximately equal to the first minimum of

autocorrelation of the distance time series, performs with an RMSE of 0.318 ± 0.005 nm. Starting from the shortest time delays available (200 ps), we construct multi-temporal Takens' vectors using time delays on successively larger orders of magnitude. The first model uses $\tau_1 = 200$ ps, $\tau_2 = 1$ ns, the second uses $\tau_1 = 200$ ps, $\tau_2 = 1$ ns, $\tau_3 = 10$ ns and so on. The best performing model is the final one with longest time delay $\tau_7 = 4\mu$ s with an RMSE of 0.274 ± 0.003 nm. This Villin protein has a characteristic folding time of 2.8μ s Lindorff-Larsen et al. [2011] so while the last 3 models using $\tau_5 = 1\mu$ s, $\tau_6 = 2\mu$ s, $\tau_7 = 4\mu$ s all use time delays on the characteristic folding time of Villin, the performance improvement only comes with the last model.

Both the univariate and multivariate models effectively capture the protein's folding dynamics, which dominate the system's behavior, especially at longer timescales and with full head-to-tail distances. While empirically we find that there is sufficient information at short time scales to reconstruct Villin with RMSEs characteristic of static methods such as cryo-electron microscopy and x-ray crystallography, improvement is achieved without the introduction of new data streams through multi-temporal Takens delay vectors. The lack of monotonic improvement with multi-temporal approaches demonstrates that timescale choice can be important to improve these models. However, while physical intuition such as characteristic folding times can be used to intuit good choices for τ_p , we are not able to determine these ideal values precisely. The multi-temporal approach offers a more nuanced reconstruction by integrating information across multiple timescales, thereby capturing dynamics at these various scales. Importantly, this method does not require an increase in the number of observables, making it a cost-effective yet robust platform for multiscale modeling provided there is a limited information cascade between timescales.

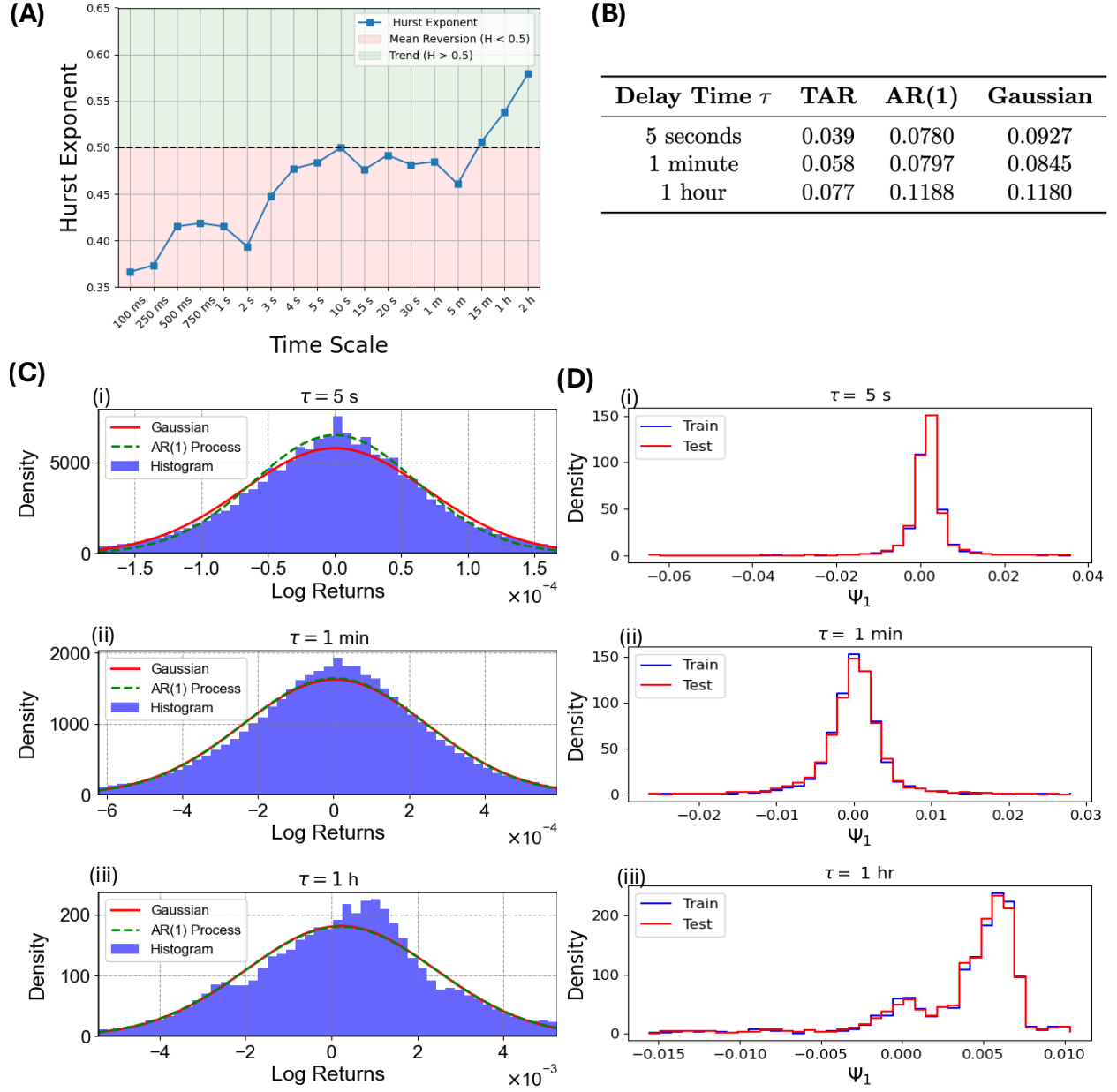


Figure 4.5: Vanguard S&P 500 ETF (VOO) returns from January 2024 Services [2024] for 5 second, 1 minute and 1 hour time delays returns. (A) measures hurst exponents over a 100 ms to 2 hour time period. (B) relays the L1 norm between the true returns distributions of VOO and (i) TAR models (ii) AR(1) (autoregressive) model (iii) Gaussian distribution. (C) Histogram of returns over $\tau = 5$ s, $\tau = 1$ min and $\tau = 1$ h time periods with both AR(1) and Gaussian fits using **all** available data. (D) histogram of returns in TAR produced phase space for both a train and test set with training performed on $\min[5000, \text{len}(\text{delay_vectors})//2]$.

4.10 Phase Space Reconstruction in the absence of training data: Vanguard S&P 500 ETF

Many experiments describe phenomena that are beyond the capacity of simulation, either because the rule sets describing them are unknown or are excessively expensive. For instance, proteins folding on time scales on the order of seconds are many orders of magnitude in timescale beyond the accessibility of atomistic simulations even on the largest super computers. Conversely, many social scientific and biological studies do not have clear physical rules that we can write to propagate these systems. However, Takens' Theorem guarantees that univariate time series can be structured in to delay vectors constructing a manifold that is diffeomorphically related to the manifold embedding the true full-dimensional dynamics. This means that the manifolds we extract with TAR in the absence of a training ground truth simulation can still give us phase spaces that embed the system dynamics. TAR manifolds can then represent them on a landscape that is bijectively related to some objective function or time series we wish to reconstruct Takens [1981]. This is sufficient to gain insight on phenomenologies driving dynamics and construct phase space descriptions of arbitrary dynamics.

In this section, we take one such experimental time series, the execution price of shares of the Vanguard S&P 500 Exchange Traded Fund (VOO) over the month of January 2024 and use TAR models to learn the phase space of returns over short time delays. It is expected that at short timescales, the price of liquid assets such as VOO will follow a random walk as traders have similar information and consequently have no ability to make decisions based on material changes in market conditions Fama [1995]. We model the returns attributable to these random walks as Gaussian processes.

$$f(x) = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(x - \hat{\mu})^2}{2\hat{\sigma}^2}\right) \quad (4.4)$$

where x are the observed returns, $\hat{\mu}$ is the estimated mean return and $\hat{\sigma}$ is the estimated standard deviation.

At very short time scales are observed to be mean reverting as a result of bid-ask bounce, liquidity effects and associated transaction costs, order-book dynamics and inventory management by market makers Madhavan [2000], Bouchaud et al. [2018], Campbell et al. [1998], O'Hara [1995]. As prices are discrete (they correspond to tick values), we use an AR(1) process to model returns:

$$x_t = \phi x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \quad (4.5)$$

$$f(X_t | x_{t-1}) = \frac{1}{\sqrt{2\pi \hat{\sigma}_\varepsilon^2}} \exp\left(-\frac{(x_t - \hat{\phi} x_{t-1})^2}{2\hat{\sigma}_\varepsilon^2}\right) \quad (4.6)$$

where ϕ is the autoregressive coefficient, determining the memory or persistence of the process, σ_ε is the standard deviation estimated from the noise term, x_t is the return at time t and ε_t is the Gaussian noise at time t . As timescale increases and processes switch from mean reverting to random walks, we expect that the value of ϕ goes to zero. At long time scales changes to economic conditions can shift equilibria and we see the emergence of trend. In this section, we show that we can use TAR to estimate probability distributions of returns for all three situations with L1 norms better than the theoretical models for systems evolving at each of these time scales.

We first construct Takens delay embedding vectors over our time series, for $\tau = [5 \text{ s}, 1 \text{ min}, 1 \text{ h}]$. We elect to use these values as they correspond to weakly mean reverting, random walk and trend following regimes and consequently are described well by either the AR(1) process, the Gaussian or neither. While prices are discrete in the order book, times are not. We avoid smaller time bins to ensure that there are multiple orders per time bin for every time in our time series, eliminating empty bins. Similarly, as there are only 6.5 trading hours per day and we need multidimensional Takens delay vectors to describe dynamics, we limit reconstructions to a maximum time delay of 1 hour to

avoid having time delays involving interday observations. There are methods of approaching both of these issues; last observation carried forward approaches, use of midpoint quotes and skipping empty bins can be used to deal with empty bins Bouchaud et al. [2018], O’Hara [1995]. Modeling overnight gaps explicitly or measuring time based on cumulative trading activity Hansen and Lunde [2006], Clark [1973] can smooth over interday effects. We choose to limit our study of time ranges to avoid the necessity of more sophisticated financial modeling techniques.

We use time delays of 3 time units as we expect $k = m = 1$ for these simple processes that are well described by Gaussian and AR(1) models. We then compute the log return over the each element of each delay vector, dividing each element by its first component. Log returns are used in securities price modeling as they better approximate a normal distribution, are additive in time and are more stationary than standard returns Campbell et al. [1998]. We then scale returns using a MinMax scaler bound by $[-1, 1]$ Pedregosa et al. [2018] and perform diffusion maps to get our latent space manifold construction of the returns landscape. Gaussian kernel width is adjusted recursively to match the implied dimensionality of diffusion maps extracted latent spaces with the embedding dimensionalities computed with the E1(d) method Cao [1997]. We train all TAR models over $\min[5000, \text{len}(\text{delay_vectors})//2]$ chosen randomly from the first half of all constructed delay vectors, ensuring that we never use more than half of available delay vectors to train our data. We then project the remain delay vectors down to the latent space extracted from real data via Nyström extension Williams and Seeger [2000]. We asses model quality by the L1 norm between the train and test probability density functions in latent space. These are compared with the L1 norms of the Gaussian and AR(1) fits to the log returns distributions. We use L1 norms as they correspond to total variation distance between distributions. These are constructed over all of the data available.

In Fig 4.5(A), we measure Hurst exponents over a 100 ms to 2 hour time period Hurst [1956, 1957], Weron [2002]. The green range represents trend following processes, while the red range is mean reverting. Hurst exponents near 0.5 correspond to random walks. We observe that on short time

scales (order of ms to s), we observe Hurst exponents indicating mean reversion while at long time scales (order of hours) and random walks occur at middle time scales (order of 10s of seconds to minutes). In Fig 4.5(B), we present the L1 norms for all tested time ranges and models. We observe that on all time ranges, TAR outperforms the AR(1) and Gaussian processes. On the 5 second timescale, the AR(1) process outperforms the Gaussian whereas on the 1 minute timescale, they perform at a similar level, indicating that ϕ takes a small value. On the 1 hour time scale, neither is capable of reconstructing the skewness in the returns distribution, as mean log returns for VOO over January 2024 are positive. In all three cases, TAR outperforms the standard models. Fig 4.5(C) displays histogram of returns over $\tau = 5$ s, $\tau = 1$ min and $\tau = 1$ h time periods with both AR(1) and Gaussian fits using all available data. We see that at short time scales, the processes are weakly mean reverting, become near random walks and eventually become trend following reflected by the skewness in subpanel (iii). Fig. 4.5(D) reflects the histogram of returns in TAR produced phase space for both a train and test set with training performed on $\min[5000, \text{len}(\text{delay_vectors})//2]$. In all 3 cases, the phase space of returns can be described in a single dimension, Ψ_1 .

We conclude that we can observe the phase space corresponding to any given time series regardless of whether or not we have a ground truth test set. TAR is agnostic to probability distribution and is able to reconstruct dynamics for all 3 tested time periods. While, future TAR models may incorporate knowledge of these distributions in the loss function constructing latent spaces, we find that for this VOO securities price time series, TAR describes the returns spectrum well.

4.11 Conclusions and Future Work

This work presents a highly interoperable framework, TAR, that integrates dynamical systems theory with machine learning models to optimize time series analysis in a physically robust manner. We introduce a pipeline that extracts phase space representations of time series data in a low-dimensional format while preserving high-dimensional dynamics through Takens' Theorem and

manifold learning. Additionally, we provide a method for mapping these representations back to their full-dimensional descriptions when a training trajectory is available. We show that TAR can be used to reconstruct arbitrary dynamical systems, starting with Lotka Volterra. Furthermore, we demonstrate that Kolmogorov-Sinai Entropies can be used to predicatively select time series inputs to optimize TAR models using molecular simulations of $C_{24}H_{50}$. With Villin, we showed that multi-temporal TAR can access dynamics across timescales and integrate the information to improve upon univariate TAR constructions using the same time series. Finally, with the security VOO, we showed that TAR can learn phase space representations of time series that accurately embed system dynamics without a ground truth full-dimensional model for training. All of the components of these pipelines including univariate, multivariate and multi-temporal TAR, nonlinear mapping and dimensionality reduction tool set developed for this paper are available on github at <https://github.com/maxtopel/TAR>.

While the TAR package can address a wide variety of dynamical systems, we seek to continue improving its accuracy. In this work we proposed the use of KSEs to predict optimal inputs for chaotic attractors. In future work we seek to expand our use of dynamics measures to characterize the utility of a given time series for learning. Like most nonlinear dimensionality reduction platforms, TAR does not provide an easily interpretable basis for the dynamical system. In the future, we seek to construct mappings from proposed bases of systems to the latent spaces we construct to understand better the physical nature of our collective variables. We also seek to expand TAR’s capabilities to apply holonomic constraints such as physical loss functions to increase the accuracy of these reconstructions. These two innovations can increase the interpretability of TAR models. The TAR pipeline promises to provide a window in to the dynamics of arbitrary time series observables. Given the wide diversity of system we seek to study with TAR, we aim to expand the types of intermanifold mapping, dimensionality reduction and data structuring methods available and investigate the effect of model choice on the quality of Takens-informed reconstructions.

4.12 Supporting Information

4.12.1 *Information Theoretic Measurements for Determining Time Delay of Embeddings*

Determining the optimal delay time for Takens Delay vectors is directly related to the characteristic time of dynamics we wish to reconstruct. Consequently, we wish to understand over which time periods these dynamics occur. Mutual information and auto-correlation allow us to make estimates about these time periods. In this section we discuss the calculation of mutual information.

Shannon Entropy: The Shannon Entropy is a measure of the average amount of information required to differentiate states in a process Shannon [1948]. Let $p(\mathbf{x})$ be the probability that our system occupies state $\mathbf{x} \in \mathcal{X}$. Then the average amount of information, $H(\mathbf{x})$ required to describe the state of $\mathbf{x}$ is the expectation value of the probability of occupying any given $\mathbf{x}$ times the log likelihood of occupying state $\mathbf{x}$:

$$H(\mathbf{x}) = \sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \log_2(p(\mathbf{x}))$$

$\mathbf{x}$ can describe any given state of our dynamical system as conveyed by our time series observables, the true full-dimensional data set, or low dimensional representations of either data set produced using manifold learning techniques such as diffusion maps.

Mutual Information: Mutual Information describes how much information time series X contains about time series Y and can be defined in terms of Shannon entropies Fraser and Swinney [1986b]:

$$I(X;Y) = H(X,Y) - H(X|Y) - H(Y|X)$$

This can be used to estimate how capable X is of predicting Y . We can also compute a time series' mutual information with a delayed version of itself to determine how far in to the future we have to look such that observable $X(t)$ and $X(t + \tau)$ share no mutual information. We use this criterion to determine optimal time delays for Takens delay vectors Fraser and Swinney [1986b].

4.12.2 2D Lotka Volterra Predator Prey Dynamics

The 2D Lotka Volterra (LV) model simulates the interaction of predator and prey species through a set of coupled differential equations.

$$\frac{dx}{dt} = \alpha x - \beta xy; \quad \frac{dy}{dt} = \delta xy - \gamma y \quad (4.7)$$

We propagate the Lotka Volterra coupled population dynamics equations for 100 000 time units with a time step of 0.001 time units and α, β, ϵ and γ of 1, 0.005, 0.8 and 0.6 respectively. The solved differential equations give us 100 000 000 observations of the populations and population changes of two species which we denote as x, y and $\frac{dx}{dt}, \frac{dy}{dt}$ respectively.

All reconstructions of Lotka Volterra models are trained on univariate time series of the population of species X, x . For each time delay, we construct 100 000 delay vectors. Each model uses 5 000 delay vectors subsampled uniformly from 1/3 of the delay vectors.

We reduce the Takens delay vectors via diffusion maps Coifman et al. [2005a] to recover a 2 dimensional intrinsic manifold using the diffusion maps method trained on the same 5 000 point training set. We take the Euclidean distance between time delay vectors as the similarity metric used to construct the pairwise distance matrix. We tune the Gaussian kernel width, ϵ , for each model such that the implied dimensionality of the resulting manifold matches the 2 dimensional full system model. Implied dimensionality is computed using the L method of Salvador and Chan Salvador and Chan [2004]. We recognize that the limit cycle attractor is itself 1 dimensional but

keep the 2D representation as the full-dimensional system we map to, $[X,Y]$ populations, is 2D. The remaining data points not included in training were Nyström extended Williams and Seeger [2000] to project them on to the extracted manifold.

Finally, we backmap from our latent space manifold to the 2 dimensional population time series X,Y using a neural network trained on the training 1/3rd of the trajectory and tested on the remaining 2 thirds. Our network has input dimension and output dimensions of 2, respecting the Takens determined diffeomorphic mapping between the Takens space and the full-dimensional dynamics space. The network has five hidden layers, each containing 50 nodes. Each hidden layer is a linear with a ReLU activation function Agarap [2019] to allow for nonlinear behavior. The training process consists of 500 epochs, using the Adam optimizer Kingma and Ba [2014] with a learning rate of 0.01. All data is scaled using MinMaxScaler Pedregosa et al. [2018] between $[0,1]$.

4.12.3 $C_{24}H_{50}$ molecular dynamics and synthetic smFRET

Simulation details

A 100 ns MD simulation of the straight chain hydrocarbon $C_{24}H_{50}$ was conducted to provide a full-dimensional trajectory to train and test the TAR model. The trajectory contains frames trained every 0.2 ps, yielding 500 001 frames. The details of this simulation are available in the supplemental information of Topel and Ferguson [2020].

All reconstructions of $C_{24}H_{50}$ models are trained on univariate time series observation of intramolecular distances evolving over time. These time series correspond to the distance between the head (1^{st}) and n^{th} carbon along the backbone of the straight chain polymer $C_{24}H_{50}$. Takens delay vectors are either taken as 5 dimensional so that $m = k = 2$ matches the intrinsic dimensionality of the system attractor or are computed via an E1(d) analysis of each of the intramolecular distance trajectory Cao [1997] with a cutoff of 0.98 depending on the model. Optimal delay times were extracted using the first minimum of the autocorrelation function for each time series. All data is

scaled using MinMaxScaler Pedregosa et al. [2018] between [0,1].

We reduce the Takens delay vectors to a 2 dimensional intrinsic manifold using the diffusion maps Coifman et al. [2005a] method trained on a 5 000 or 20 000 point training set selected. These training points are subsampled from 1/3 of the available delay vectors with test data coming from the remaining 2/3. A simple Euclidean distance between time delay vectors is taken to be the distance metric used to construct the pairwise distance matrix. We tune the Gaussian kernel width ϵ for each model such that the implied dimensionality of the resulting manifold matches the 2 dimensional full system model. Implied dimensionality is computed using the L method of Salvador and Chan Salvador and Chan [2004]. All remaining data points were Nyström extended Williams and Seeger [2000] extended on to the extracted manifold. We then map from our latent space manifold extracted by diffusion maps to a manifold extracted by diffusion maps run over a matrix constructed from the Euclidean distances between each frame of the trajectory, learned over the first half of all constructed delay vectors. This latent space manifold, $\mathcal{M}'$, embeds within it the dynamics displayed by the time delay vectors used to construct it.

We reconstruct a latent space representation of the $\text{C}_{24}\text{H}_{50}$ molecular dynamics directly from the full-dimensional data. Constructing a pairwise distance matrix over the all atom representations of the frames corresponding to the training points used to construct the latent space manifold from time series. These pairwise distances correspond to the rotationally and translationally aligned root mean squared error (RMSE) of each atomic position Kabsch [1976] with minimization of RMSE about mirror symmetries arising from the uniformity of this straight chain hydrocarbon and those due to head to tail inversion. Details of this process are available in Topel and Ferguson [2020]. We use the same Gaussian kernel width estimation and Nyström extension method to recover the latent space manifold $\mathcal{M}$ that embeds the dynamics displayed by the $\text{C}_{24}\text{H}_{50}$ simulation. Following the L method of Salvador and Chan Salvador and Chan [2004], we determine that this manifold is intrinsically 2 dimensional.

To recover the diffeomorphic mapping between $\mathcal{M}'$ and $\mathcal{M}$, we use a neural network trained on the first half of delay vectors constructed from the time series and tested on the last half. Our network has input and output dimensions of 2 with five hidden layers, each containing 50 nodes. Each hidden layer is a linear with a ReLU activation function Agarap [2019] to allow for nonlinear behavior. The training process consists of 500 epochs, using the Adam optimizer Kingma and Ba [2014] with a learning rate of 0.01. We call the reconstruction of $\mathcal{M}$ produced by this last mapping from $\mathcal{M}'$, $\Theta(\mathcal{M}')$.

To recover the full-dimensional coordinates of each of the 24 heavy atoms of $\text{C}_{24}\text{H}_{50}$, we employ a second network, also with five hidden layers, each containing 100 nodes using the ReLU activation function. Models are trained over 500 epochs, using the Adam optimizer Kingma and Ba [2014] with a learning rate of 0.01 on the first half of points corresponding to the same training set at the previous step. This network takes in the 4 coordinates of each point on $\Theta(\mathcal{M}')$ and outputs the 3×24 coordinates corresponding to the x,y,z positions of each heavy atom of $\text{C}_{24}\text{H}_{50}$. This process is repeated twice more using the other 2/3 of the data to bootstrap uncertainty estimates.

Calculation of Lyapunov Exponents and Kolmogorov-Sinai Entropies

All Lyapunov exponent are computed using the nolds package which implement's Eckmann's method for computing the Lyapunov Exponents Schölzel [2020], Eckmann et al. [1986]. A matrix and embedding dimensionality of 5 was used to be consistent with the intrinsic dimensionality $m = k = 2$ of the $\text{C}_{24}\text{H}_{50}$ system.

Fig.4.6 shows how varying the number of points used in these computations alters the computed Lyapunov exponents. This is a well documented phenomenon Rosenstein et al. [1993]. We note that for all computations, the trend in Lyapunov exponents increases monotonically with points used. In all cases except for $n < 5000$, rank ordering in KSEs and leading Lyapunov Exponents is preserved. Moreover, the spacing between values computed for each choice of time series remain largely

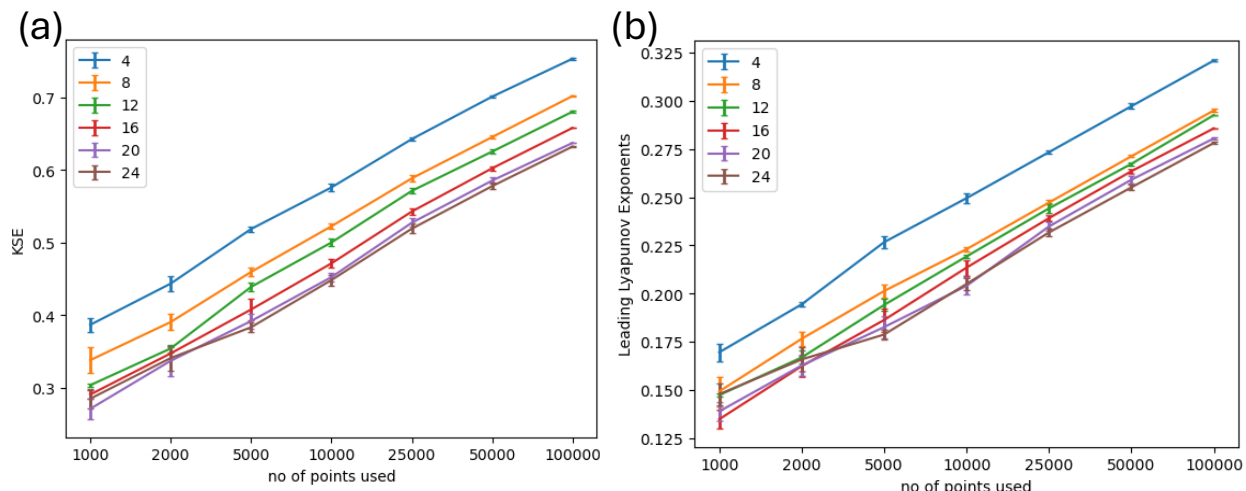


Figure 4.6: (a) Kolmogorov-Sinai Entropies (b) Leading Lyapunov Exponents as a function of number of points used in training for time series of the distance between the first and i^{th} atom in $C_{24}H_{50}$.

static across number of points used. As we are not interested in the converged values of KSEs and Lyapunov Exponents but rather this rank ordering as a way to determine optimal time series choice, we find that any sufficiently large quantity of points used in this calculation is sufficient to make this determination. The 2 000 data points we use are sufficient to make this determination. Errors are computed by bootstrapping; we compute both quantities from 3 different parts of the $C_{24}H_{50}$ MD trajectory.

4.12.4 Villin molecular dynamics and synthetic smFRET

We use a simulation of the Villin polypeptide, a 35-residue 577 atom protein, with protonated HIS residue (PDB ID: 2F4K) conducted by D.E. Shaw Research on the supercomputer Anton Lindorff-Larsen et al. [2011] as our ground truth model. This 125.6 μs MD simulation has 2 ps time resolution yielding 627 907 frames.

Reconstructions of Villin univariate and multi-temporal TAR models based on time series observation of intramolecular distances between atoms in the protein over time. These distances are extracted from the molecular dynamics simulation order to create synthetic smFRET signals. These time series correspond to the distance between the head (1st) and 286th carbon along the backbone of the protein. We learn each latent space from 20 000 time delay vectors subsampled uniformly from 1/3 of all constructed delay vectors. As we do three reconstructions to bootstrap uncertainty, we cycle the training region between the first, middle and last thirds of the training time series. Testing is performed on remaining points. Dimensionalities of each delay vector are chosen in accordance with the E1(d) analysis suggested dimensionality for each intramolecular distance trajectory with a cutoff of 0.98 Cao [1997]. Optimal delay times were extracted using the first minimum of the autocorrelation function for each time series. All time series are scaled using MinMaxScaler Pedregosa et al. [2018] between [0,1].

We reconstruct a latent space representation of Villin molecular dynamics trajectory directly from the full-dimensional data. Next, we construct a pairwise distance matrix over the all atom representations of the frames corresponding to the 20 000 training points used to learn the latent space manifold from time series. These pairwise distances correspond to the rotationally and translationally aligned root mean squared deviation of each atomic position Kabsch [1976] with minimization of RMSE about mirror symmetries arising from head to tail inversion. Details of this process are available in Topel and Ferguson [2020]. We use the same Gaussian kernel width estimation and Nyström extension method to recover the latent space manifold $\mathcal{M}$ that embeds the dynamics displayed by the Villin simulation. Following the L method of Salvador and Chan Salvador and Chan [2004], we determine that this manifold is intrinsically 3 dimensional. The method for the extraction of $\mathcal{M}'$ from Takens delay vectors is identical to that for $C_{24}H_{50}$ and the network used to map from $\mathcal{M}'$ and $\mathcal{M}$ is also identical.

To recover the full-dimensional coordinates of each of the 287 heavy atoms of Villin, we employ a

second network, also with five hidden layers, each containing 100 nodes using the ReLU activation function. Models are trained over 500 epochs, using the Adam optimizer Kingma and Ba [2014] with a learning rate of 0.01 on the first half of points corresponding to the same training set as the previous step. This network takes in the 3 coordinates of each point on $\Theta(\mathcal{M})$ and outputs the 3×287 coordinates corresponding to the x,y,z positions of each heavy atom of $\text{C}_{24}\text{H}_{50}$.

CHAPTER 5

KOLMOGOROV-SINAI ENTROPIES IDENTIFY OPTIMAL OBSERVABLES FOR PREDICTION AND DYNAMICS RECONSTRUCTION IN CHAOTIC SYSTEMS

5.1 Abstract

Embedding theorems provide theoretical insight into the relationship between time series observations of a system and the dynamics that generate them. Notably, Takens' Delay Embedding Theorem guarantees that there exists a bijective map between vectors constructed from arbitrary time series observables and the full-dimensional dynamics of that system. However, Takens' Theorem is agnostic to the choice of observable and formal bounds on the limitations of these relationships in real systems are lacking. Here we show that under modest technical conditions, the Kolmogorov-Sinai entropy is predictive of the time series that minimizes reconstruction errors in the full dimensional dynamics for chaotic ergodic systems. Using Oseledets Multiplicative Ergodic Theorem, we demonstrate that, if the conditions of Takens' Theorem are met, manifolds constructed from Takens' Theorem can be described by a filtration of subspaces that are diffeomorphically related for any observable that respects these conditions. Moreover, the exponential rates of growth and shrinkage of their subspaces can be used to estimate the sensitivity of each to error. Finally, we bound reconstruction errors as a function of the error contributions from each Oseledets subspace for a given observable, showing that they are proportional to the positive Lyapunov exponents of the system and, by the Ruelle inequality, the Kolmogorov-Sinai entropy. This work provides analytical intuition into observable selection for arbitrary dynamical systems and can be used to this effect for the modeling of dynamical systems via statistical learning.

5.2 Introduction

Advances in the fields of complex systems Strogatz [2015], Barabási [2002], Mitchell [2009] and machine learning Goodfellow et al. [2015] alike have enabled the study of complicated, interconnected and noisy systems from population dynamics May [1976] to consumer demand functions Varian [2014]. These tools have put the study and prediction of complex systems comfortably within the reach of anyone with access to compute. However, with these marked advances in learning technologies, much of the theoretical and analytical insight provided by the fields of dynamical systems and embedding theorems has gone unused. Within a sea of plausible time series observations of dynamics coming from any number of data sources, deciding which to use can be tricky. In this work, we propose how and prove why one such tool of dynamics, the Kolmogorov-Sinai Entropy (KSE), can be used to predictively select optimal time series to feed into machine learning platforms modeling chaotic ergodic systems.

We apply a combination of geometry, embedding theorems and dynamics to first define the problem of time series selection, introduce the notion of geometries from time series, discuss differences in the manifolds describing dynamics of each and finally prove how the KSE can be used to optimize time series selection. In Sect. 5.3, we introduce the notion of ergodic dynamics and invariant manifolds arising from complex dynamics. In Sect. 5.4, we discuss Takens' Delay Embedding Theorem and its application in ergodic dynamics. In Sect. 5.5, we describe types of geometrical distortion to attractors constructed from time series. In Sect. 5.6, we introduce Oseledec's Multiplicative Ergodic Theorem and its application to Takens embeddings under certain conditions. In Sect. 5.7, we define Lyapunov exponents and the Kolmogorov-Sinai entropy. Finally, in Sect. 5.8, we prove that under certain modest technical conditions, Kolmogorov-Sinai entropy predictively selects time series that minimize reconstruction errors when backmapping to full dimensional dynamics.

5.3 Ergodic Dynamics and Attractors

We can define the accessible phase space of a dynamical system $\mathbb{S}$ in terms of its measurable phase space given by the triple (X, Σ, μ) where X defines accessible states, Σ specifies the σ -algebra consisting of all measurable subsets in X , and μ is the probability measure defining the probability of measurable sets in X Walters [1982], Cornfeld and Fomin [1982]. We can also define some measure preserving transformation $T : X \rightarrow X$ that propagates X through $\mathbb{S}$. T is said to be ergodic if for every invariant set $A \in \Sigma$ satisfying $T^{-1}(A) = A$, $\mu(A) = 0$ or $\mu(A) = 1$ Walters [1982].

For instance, in classical mechanics, Liouville's Theorem guarantees that there exists a time evolution map given by Hamilton's equations of motion where phase space volume is preserved under the time evolution Arnold [1989]. More formally, for any measurable set $A \subseteq X$, $\mu(A)$ remains constant under time evolution, $\mu(T^t(A))$. This transformation is said to be symplectic (volume-preserving) Arnold [1989]. It may also be ergodic with respect to the Liouville measure; however, constraints on dynamics can result in the emergence of restrictions to accessible phase space.

The existence of conserved quantities such as energy in classical physical systems results in the emergence of lower dimensional surfaces (invariant manifolds) in phase space that can effectively describe system dynamics Noether [1918]. For instance, an N -atom molecular system evolving in some high dimensional phase space $\mathbb{R}^{6N}$ has an effective description on an invariant manifold of $k \ll 6N$ dimensions. These dynamics are restricted to $\mathbb{R}^{6N-1}$ by energy constraints and then can be further confined to $\mathbb{R}^k$ as a result of atomic interactions such as electrostatic forces Wales [2003]. We term this subset of phase space the *attractor*, although this term is also used for manifolds emerging from dissipative dynamics where trajectories converge. We extend this usage to describe any low dimensional and invariant subset of an original high dimensional description. In this work, we consider the broader context of both dissipative and chaotic systems giving rise to attractors. Such systems may have an invariant Sinai–Ruelle–Bowen (SRB) measure describing the statistical

behavior of chaotic attractors which are typically ergodic on systems where they are defined Young [2002], Ruelle [1979]. These attractors have been shown to be robust to perturbations as in Kolmogorov–Arnold–Moser (KAM) theory Arnold [1989] and in the works of Stark and Broomhead under both stochastic and deterministic forcing Stark [1999], Stark et al. [2003], Broomhead and King [1986]. If we constrain X to these invariant attractors, we can often define some transformation T that is ergodic with respect to a measure μ on these invariant manifolds Walters [1982], Eckmann and Ruelle [1985].

5.4 Takens’ Delay Embedding Theorem

Takens’ Delay Embedding Theorem states that for any dynamical system, $\mathbb{S}$, with smooth dynamics describable by a compact attractor (i.e. on an invariant manifold), for which we can record a time series observation of its dynamics, vectors constructed by recording time delayed scalar observables of $\mathbb{S}$ constitute an embedding of the attractor of $\mathbb{S}$ onto $\mathbb{R}^{2m+1}$ where m is the dimensionality of the attractor under modest technical conditions Takens [1981]. These vectors have form $y(t) = [o_t, o_{t-\tau}, \dots, o_{t-2m\tau}] \in \mathbb{R}^{2m+1}$ and their embedding describes the geometry and dynamical properties of $\mathbb{S}$ up to an *a priori* unknown diffeomorphism Takens [1981].

This powerful theorem can be used to extract robust descriptions of a dynamical system from time series observations of that system. This is particularly convenient for complex dynamics where the forces or rules that propagate that system are unknown, or for large interacting systems like molecules, where propagating those rules in a full-dimensional form is very expensive. While this theorem allows us to identify these intrinsic manifolds describing dynamics, these manifolds do not necessarily have ergodic measures. There may exist multiple invariant components to these manifolds and ergodicity requires that these systems not be decomposable. That is to say, for an ergodic measure μ on the map T describing all accessible states X for (X, Σ, μ) , there cannot be some $A \subset X$ that is not accessible by the propagation of any other subset $A' \subset X$, $T^t(A')$. If there

were multiple closed sets in X , X could be subdivided into multiple subsets where $\mu(A) = 0, 1$, violating ergodicity. In this case, an invariant measure could be composed on each closed subset of X , however the decomposability would violate the requirements of ergodicity.

Here we will appeal to two facts. 1) If we collect continuously observed time series observations of a dynamical system S then we typically only capture the states within an invariant subset of the phase space Strogatz [2015]. We can then apply ergodicity to this subset of phase space if there is an ergodic invariant measure on it. It should be noted that stochastic or deterministic forcing causing a jump to a different invariant manifold is possible in the case where there exist multiple invariant manifolds describing dynamics for forced systems Grebogi et al. [1983]. 2) It is often the case that the chaotic dynamical systems we discuss in this work enjoy an invariant Sinai–Ruelle–Bowen measure describing that is typically ergodic on systems where they are defined Young [2002]. This SRB measure is considered to be physical because it describes statistical behavior congruent on almost any initial conditions characteristic of the attractor Young [2002], Eckmann and Ruelle [1985].

5.5 Distortions to Embedded Manifolds

Our discussion of both Takens' Theorem and dynamics has been somewhat idealized. In the real world, our time series observations may be subject to noise or deterministic forcing. Moreover, chaotic systems are subject to "deterministic noise", small differences in initial conditions resulting in exponentially different trajectories over time Eckmann and Ruelle [1985]. Consequently, although a system may evolve deterministically, the time series observations of that system may resemble stochastically defined noise. This high degree of sensitivity to initial conditions for chaotic systems means that any imprecision in our definition of a state $x \in X$ can lead to a mischaracterization of the embedding $\mathcal{M} \subseteq \mathbb{R}^{2m+1}$ we can produce from Takens' Theorem. Possible sources of error abound. Stochastic noise from recording equipment, deterministic forcing from a second and separable form

of dynamics influencing our system or even errors determining the correct value of x due to errors accrued solving differential equations numerically over long time scales can limit our ability to correctly and uniquely define the set of accessible states X on an attractor over time Kantz and Schreiber [2004].

It is fortunate, then, that the works of Stark and Broomhead Broomhead and King [1986], Stark [1999], Stark et al. [2003] have demonstrated the robustness of Takens' Theorem to such deterministic and stochastic forcing. However, we expect that the degree of sensitivity of any one of our potential time series observations to noise will result in distortions to the embedding $\mathcal{M}' \subseteq \mathbb{R}^{2m+1}$ that will be a function of individual manifold topography. Consequently, while Takens' Theorem applies to generic observables, in practice, observable choice can affect the quality of modeling performed using Takens delay vectors constructed from that time series Kantz and Schreiber [2004], Takens [1981], Sauer et al. [1991]. To our knowledge, formal guidelines surrounding the selection of such time series observables do not exist.

We can, however, classify the types of distortions to latent space representations of dynamics that can arise in manifolds constructed from Takens' Theorem. While Takens delay embeddings are guaranteed to be diffeomorphically related to manifolds describing full dimensional dynamics, they are subject to smooth deformations of their topography including stretching, squashing, bending, local warping, twisting and shearing Hirsch et al. [2003]. These will result in deformations to the intrinsic geometry of the manifold, meaning that there will be changes to the distances between embedded points on $\mathcal{M}'$ produced by Takens' Embedding.

Any *intrinsic* geometrical deformation alters distances on the manifold while *extrinsic* ones may visually affect the manifold but leave distances unchanged Hirsch et al. [2003]. Minimizing intrinsic distortions leads to more faithful representations of system dynamics. Moreover, larger deformations present the opportunity for greater sensitivity to mischaracterization of $r'(t) \in \mathcal{M}'$ and consequently, observable selection should be done in a fashion that minimizes such deformations. In the following

Deformation Type	Description	Intrinsic Geometry Conserved
Stretching/Squashing	Expansion or compression of regions of phase space.	No
Bending	Smooth curving that does not induce folds.	Yes
Local Warping	Small deformations on the surface.	No
Twisting	Smooth rotation.	Yes
Shearing	Deformation in angles and distances.	No

Table 5.1: Types of deformations of the manifold $\mathcal{M}'$ under C^1 equivalence as guaranteed by Takens' Theorem that may vary with observable choice Hirsch et al. [2003], Takens [1981], Sauer et al. [1991].

section, we apply Oseledets Multiplicative Ergodic Theorem to study the rate of such distortions.

5.6 Oseledets Multiplicative Ergodic Theorem

It is now incumbent upon us to study the rate at which the reconstructed manifold $\mathcal{M}' \subseteq \mathbb{R}^k$ produced by Takens' Theorem is distorted. We assume first that our observations describe an effective phase space described by an attractor, $\mathcal{M} \subseteq \mathbb{R}^k$, and that the time evolution of state $r(t) \in \mathcal{M}$ is described by transformation T ergodic on the invariant measure μ . We contend that this is a fair assumption for a wide variety of chaotic dynamical systems as discussed in the previous sections. Moreover, ergodicity ensures that time averages converge to ensemble averages via Birkhoff's Ergodicity Theorem, ensuring reliability of the statistical properties derived from observations Birkhoff [1931].

Oseledets multiplicative ergodic theorem (OMET) provides the convergence criterion in the rates of growth or shrinkage of the tangent space (or bundle) at almost every point on the attractor $\mathcal{M} \subseteq \mathbb{R}^{2m+1}$ Oseledets [1968]. It is multiplicative as it deals with the product of linear operators and is vector-valued; that is to say it describes time averaged behavior of the evolution of a vector under a sequence of linear transformations to our system Arnold [1998]. For some linear operator $A(r)$ operating on the orbit $T^j(r)$ acting on vector $r(t) \in \mathcal{M} \subseteq \mathbb{R}^{2m+1}$, there exist statistical limits to the evolution of r propagated by $A_t(r) = A(T^{t-1}(r)) \cdots A(T(r))A(r)$. This operator corresponds to the cocycle over the base dynamics, T . OMET tells us that we can subdivide the tangent space of

the attractor $\mathcal{M} \subseteq \mathbb{R}^{2m+1}$ into a filtration of subspaces $\{0\} \subseteq V_0 \subseteq V_1 \subseteq \dots \subseteq V_q \subseteq V_{q+1} \subseteq \dots \subseteq V_K \subseteq \mathbb{R}^{2m+1}$ for which there exist a corresponding set of Lyapunov exponents $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K$ measuring rate of growth or shrinkage of each subspace under the transformation A Oseledets [1968], Arnold [1998]. Moreover, these rates of growth and shrinkage are given by the limit behavior of the evolution of our system at long timescale:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log \|A_t(r)v\| = \lambda_q \quad (5.1)$$

where $v \in V_q \setminus V_{q-1}$ is a nonzero vector in subspace V_q but not V_{q-1} Oseledets [1968], Arnold [1998].

We can equivalently apply OMET to some manifold extracted via Takens' Delay Embedding Theorem $\mathcal{M}' \subseteq \mathbb{R}^{2m+1}$ that enjoys a smooth, invertible and bijective map between manifolds, $\Phi: \mathcal{M}' \mapsto \mathcal{M}$. leading to its own filtration of subspaces $\{0\} \subseteq V'_0 \subseteq V'_1 \subseteq \dots \subseteq V'_q \subseteq V'_{q+1} \subseteq \dots \subseteq V'_K \subseteq \mathbb{R}^{2m+1}$ for which there exist a corresponding set of Lyapunov exponents $\lambda'_1 \geq \lambda'_2 \geq \dots \geq \lambda'_K$ produced by the limiting behavior:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log \|A'_t(r')v'\| = \lambda'_q \quad (5.2)$$

where r' is a point on the reconstructed manifold $\mathcal{M}'$, $A'_t(r')$ describes the linear map of propagating r' over base dynamics T' with orbit T'^j and $v' \in V'_q \setminus V'_{q-1}$ is a vector in the tangent space at r' .

Takens theorem guarantees that the manifolds $\mathcal{M}$ and $\mathcal{M}'$ are diffeomorphically related Takens [1981]. As we have assumed that base dynamics T and T' are ergodic on the invariant measure μ and μ' respectively, we can apply OMET. Moreover, there exists a bijective map between any given point on the orbits $T^j(r)$ and $T'^j(r')$ in the limit that the requirements of Takens' theorem are

fulfilled. The induced diffeomorphism between $\mathcal{M}$ and $\mathcal{M}'$ holds for their tangent bundles; this ensures invariant splittings defined by the filtration in OMET, V_q and V'_q , also enjoy a diffeomorphic relation.

Earlier, we argued that the fidelity of latent space manifold extracted via Takens' theorem, $\mathcal{M}'$, is a function of stochastic and deterministic forcing as well as inconsistencies in the recording of time series O . Applying this to our understanding of OMET, it follows that reconstruction of any given invariant subspace V_q will then be a function of the deformation to the intrinsic geometry of each subspace induced by noise and forcing. It then follows that the sensitivity of each subspace V'_q to infinitesimal errors in our estimation of $v'_q \in V'_q$ is given by λ'_q . This fact is the basis of our proof in Sect. 5.8. In the next section we explore the interpretation of these Lyapunov exponents and the aggregate effect of their values across all subspaces.

5.7 Lyapunov Exponents and the Kolmogorov Sinai Entropy

The Lyapunov exponents of a dynamical system describe the average exponential rate of divergence or convergence of infinitesimally close trajectories on a phase space describing the evolution of a dynamical system. Empirically, they are extractable by fitting a function of mean distance between trajectories $\delta(t) = \delta(0)e^{\lambda t}$ where $\delta(t)$ describes the mean distance between trajectories at time t with Lyapunov exponent λ . Positive Lyapunov exponents correspond to exponential divergence in trajectories, negative ones to convergence or dissipation while a zero Lyapunov exponent typically characterizes motion along a limit cycle attractor Eckmann et al. [1986], Rosenstein et al. [1993].

In the previous section, we discussed how the rate of growth and shrinkage of subspaces for systems satisfying OMET are related to these exponents. They can be extracted directly from time series using the method of Eckmann *Eckmann et al.* [1986]. Now we consider the aggregate effect of these deformations on the time series.

The **Kolmogorov-Sinai Entropy** The Kolmogorov-Sinai Entropy (KSE), h^{KS} , quantifies the average rate of information production in the system. It is defined as

$$h^{\text{KS}} = \lim_{\varepsilon \rightarrow 0} \lim_{n \rightarrow \infty} \frac{1}{n} H(\varepsilon, n), \quad (5.3)$$

where $H(\varepsilon, n)$ is the Shannon entropy of trajectories defined with resolution ε over n -step intervals Walters [1982]. However, it is also extractable directly from the Lyapunov exponents. Intuitively, this is because positive Lyapunov exponents describe chaotic behavior, and consequently information production. Ruelle's inequality allows us to bound this directly from the positive Lyapunov exponents of our time series Ruelle [1978]:

$$h^{\text{KS}} \leq \sum_{\lambda_i \geq 0} \{\lambda_i\} \quad (5.4)$$

Under appropriate conditions (e.g., for smooth hyperbolic systems with an SRB measure), Pesin's Theorem shows these values are indeed equal Pesin [1977], Young [2002].

As the upper bound limit of the KSE is computed from the Lyapunov exponents of a given time series and these are computed by the mean divergence in the distance of infinitesimally close points on the manifold extracted from this time series ($\mathcal{M}'$), we hypothesize that manifolds with larger KSEs will be more sensitive to noise, forcing and errors in specifying a given $r' \in \mathcal{M}'$

5.8 Theorem and Proof

In previous work, we demonstrated that minimizing Kolmogorov-Sinai entropies of input time series minimize error in a reconstruction of a backmapping from scalar time series observations of the straight chain hydrocarbon $\text{C}_{24}\text{H}_{50}$ to configurational representations of each of its 24 carbons in Euclidean space. Here we prove that time series exhibiting smaller Kolmogorov-Sinai entropies

can produce reconstruction of full dimensional dynamics with smaller mean errors than those with large errors.

Theorem 1. *Let $\mathbb{S}$ be a dynamical system whose dynamics are chaotic and evolve on the intrinsic attractor $\mathcal{M} \subseteq \mathbb{R}^{2m+1}$. Let $O_i(t), i \in \{1, \dots, n\} \in \mathbb{N}$ be a set of n true scalar time series observables of $\mathbb{S}$. Due to imperfect measurement, external forcing and stochastic noise, the observed time series is then $\tilde{O}(t) = O(t) + \varepsilon(t)$ where $\varepsilon(t)$ represents the perturbation at time t . Furthermore, we assume that $\varepsilon(t)$ is drawn from some probability distribution Ω . Ω can be time dependent in the case of deterministic forcing. Let $\|\varepsilon(t)\| \ll \|O(t)\|$ and let this noise be uncoupled with the nonlinear dynamics of $\mathcal{M}$. Assume that the trajectories of each $O_i(t)$ produce embeddings $\mathcal{M}_i \subseteq \mathbb{R}^{2m+1}$ via Takens' Theorem upon which the time-evolved trajectories on the embedded manifolds, $T_i(t)$, exist. Assume that all of the requirements of Takens' Theorem be fulfilled including sufficient sampling time such that all $\mathcal{M}_i$ will be diffeomorphically related to each other and the intrinsic attractor $\mathcal{M}$. Assume each orbit $T_i(t)$ on $\mathcal{M}_i$ be ergodic with respect to measure μ_i pertaining to the dynamics on each manifold.*

Let h_i^{KS} be the effective Kolmogorov-Sinai entropy whose upper bound limit under Ruelle's Inequality Ruelle [1978], $h_i^{KS,UB}$, is taken as the sum of positive Lyapunov exponents of $\mathcal{M}_i$. Under noise, the effective Kolmogorov-Sinai entropies of each observation are not guaranteed to be invariant as this condition requires measure-preserving isomorphisms (i.e. smooth, invertible changes of coordinates) between all $\mathcal{M}_i$ and $\mathcal{M}$ Walters [1982] which may be violated by these perturbations. Let ε_i be the lowest possible mean root mean squared error in reconstruction of all points $r(t) \in \mathcal{M}$ from $r_i(t) \in \mathcal{M}_i$ performed by any backmapping procedure $\hat{\phi}$ that is locally Lipschitz continuous about any point $r_i(t) \in \mathcal{M}$ with $L > 0$. Then, among the candidate observables, $O_i(t)$, the one with the smallest effective Kolmogorov-Sinai entropy upper bound limit produces the minimum possible ε_i . That is to say if

$$h_i^{KS,UB} \leq h_k^{KS,UB}$$

then

$$\varepsilon_i \leq \varepsilon_k$$

$\forall k \in \{1, \dots, n\}$ as these correspond to minimally chaotic embeddings $\mathcal{M}_i$ which are, consequently, the least susceptible to error due to deterministic forcing and stochastic noise.

Proof. As the orbits $T_i(t)$ are ergodic with respect to their invariant measures μ_i , Birkhoff's Ergodicity Theorem guarantees that time averages computed along these orbits are equivalent to ensemble averages. This justifies the use of the Kolmogorov–Sinai entropies and the application of Oseledets Multiplicative Ergodic Theorem that follow.

Step 1: Establishing diffeomorphic relationships

Oseledets Multiplicative Ergodic Theorem (OMET) provides the convergence criterion in the rates of growth or shrinkage of the tangent space (or bundle) at almost every point on the attractor $\mathcal{M} \subseteq \mathbb{R}^{2m+1}$. For some linear operator $A_t(r) = A(T^{t-1}(r)) \cdots A(T(r))A(r)$ operating on the orbit $T^j(r)$ acting on vector $r(t) \in \mathcal{M} \subseteq \mathbb{R}^{2m+1}$, OMET tells us that we can subdivide the tangent space of the attractor $\mathcal{M} \subseteq \mathbb{R}^{2m+1}$ into a filtration of subspaces $\{0\} \subseteq V_0 \subseteq V_1 \dots V_q \subseteq V_{q+1} \dots \subseteq V_K \subseteq \mathbb{R}^{2m+1}$ for which there exist a corresponding set of Lyapunov exponents $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K$ measuring rate of growth or shrinkage of each subspace under the transformation A :

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log \|A_t(r)v_q\| = \lambda_q \quad (5.5)$$

where $v \in V_q \setminus V_{q-1}$ is a nonzero vector in subspace V_q but not V_{q-1} .

We can equivalently apply OMET to $\mathcal{M}_i \subseteq \mathbb{R}^{2m+1}$ constructed from observable $O_i(t)$ via Takens' Delay Embedding Theorem, leading to its own filtration of subspaces $\{0\} \subseteq V_0^i \subseteq V_1^i \dots V_q^i \subseteq$

$V_{q+1}^i \dots \subseteq V_K^i \subseteq \mathbb{R}^{2m+1}$ for which there exist a corresponding set of Lyapunov exponents $\lambda_1^i \geq \lambda_2^i \geq \dots \geq \lambda_K^i$.

Takens' theorem guarantees that the manifolds $\mathcal{M}$ and $\mathcal{M}_i$ are diffeomorphically related; that is to say there exists a smooth, invertible mapping between the two. Moreover, there exists a bijective map between any given point on the orbits $T(r)$ and $T^i(r_i)$ in the limit that the requirements of Takens' theorem have been fulfilled. The induced diffeomorphism between $\mathcal{M}$ and $\mathcal{M}_i$ holds for their tangent bundles; this ensures invariant splittings defined by the filtration in OMET also enjoy a diffeomorphic relation. Consequently the reconstruction of any given invariant subspace V_q will be a function of the deformation to the intrinsic geometry of each subspace induced by noise and forcing. That is to say, the sensitivity of each subspace V_q^i to infinitesimal errors in our estimation of $v_q^i \in V_q^i$ is given by λ_q^i .

Summary: Provided the requirements of Takens' Delay Embedding Theorem are met, Oseledets multiplicative ergodic theorem guarantees that both the true intrinsic manifold describing the dynamics of $\mathbb{S}$, $\mathcal{M}$ and its reconstruction from time series $O_i(t)$ are diffeomorphically related as are the subspaces in their OMET filtrations.

Step 2: Relating Lyapunov exponents to error sensitivity

Let the set of true scalar observables be denoted by $O_i(t)$ and due to imperfect measurement, external forcing and stochastic noise, the observed time series is then $\tilde{O}(t) = O(t) + \varepsilon(t)$ where $\varepsilon(t)$ represents the perturbation at time t . Furthermore, we assume that $\varepsilon(t)$ is drawn from some probability distribution Ω . Note that Ω can be time dependent in the case of deterministic forcing. Let $\|\varepsilon(t)\| \ll \|O(t)\|$.

Reconstruction of the embedding $\mathcal{M}_i$ from $\tilde{O}(t)$ via Takens' Delay Embedding Theorem will propagate these perturbations leading to an error in reconstructed states. We denote the true state of

the system at time t on the attractor $\mathcal{M}$ as $r(t)$ and state on the reconstructed attractor $\mathcal{M}_i$ extracted from $\tilde{\mathcal{O}}_i(t)$ is given as $r_i(t)$. Then the reconstruction error at time t is given by:

$$\varepsilon_i(t) = \|r(t) - \hat{r}_i(t)\|$$

where $\hat{r}_i(t)$ is the output of the backmapping from $r_i(t)$ to $r(t)$.

We consider a linearization of these dynamics around a given point $r_i(t)$. Let $A(t)$ be a linear operator propagating $r_i(t)$ to a new $r_i(t + \delta t)$. Then a small perturbation to $r_i(t)$, $\delta(t)$ will result in perturbation

$$\delta(t + \tau) \approx A(t)\delta(t)$$

.

Given an initial perturbation $\delta r(0) = \varepsilon(0)$ and assuming a single positive Lyapunov exponent, λ , then

$$\delta r(\tau) \approx \varepsilon(0)e^{\lambda\tau}$$

Extending this to $\mathcal{M}_i$ with its K Lyapunov exponents, λ_q^i , corresponding to the q^{th} invariant subspace V_q^i , then the sensitivity of dynamics in each subspace is given by:

$$\delta_q^i(\tau) = \varepsilon_q(0)e^{\lambda_q^i\tau}$$

where $\varepsilon_q(0)$ describes the initial perturbation in subspace V_q^i . Summing over all subspaces, the total error is given by the Euclidean distance between the perturbed and unperturbed states:

$$\delta_{\text{total}}^i(\tau) = \sqrt{\sum_{q=1}^K \left(\varepsilon_q(0) e^{\lambda_q^i \tau} \right)^2}$$

This effect is dominated by those subspaces whose $\lambda_q^i > 0$ as these experience exponential divergence.

We can then relate the reconstruction error ε_i to $\delta_{\text{total}}^i(\tau)$. As the sum of the positive Lyapunov exponents (KSE) decreases, so too does $\delta_{\text{total}}^i(\tau)$. In the limit of Takens' Theorem holding, $\mathcal{M}_i$ and $\mathcal{M}$ enjoy a diffeomorphic relationship. The limitation to learning this mapping with a universal functional approximator in this limit comes from errors in the estimation of any $r_i(t) \in \mathcal{M}_i$. Then the reconstruction error ε_i is a function of $(\delta_{\text{total}}^i(\tau))$.

Summary: For the noisy time series $\tilde{\mathbf{O}}(t)$, points on the reconstructed manifold $\mathcal{M}_i$ will be reconstructed with mean error $\delta_{\text{total}}^i(\tau) = \sqrt{\sum_{q=1}^K \left(\varepsilon_q(0) e^{\lambda_q^i \tau} \right)^2}$

Step 3: Error in the backmapping of $r_i(t)$

There exists a smooth, continuous, invertible and diffeomorphic mapping between the manifolds $\mathcal{M}_i \rightarrow \mathcal{M}$ given by Takens' Theorem:

$$\phi : \mathcal{M}_i \rightarrow \mathcal{M}$$

approximated by continuous backmapping function $\hat{\phi} : \mathcal{M}_i \rightarrow \mathcal{M}$. Then for every $\varepsilon > 0 \exists \delta > 0$ such that for any $r_i(t) \in \mathcal{M}_i$ perturbation δ_{total}^i with

$$\|r_i(t) - (\hat{r}_i(t) + \delta_{\text{total}}^i)\| < \delta$$

when the there $r_i(t)$ and $r_i(t) + \delta_{\text{total}}^i \in \mathcal{M}_i$ are mapped onto $\mathcal{M}$, they will be at some distance

$$\|\hat{\phi}(r_i(t)) - \hat{\phi}(r_i(t) + \delta_{\text{total}}^i)\| < \varepsilon$$

away from each other. Then if $\hat{\phi}$ is locally Lipschitz continuous about any given $r_i(t)$ with Lipschitz constant $L > 0$:

$$\|\hat{\phi}(r_i(t)) - \hat{\phi}(r_i(t) + \delta_{\text{total}}^i)\| \leq L \|\delta_{\text{total}}^i\|$$

Consequently, small errors in the estimation of $r_i(t)$ translate monotonically into small errors in the reconstructed state $\hat{\phi}(r_i(t))$.

Summary: Small errors in estimation of $r_i(t)$, $(\delta_{\text{total}}^i(\tau))$ result in reconstruction errors in the estimation of $\hat{\phi}(r_i(t))$ that are monotonically related to $(\delta_{\text{total}}^i(\tau))$

Step 4: Bounding the mean error, ε_i

We define the mean reconstruction error ε_i as the mean of the RMSEs computed for each frame. That is,

$$\varepsilon_i = \frac{1}{N} \sum_{t=1}^N \|r(t) - \hat{r}_i(t)\| = \frac{1}{N} \sum_{t=1}^N \varepsilon_i(t)$$

where N is the total number of time frames over which the reconstruction is computed.

Then substituting the error bound for a single time step in **Step 3**, ε_i evaluates to

$$\|r(t) - \hat{r}_i(t)\| = \|\hat{\phi}(r_i(t)) - \hat{\phi}(r_i(t) + \delta_{\text{total}}^i(t))\| \leq L \|\delta_{\text{total}}^i(t)\|$$

The overall reconstruction error ε_i can be bounded by the average of these local errors:

$$\varepsilon_i \leq \frac{L}{N} \sum_{t=1}^N \|\delta_{\text{total}}^i(\tau)\|$$

In **Step 2**, we showed that

$$\|\delta_{\text{total}}^i(\tau)\| = \sqrt{\sum_{q=1}^K \left(\varepsilon_q(0) e^{\lambda_q^i \tau} \right)^2}$$

Consequently, as ε_i grows monotonically with $\frac{L}{N} \sum_{t=1}^N \|\delta_{\text{total}}^i(\tau)\|$, $\frac{L}{N} \sum_{t=1}^N \|\delta_{\text{total}}^i(\tau)\|$ grows monotonically with increased λ_q^i and $h_i^{KS,UB} \leq \sum_n \lambda_n^i$ which grows monotonically with increases in each λ_q^i , then if

$$h_i^{KS,UB} < h_k^{KS,UB}$$

then

$$\varepsilon_i \leq \varepsilon_k$$

Conclusion: Then if $h_i^{KS,UB} < h_k^{KS,UB}$, $\varepsilon_i \leq \varepsilon_k$.

□

5.8.1 Limitations

We recognize the following limitations to this theorem and proof. We assume that errors along different invariant subspaces combine in a Euclidean manner. This assumption of local linearizability will break down under strong enough nonlinearity or large enough noise Strogatz [2015], Kantz

and Schreiber [2004]. We have treated noise as being separable from dynamics. In many cases this assumption holds, however if noise interacts with the system’s nonlinear dynamics, the linearized error propagation may not be sufficient to bound errors in recorded $r_i(t)$. Examples of this include accrued errors from the propagation of differential equations over long time periods with numerical means Frenkel and Smit [2002] and experimental observations with error that is a function of the recorded observable such as single molecule fluorescence energy transfer (smFRET Roy et al. [2008]).

We reiterate that the proof applies to functional approximators respecting this condition of local Lipchitz continuity. We note that this condition is required for any smooth and diffeomorphic map between $\mathcal{M}_Y$ and $\mathcal{M}$ Rudin [1976], Spivak [1965]. While we cannot make such a guarantee for any arbitrary universal functional approximator, if such an approximator is close to the true map, it must also respect this condition of a smooth and diffeomorphic mapping as guaranteed by Takens’ Theorem.

Finally, as noted earlier, under appropriate conditions (e.g., for smooth hyperbolic systems with an SRB measure), Pesin’s Theorem shows that the Ruelle Inequality actually becomes an equality Ruelle [1978], Pesin [1977], Young [2002].

5.9 Conclusion

We have now proven that manifolds constructed from time series with greater Kolmogorov-Sinai entropies experience greater mean level of reconstruction errors under modest technical conditions. In Chapter 4, we demonstrated this finding computationally with the molecular system $C_{24}H_{50}$. This proof represents an analytical description for the physical intuition surrounding observable selection. It provides a means of selecting observables to model arbitrary dynamical systems before performing costly modeling.

There are many other bounding problems in the embedding theorems space that could inform input design in the statistical learning space. In future work, we seek to extend this proof to describe the size of these reconstruction errors as a function of sampling time. Moreover, we would like to form a deeper understanding of the nature of topographical deformations in manifolds constructed from Takens' Theorem. Advances in these and similar areas will provide a mathematical foundation for observable selection in complex systems where such problems are beyond physical intuition.

CHAPTER 6

DATA-DRIVEN DISCOVERY OF LINEAR MOLECULAR PROBES WITH OPTIMAL SELECTIVE AFFINITY FOR PFAS IN WATER

6.1 Abstract

Approaches to tackle the wide and growing variety of highly persistent per- and polyfluoroalkyl substances (PFAS) are of pressing global need because of their detrimental human health effects such as cancer, birth defects, and hormone imbalance. Sensitive, selective, and easy-to-use real-time sensors to monitor and detect PFAS and sorbents to extract them are critical to meeting government-mandated environmental concentrations. In this work, we combine all-atom molecular dynamics simulations, enhanced sampling, deep representational learning, and Bayesian optimization to perform high-throughput virtual screening for highly sensitive and selective molecular probes. Our molecular design space consists of 3850 linear hydrocarbon chains with varying degrees of halogenation with and without amine and phosphine-based head groups. By employing a data-driven search process, we efficiently explore the molecular design space to optimize sensitivity to perfluorooctanesulfonic acid (PFOS) as a prototypical PFAS analyte and selectivity relative to a sodium dodecyl sulfate (SDS) interferent. We calculate 504 Gibbs free energies of probe-analyte and probe-interferent interactions and identify probes with PFOS association free energies of up to $(-\Delta G_{PFOS}) = (9.8 \pm 0.2)$ kJ/mol and selectivities relative to SDS of $(-\Delta \Delta G_{PFOS-SDS}) = (3.1 \pm 1.5)$ kJ/mol. A $C_{11}Br_{23}P(CH_3)_2$ probe containing 11 backbone brominated carbons and a tertiary phosphine head group possesses the most sensitive binding constant to PFOS within the defined search space of $K_b^{PFOS} = (177.4 \pm 12.7) M^{-1}$, and a semibrominated probe $C_5H_{11}C_7Br_{14}N(CH_3)_2$ containing twelve backbone carbons and a tertiary amine head group possesses the highest selectivity relative to SDS $K_b^{PFOS}/K_b^{SDS} = (4.6 \pm 1.7)$. A retrospective analysis of our data to extract interpretable design rules reveals that the sensitivity of linear hydrogenated probes increases by approximately 1

kJ/mol per C-C bond. Addition or removal of halogen atoms and amine or phosphine head groups produces non-monotonic changes in both sensitivity and selectivity with changes to sensitivity of up to 2.5 kJ/mol. This work places empirical limitations on the performance of a wide range of linear probes for PFOS detection, and offers a generic strategy for the high-throughput computational screening to promote selective and sensitive binding.

6.2 Introduction

Per- and polyfluoroalkyl substances (PFAS) represent a large group of synthetic chemicals defined by the U.S. Environmental Protection Agency (EPA) as compounds containing $R-(CF_2)_n-C(F)(R')R''$ units, where CF_2 and CF groups are saturated carbons and the R groups (R , R' , or R'') are not hydrogens. Wallington et al. [2021], U.S. Environmental Protection Agency [2022e] Due to their excellent properties including heat resistance, hydrophobicity and oleophobicity, PFAS have been extensively used in numerous commercial applications such as non-stick cookware, firefighting foam and flame retardant, stain resistant fabrics, and fast food packaging. Baker and Knappe [2022], Glüge et al. [2020], Xia et al. [2022], Sim et al. [2021] PFAS compounds enter the environment during the production, use, or waste of these consumer products and, due to the highly stable carbon-fluorine covalent bonds – one of the strongest single bonds with a dissociation energy of ~ 440 kJ/mol due to the high electronegativity of fluorine Wang et al. [2021b] – are extremely persistent, resulting in PFAS accumulation in both the environment and living organisms and leading to them often referred to as “forever chemicals”. Ghisi et al. [2019], De Silva et al. [2021], Cousins et al. [2022], Kwiatkowski et al. [2020] PFAS exposure can cause adverse health effects in humans such as increased cholesterol levels, birth defects, disruption in thyroid hormone balance, and potentially a high risk of cancer. Fenton et al. [2021] Due to these detrimental properties, PFAS attracted broad interest from various regulatory agencies worldwide Brennan et al. [2021]. It is one of the pressing global challenges to address the growing quantities of PFAS in the environment, especially in water resources from which these molecules are readily disseminated to living organisms. Ghisi et al.

[2019], Wang et al. [2021b], Cousins et al. [2022] This requires an approach to effectively and efficiently detect and eliminate the growing number (> 4700 as per the EPACompTox Chemicals Dashboard, U.S. Environmental Protection Agency [2022]) of PFAS variants and to do so at or below the very dilute concentrations of approximately 4 ng/L mandated by EPA guidelines for safe drinking water U.S. Environmental Protection Agency [2023].

Current standard approaches available for regulatory or guidance activities of PFAS in water sources by the EPA utilize solid-phase extraction (SPE) and liquid chromatography/tandem mass spectrometry (LC-MS/MS) based analytical tools. Wang et al. [2021b], U.S. Environmental Protection Agency [2022d] These approaches formally referred as Method 537.1 and 533 can measure 18 and 25 types of PFAS, respectively with a common total of 29 PFAS molecules in potable water sources. U.S. Environmental Protection Agency [2022c,b] The lowest concentration minimum reporting level (LCMR) for Method 537.1 and 533 ranges from 0.53-6.3 ng/L and 1.4-16 ng/L, respectively depending on the PFAS variant. U.S. Environmental Protection Agency [2022c,b] Both methods offer LCMR well below the established health advisory levels – 70 ng/L (70 ppt) for lifetime exposure set by the USEPA in 2016 for the most commonly used and studied PFAS – perfluorooctanoic acid (PFOA) and perfluorooctanesulfonic acid (PFOS). However, the proposed legal maximum contaminant levels (MCL) and non-enforceable maximum contaminant level goals (MCLG) of PFAS in drinking water by EPA in early 2023 U.S. Environmental Protection Agency [2023] for both PFOA and PFOS (MCL = 4 ng/L and MCLG = 0 ng/L) are approaching the limits of LC-MS/MS. In addition, LC-MS/MS require expensive instruments intended for skilled analysts with complex and time consuming procedures for sample preparation, data collection and interpretation. Wang et al. [2021b] In contrast to LC-MS/MS based tools, portable sensors engineered with molecular probes for detecting PFAS are a promising alternative to standard methods with potential advantages such as lower cost, simpler operation and real-time analysis. Wang et al. [2021b], Cheng et al. [2020], Allonia [2022] However, given the vast and growing number of PFAS variants, it is a challenge to design molecular probes that can sensitively and selectively detect PFAS molecules in water Wang

et al. [2021b].

In this work, we engage this challenge by developing a high-throughput machine learning (ML) guided computational screening protocol to efficiently navigate the chemical design space to find probes with optimal sensitivity and selectivity. We target perfluorooctanesulfonic acid (PFOS) as one of the major PFAS molecules of concern due to its wide initial production and adverse effects such as low birth weight, tissue damage and cancer. Agency for Toxic Substances and Disease Registry, United States Department of Health and Human Services [2022], Meegoda et al. [2020], Buck et al. [2011] PFOS is a linear surfactant molecule with eight fluorinated carbon groups and a sulfonate head group (Fig. 6.1) that remains in an anionic state at the pH values in typical environmental water courses. U.S. Environmental Protection Agency [2022a], Meegoda et al. [2020], Wang et al. [2021b] While this work focuses on PFOS, our approach is equally applicable to any of the PFAS variants. The particular choice of PFOS is made because it is one the earliest PFAS known for its toxic properties. Agency for Toxic Substances and Disease Registry, United States Department of Health and Human Services [2022]. The fluorinated PFOS tail is hydrophobic while the anionic sulfonate head group is hydrophilic and elevates solubility in water. Meegoda et al. [2020] We search for optimal PFOS probes over a design space consisting of 3850 linear hydrogenated and halogenated molecules of length ranging from three to fourteen carbons with and without an amine or phosphine based head group (Fig. 6.1). We are motivated to define our design space as such based on the synthetic accessibility of these molecules and their similar chemical nature to PFOS, wherein we can exploit hydrophobic binding to the aliphatic tail and potentially favorable electrostatic interactions with a head group. For synthesizability reasons, we consider only semi-halogenated probes wherein the halogenated groups form contiguous blocks of fluorinated/chlorinated/brominated methylenes. While some of these probe molecules show a strong resemblance to PFAS, they will be tethered to surfaces and deployed at such small scales as not to themselves represent any significant threat to the PFAS contamination challenge they are designed to help solve in their potential future uses within portable sensors.

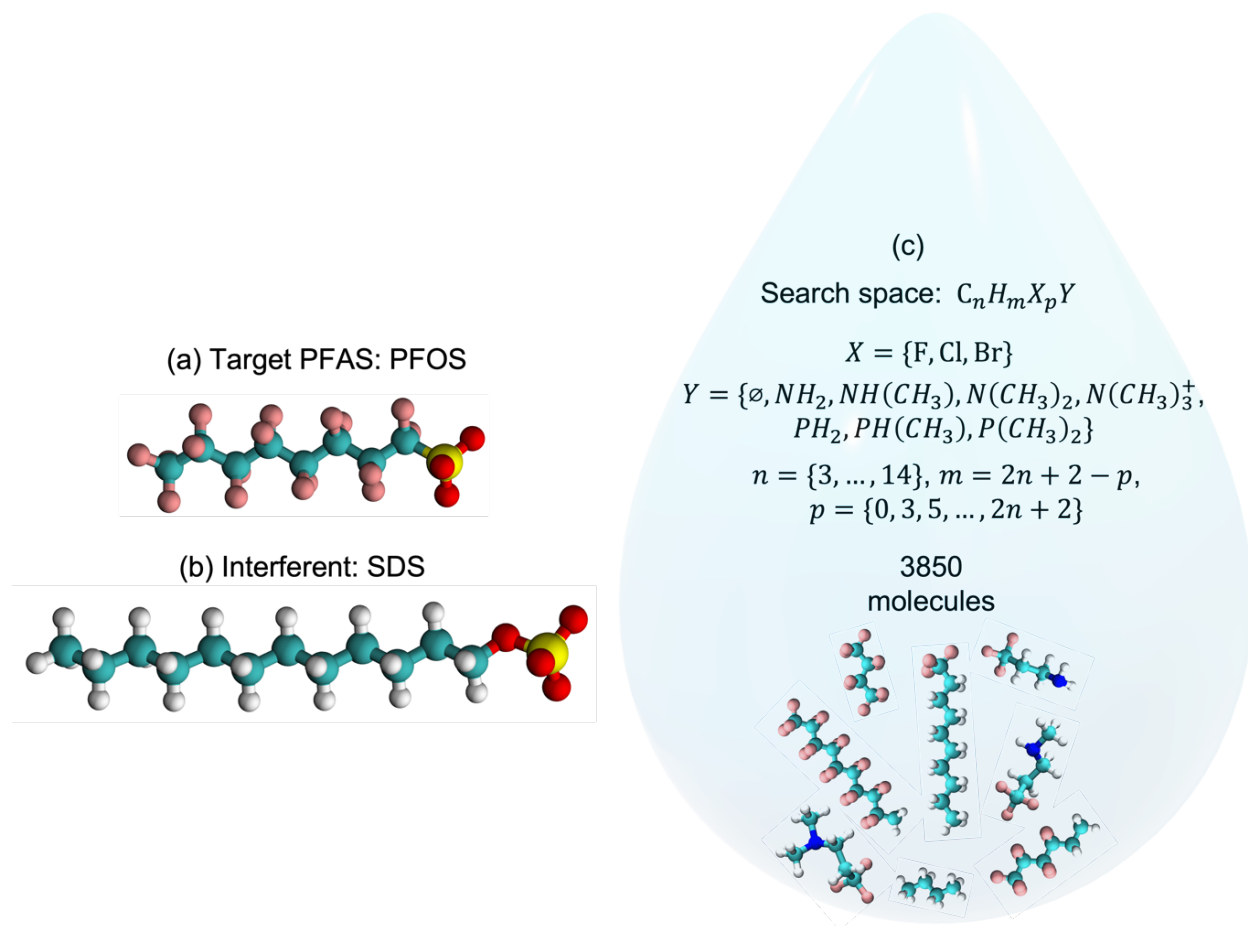


Figure 6.1: Illustrations of the target PFAS analyte, interferent, and molecular search space over which the probes with optimal sensitivity and selectivity are discovered. (a) 3D structure of the target PFAS variant (perfluorooctanesulfonic acid, PFOS), (b) 3D structure of the template interferent (sodium dodecyl sulfate, SDS), and (c) search space comprising linear hydrogenated or halogenated chains of three to fourteen carbons with and without amine or phosphine based head groups. All semi-halogenated probes considered contain contiguous blocks of halogenated methylene groups for experimental characterizability. We explored primary, secondary, tertiary and quaternary states for the amine based head group and only the primary, secondary and tertiary states of the phosphine head groups. In total, our search space comprises 3850 molecules over which we employ our computational active learning and high-throughput screening approach to discover optimal molecular probes. The 3D structures are rendered using VMDHumphrey et al. [1996], where carbon, hydrogen, oxygen, nitrogen, sulfur, fluorine are colored in cyan, white, red, blue, yellow, and pink, respectively.

Having specified PFOS as our target analyte, we can compute the *sensitivity* of candidate molecular probes in our design space by measuring the PFOS-probe binding free energy, ΔG_{PFOS} . Additionally, we adopt sodium dodecyl sulfate (SDS) as an interferent molecule representative of the

diversity of other potential absorbants that may be present in a water sample (Fig. 6.1). This allows us to quantify the *selectivity* of candidate probes via the relative PFOS-probe and SDS-probe binding free energies, $\Delta\Delta G_{PFOS-SDS} = \Delta G_{PFOS} - \Delta G_{SDS}$. It is computationally intractable to measure the selectivity of a candidate probe against all possible competing interferents, so we selected SDS as a representative interferent that is often found in waterways with PFAS. The chemical similarity of PFOS and SDS means that it is expected to be a challenging task for a molecular probe to discriminate the PFOS analyte from the SDS interferent based on differential thermodynamic binding affinity.

It is the goal of our computational search to simultaneously minimize ΔG_{PFOS} and $\Delta\Delta G_{PFOS-SDS}$ – or, equivalently, simultaneously maximize $(-\Delta G_{PFOS})$ and $(-\Delta\Delta G_{PFOS-SDS})$ – to discover highly sensitive and selective probe candidates for experimental testing. To achieve this, we employ a computational active learning approach involving deep representational learning of the molecular probes, training of surrogate Gaussian process regression models, and the application of a multi-objective Bayesian optimization with random scalarizations to identify Pareto optimal probes with respect to sensitivity and selectivity. We train our active learning platform over 10 rounds and ~ 250 probe molecules ($\sim 6\%$ of the design space) to identify probes on the sensitivity-selectivity Pareto frontier with sensitivities as high as $(-\Delta G_{PFOS}) = (9.8 \pm 0.2)$ kJ/mol, with a corresponding binding constant of $K_b^{PFOS} = (177.4 \pm 12.7)$ M⁻¹, and selectivities as high as $(-\Delta\Delta G_{PFOS-SDS}) = (3.1 \pm 1.5)$ kJ/mol, with a corresponding binding constant ratio of $K_b^{PFOS}/K_b^{SDS} = (4.6 \pm 1.7)$. A retrospective analysis of our data reveals that marginal increases to chain length provide the significant increase in sensitivity, whereas the incorporation of amine head groups and halogenation can improve sensitivity and selectivity of probes. However, there is no clear trend as to which head group and halogenation combination results in better sensitivity and selectivity for a given length of hydrogenated or fluorinated probe. A C₁₁Br₂₃P(CH₃)₂ probe containing 11 backbone brominated carbons and a tertiary phosphine head group possesses the highest computed sensitivity to PFOS within the design space, and a semibrominated probe C₅H₁₁C₇Br₁₄N(CH₃)₂ containing twelve

backbone carbons and a tertiary amine head group possesses the highest selectivity.

6.3 Methods

We employ a machine learning guided approach to efficiently navigate the molecular design space to discover probes with high sensitivity and selectivity. This approach is similar to the active learning frameworks we have employed in recent high-throughput screening campaigns for self-assembling π -conjugated peptides Shmilovich et al. [2020], switchable nanostructured materials Dasetty et al. [2022] and small organic molecules to selectively permeate cardiolipin membranes Mohr et al. [2022]. Fig. 6.2 illustrates our employed active learning framework that involves four key components: (i) all-atom molecular dynamics (MD) simulations and enhanced sampling using parallel bias metadynamics to estimate sensitivity ($-\Delta G_{PFOS}$) and selectivity ($-\Delta\Delta G_{PFOS-SDS}$), (ii) encoding of probes in search space using a pre-trained variational autoencoder with Self-Referencing Embedded Strings (SELFIES) Krenn et al. [2020] representation, (iii) construction of Gaussian process regression (GPR) surrogate models over the collected simulation data to predict sensitivity and selectivity for the encoded probes in search space, and (iv) multi-objective Bayesian optimization (BO) using random scalarizations using the trained GPR surrogate models to identify the next round of probes in search space with optimal sensitivity and selectivity.

6.3.1 Estimation of sensitivity ($-\Delta G_{PFOS}$) and selectivity ($-\Delta\Delta G_{PFOS-SDS}$)

All-atom molecular dynamics simulations of PFOS-probe and SDS-probe systems

We perform all-atom MD simulations of PFOS-probe and SDS-probe interactions accelerated with parallel bias metadynamics Pfaendtner and Bonomi [2015] (Fig. 6.2a). Although, constant pH MD simulations de Oliveira et al. [2022] could dynamically capture the protonation state of a molecule, we fix the protonation state and perform standard MD simulations because of their low computational cost and higher computational efficiency that are paramount for our high-

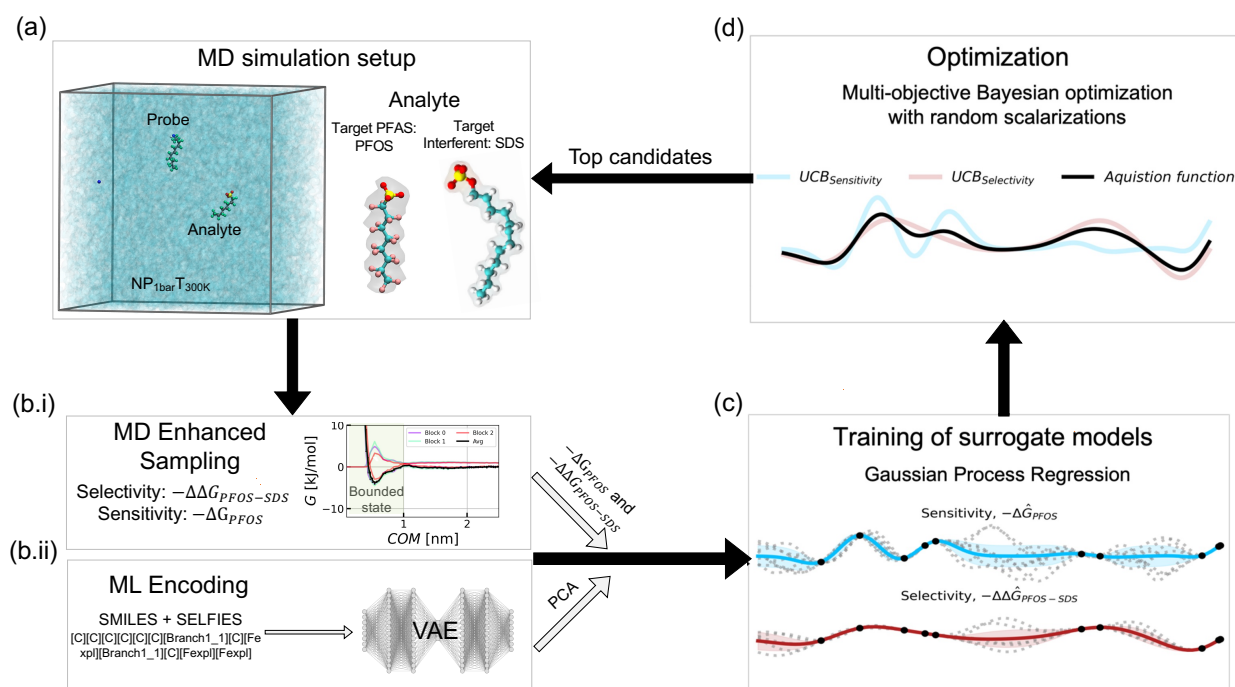


Figure 6.2: Illustration of the computational active learning approach employed to identify linear molecular probes with optimal sensitivity and selectivity. (a) Set up of all-atom and explicit molecular dynamics simulations and enhanced sampling using parallel bias metadynamics to estimate sensitivity ($-\Delta G_{PFOS}$) and selectivity ($-\Delta\Delta G_{PFOS-SDS}$), (b) embedding sensitivity and selectivity information generated through (b.i) performance of these enhanced sampling simulations and (b.ii) an encoding of molecular probes into a smooth low-dimensional latent space using a pre-trained variational autoencoder, (c) construction of Gaussian process regression (GPR) surrogate models over the latent space to predict probe sensitivity and selectivity, (d) multi-objective Bayesian optimization (BO) using random scalarizations to guide selection of the next molecular probes within the design space for molecular simulation.

throughput virtual screening. We model PFOS in its anionic state Cheng et al. [2009], Lindim et al. [2016] because of its extremely low pKa Lampert et al. [2007] (<1) and consider a target pH range of 6.5-8.5 corresponding to typical tap and drinking water sources. Yin et al. [2023], City of Chicago, Department of Water Management - Bureau of Water Supply [2023], WHO [2022] Similarly, we model SDS in anionic state that has a low pKa Moosavi-Movahedi et al. [2005] of 3.3. Details on the generation of initial structures in PDB format of PFOS, SDS, and the probes are provided in the Supporting Information.

Calculations to determine partial charges are performed using the restrained electrostatic potential

(RESP) method Bayly et al. [1993] with Gaussian 16 Rev A.03 Frisch et al. [2016] and force field parameters are taken from General Amber Force Field (GAFF) Wang et al. [2004] by antechamber Wang et al. [2006]. Specifically, we performed geometry optimization and the partial charge calculations in vacuum using density functional theory (DFT) at B3LYP/6-31G(d) basis level following the conventional RESP procedure Bayly et al. [1993]. A more computationally burdensome but more physically accurate scheme for partial charge assignments would employ a solvation model in the charge calculation Marenich et al. [2009], Ehlert et al. [2021], Marenich et al. [2013]. System topology generation in GROMACS Abraham et al. [2015] format is facilitated by ACPYPE Sousa da Silva and Vranken [2012]. Each simulation system comprises a single probe and a single target analyte in water, which can be interpreted as a system at infinite dilution concentration limit. This follows the primary objective of this work in designing probes for detecting target PFAS at extremely low target concentrations enforced by EPA.

Each system was first energy minimized and then equilibrated for 1 ns in a NVT ensemble at 300 K followed by a 1 ns equilibration in a NPT ensemble at 300 K and 1 bar. MD simulation parameters applied during energy minimization and the equilibration runs are provided in Supporting Information at <https://pubs.acs.org/doi/10.1021/acs.jced.3c00404>.

Enhanced sampling using parallel bias well-tempered metadynamics

Accelerated sampling of probe-SDS and probe-PFOS interactions makes high throughput simulation of targeted design space computationally accessible by reducing simulation costs while ensuring good sampling of the intermolecular free energy landscape. In this work, we employ parallel bias metadynamics (PBMetaD) method Pfandtner and Bonomi [2015] to accelerate the sampling of probe-SDS and probe-PFOS interactions and estimate their equilibrium binding free energies and binding constants. Details on the PBMetaD procedure, the choice of collective variables (CVs), calculation of the potential of mean force curves as a function of the center of mass separation $PMF(r)$, and binding free energies ΔG between the probes and PFOS or SDS, are provided in

Supporting Information.

All the molecular simulations with parallel bias metadynamics are performed using the GROMACS-2018.6 suite Abraham et al. [2015] with PLUMED-2.5.2 Bonomi et al. [2009] library. These enhanced sampling calculations took ~ 3 -5 days for each PFOS or SDS and probe system depending on the system size on a shared supercomputing node utilizing $1 \times$ V100 GPU and $20 \times$ CPU cores of Intel Skylake processor.

6.3.2 *Encoding of probe molecules using pre-trained variational autoencoder with SELFIES representation*

Molecular design space

In this work, we focus our search for optimal probes by considering linear hydrogenated and halogenated molecules with the number of carbons ranging from three to fourteen (Fig. 6.1). We consider three halogens — fluorine, chlorine, and bromine for the halogenated molecules with a constraint that the halogenated methyl or methylene groups are contiguous along the backbone of the probe molecule. The halogenated methylene and methyl groups were primarily considered to take advantage of the halogen mediated interactions between the probe and the target PFAS Tsuzuki and Uchimarui [2020], Robalo et al. [2019, 2020]. For each of the hydrogenated or halogenated molecules, we also include amine and phosphine based head groups to utilize electrostatic interactions for binding the head group of probe with the anionic sulfonate head group of PFOS. These head groups replace terminal hydrogenated or halogenated methyl group in a probe. For amine head groups, we include primary (NH_2), secondary ($\text{NH}(\text{CH}_3)$), tertiary ($\text{N}(\text{CH}_3)_2$), and quaternary ($\text{N}(\text{CH}_3)_3^+$) states. For phosphine head groups, only primary (PH_2), secondary ($\text{PH}(\text{CH}_3)$), and tertiary ($\text{P}(\text{CH}_3)_2$) states are considered. Collectively with the head group combinations of the linear hydrogenated and halogenated molecules, there are 3850 molecules in our search space (Fig. 6.1).

Molecular featurization and embedding using a pre-trained small molecule variational autoencoder (VAE)

The variational autoencoder (VAE) Kingma and Welling [2022], Gómez-Bombarelli et al. [2018] model utilized in this study was previously trained on more than 1.2M small molecules contained within the ZINC dataset Irwin and Shoichet [2005 Jan-Feb], Irwin et al. [2012, 2020] plus a number of common chemical screening libraries. Tang et al. [2023] It was our anticipation that this pre-trained VAE over a large class of small molecules would provide good representations of the 3850 linear probe molecules considered in this work as rich but interpretable featurizations appropriate for our active learning search. The training process first requires unique representation of the molecules using SELFIES Krenn et al. [2020] and optimization of a loss function to learn a continuous representation of these molecules within a low-dimensional latent space constituting the information bottleneck layer between the encoder and decoder. This latent space provides a smooth and low-dimensional representation of the molecular design space that is well-suited to the construction of surrogate structure-property models and enables a Bayesian-guided traversal and optimization of molecules within the design space Gómez-Bombarelli et al. [2018], Shmilovich et al. [2020], Mohr et al. [2022], Tang et al. [2023]. We provide a brief description of the VAE model construction and training in the Supporting Information; full details are available in Tang et al. [2023].

The 3850 candidate probe molecules in the design space were created and stored as SMILES strings using RDKit Landrum et al. [2023]. We use selfies 1.0.4 Krenn et al. [2020] to encode each molecule into SELFIES representation using the generated SMILES strings. The molecules in SELFIES representation are then projected onto the molecular latent space of the pre-trained VAE model (Fig. 6.2b.ii). For this, we apply the same one-hot encoding SELFIES representation used during training process of the VAE model. To reduce the 100-dimensional latent space of molecules in search space, we then apply principal component analysis (PCA) Jolliffe and Cadima [2016]. The

top 5 principal components (PC) capture more than 95% of the cumulative variance of the search space molecules in the latent space, allowing us to further compress the latent space representation for the purposes of GPR training and BO active learning with limited loss of information. We also verify that the 3850 linear probe molecules are smoothly embedded into the five leading PCs of the latent space, exhibiting smooth transitions in key molecular properties such as molecular weight and number of carbon or halogen atoms in the probe. This implies that the molecules in a given neighborhood of latent space have similar properties and might therefore also be expected to have similar observables such as binding affinity to a target analyte. The present work considers a finite design space of 3850 candidate molecules that we project into the VAE latent space to provide a featurization of these molecules suitable for our active learning search. As such, we do not exploit the generative capacity of the VAE decoder to produce novel molecules conditioned on a particular location in the latent space, although note that this capability could be useful for expanding the search into new regions of molecular design space.

6.3.3 Construction of Gaussian process regression (GPR) surrogate models

We train the GPR surrogate models Deringer et al. [2021] to predict $(-\Delta G_{PFOS})$ and $(-\Delta\Delta G_{PFOS-SDS})$ as a function of the 5D vector x specifying the projection of each candidate probe molecule into the leading five principal components of the VAE latent space (Fig. 6.2c). We construct our GPR kernel using the widely used and infinitely differentiable radial basis function (RBF) for the GPR covariance function. The RBF length scale parameter l that sets the distance for two points to be correlated is treated as hyperparameter that is optimized separately for each of the five dimensions of x during training by maximizing the log marginal likelihood Rasmussen and Williams [2005]. The bounds of RBF length scale parameter l during optimization were set to 0.001 and 200, which are close to the minimum and maximum distances between the search set molecules along the top five principal components. The GPR models are trained using scikit-learn Pedregosa et al. [2011]. The models are trained over the $(-\Delta G_{PFOS})$ and $(-\Delta\Delta G_{PFOS-SDS})$ values collected

for all probes to date, and return predictions of an estimated mean μ and standard deviation σ for $(-\Delta G_{PFOS})$ and $(-\Delta\Delta G_{PFOS-SDS})$ for any vector x . In this manner, the GPR serves as a surrogate model for the sensitivity and selectivity of new probe molecules that have not yet been simulated.

6.3.4 Multi-objective Bayesian optimization (BO)

The trained GPR surrogate models are passed to a multi-objective Bayesian Optimization (BO) protocol using random scalarizations Paria et al. [2020] (Fig. 6.2d). The aim of the BO-guided search is to prospectively identify probe molecules residing on the high-sensitivity–high-selectivity Pareto frontier. A single acquisition function for the sensitivity (i.e., maximize $(-\Delta G_{PFOS})$) and selectivity (i.e., maximize $(-\Delta\Delta G_{PFOS-SDS})$) objectives is constructed from a random scalarization of the two GPR surrogate models Shmilovich et al. [2022b], Paria et al. [2020]. In this approach, two independent acquisition functions are first employed for each of the objectives and then merged to obtain a single scalarized acquisition function.

As is standard practice when employing random scalarizations, we employ upper confidence bound (UCB) acquisition functions $\alpha_{UCB,sens}(x) = \mu(-\Delta G_{PFOS}(x)) - \beta\sigma(-\Delta G_{PFOS}(x))$ and $\alpha_{UCB,select}(x) = \mu(-\Delta\Delta G_{PFOS-SDS}(x)) - \beta\sigma(-\Delta\Delta G_{PFOS-SDS}(x))$ Auer [2002] for each of the two design objectives Paria et al. [2020], Cox and John [1997], where the mean and standard deviations in the sensitivity and selectivity at a particular x are extracted from the most recently trained GPR models and β is a random variable drawn from a log uniform distribution on the range $[\log(10^{-4}), \log(10^4)]$ that controls the degree of exploitation vs. exploration Shmilovich et al. [2022b]. The two UCB acquisition functions are merged via a randomly weighted linear combination into a single scalarized acquisition function $\alpha(x) = \gamma\alpha_{UCB,sens}(x) + (1 - \gamma)\alpha_{UCB,select}(x)$ (Fig. 6.2d). Colloquially, $\alpha(x)$ provides a measure of desirability for each potential candidate molecule with embedding vector x under a random sample of the relative weighting $\gamma \sim U(0, 1)$ between the two objectives, where $U(0, 1)$ is a uniform distribution over $(0, 1)$. Under sufficient numbers of trials, γ will explore various weightings of the sensitivity-selectivity design objectives

and guide sampling to explore the entire Pareto frontier. We calculate $\alpha(x)$ for all as-yet-unsampled candidate molecules in the design space and employ the Kriging believer Shahriari et al. [2015] to perform batched selection of molecules for the enhanced sampling calculations.

We seed the active learning search process with 45 manually-selected candidates from the design space designed to comprise a diversity of molecules consisting of hydrogenated and halogenated probes of different lengths with and without amine head groups. Rather than distributing the initial probes over the search space using a diversity maximizing strategy Shmilovich et al. [2022a], we instead adopted a hypothesis driven approach motivated by experimental insights in an attempt to inject this prior knowledge into the initialization of the search. Specifically, we hypothesized that the fluorophilic interactions can be critical in binding the probe with the target PFOS containing fluorinated backbone relative to the interferent SDS that has a hydrogenated backbone. For this reason, we initially selected probes of different lengths ranging from 3 to 14 backbone carbons with different degrees of contiguous blocks of fluorination. In addition, we added amine (including cationic) head groups to the probes to explore the feasibility of stacking (electrostatic interactions) with the head (anionic) group in PFOS and SDS. For a selected few probes that are found to be optimal with respect to sensitivity and selectivity in the fluorinated probes, we studied chlorinated variants of the fluorinated probes. We performed a total of 10 rounds of active learning over the course of which we simulated a total of 252 probe molecules ($\sim 6\%$ of the 3850 candidate design space) at a cost of $\sim 25,000$ GPU-h and $\sim 5,000$ CPU-h. Approximately $N = 20$ molecules were sampled within each active learning cycle, but this number varied in each round in order to make efficient use of parallel compute resources. Specifically, molecules whose $1\ \mu\text{s}$ simulation completed within the ~ 2 week time horizon of a single active learning round participated in the active learning cycle. The remaining molecules whose simulations were incomplete at beginning of the new cycle were added to the next cycle. In addition, we added five hand selected probes with amine head group in active learning cycle 8 within a human-in-the-loop intervention into the otherwise automated search protocol. Specifically, we were interested in understanding the role of

amine head groups upon probe performance and appreciated that such molecules had heretofore been undersampled in the active learning screen. Incorporating these molecules within our screen was valuable for our retrospective analyses of the role of probe length, degree of fluorination, and presence of amine head groups upon probe performance.

A notebook containing the computational active learning process with multi-objective BO is available at <https://github.com/Ferg-Lab/activeLearningPFASLinear.git>. Sensitivities ($-\Delta G_{PFOS}$), selectivities ($-\Delta\Delta G_{PFOS-SDS}$), and equilibrium binding constants of the probes explored in each round of the active learning search along with their 2D structures, molecular weights, and IUPAC names. In addition, the SMILES strings, sensitivities, selectivities, and equilibrium binding constants (K_b^{PFOS} , K_b^{SDS}) of each cycle are available in machine-readable format at

<https://github.com/Ferg-Lab/activeLearningPFASLinear.git> along with helper notebooks and scripts to help facilitate ready uptake and use by the community.

6.3.5 Training of LASSO regression models to infer important chemical substructures of probes

We train and analyze LASSO (Least Absolute Shrinkage and Selection Operator) regression models to extract important chemical substructures of probes that correlate with their sensitivity and selectivity to PFOS. In contrast to the relatively opaque GPR models, these simple and interpretable linear models are commonly applied for feature selection and to understand substructures or sub-graphs that play an important role in predicting observed molecular properties and responses. Bhattacharjee and Vlachos [2020], Mohr et al. [2022] LASSO regression employs standard least squares regressions with an L1 regularization over the weights to minimize the cost function $\ell = \sum_i^N (y_{true,i} - \sum_j^K w_j x_{ij})^2 + \lambda \sum_j^K ||w_j||_1$, where $y_{true,i}$ corresponds to the measured response, x_{ij} corresponds to the value of the j^{th} feature of a sample i , w_j is the linear regression coefficient

accorded to feature j within the LASSO model, N is the number of samples, K is the number of features, and λ is a hyperparameter controlling the strength of the L1 regularization. The hyperparameter λ is typically tuned by cross-validation. The effect of the regularization term is to shrink the regression coefficients of the least important features to zero, thereby inducing sparsity in the model to include only those features with the greatest explanatory power. The magnitude and sign of the weights w_j lend themselves to simple interpretability as to the strength and direction of the effect of that feature upon the response.

We train separate LASSO regression models to predict the measured negative sensitivity ($y_{true,i} = \Delta G_{PFOS}^i$) or negative selectivity ($y_{true,i} = \Delta \Delta G_{PFOS-SDS}^i$) of the 252 simulated probe molecules using an input featurization enumerating the presence or absence of particular chemical substructures (i.e., molecular subgraphs) within each molecular probe. We follow the procedure described by Bhattacharjee and Vlachos [2020] to extract the substructures of the molecular probes explored in active learning process using RDKitLandrum et al. [2023] by setting the minimum and maximum edge lengths as 1 and 8, respectively. Hydrogen atoms are not considered part of the substructures. We note that the substructures mined consider only edge connectivity between atoms but not the atom properties such as their charge or amine state. A matrix with each molecule and the count of each unique substructure mined from RDKitLandrum et al. [2023] as rows and columns, respectively are used as features for training the LASSO regression models. Standard scalers in scikit-learnPedregosa et al. [2011] are applied to normalize the features or matrix representing substructures count in each molecule prior to training. The hyperparameter λ in the LASSO regression model is tuned by 10-fold cross validation to minimize the validation prediction error on the validation set. The trained models are evaluated on a randomly-selected 20% hold-out test set that was not exposed to the model during any part of training.

6.4 Results and Discussion

6.4.1 Discovery of optimal linear molecular probes using computational active learning

Our goal is to discover probes in the design space of 3850 linear molecules those possessing optimal sensitivity and selectivity to PFOS (Fig. 6.1). We efficiently traverse the design space using our computational active learning pipeline (Fig. 6.2) that we seed with an initial batch of 45 probes and execute 10 cycles of active learning sampling $N = 13$ -29 molecules per round. To assess convergence of the active learning campaign, we calculate,

$$d_i = \sqrt{(\text{sensitivity}_i)^2 + (\text{selectivity}_i)^2} = \sqrt{(-\Delta G_{PFOS}^i)^2 + (-\Delta \Delta G_{PFOS-SDS}^i)^2}, \quad (6.1)$$

as the distance of each probe i from the origin in the 2D sensitivity-selectivity design space. Recalling that it is our objective to simultaneously maximize sensitivity ($-\Delta G_{PFOS}$) and selectivity ($-\Delta \Delta G_{PFOS-SDS}$), this presents a quantitative measure of the advancement of the active learning screen in discovering desirable molecular probes. We further compute,

$$d_{all}^{(j)} = \overline{d}_i, \quad i \in \{(0), \dots, (j)\}, \quad (6.2)$$

$$d_{Pareto}^{(j)} = \overline{d}_i, \quad i \in \text{Pareto}_{\{(0), \dots, (j)\}}, \quad (6.3)$$

where $d_{all}^{(j)}$ reports a cumulative average over the d_i values of all probes sampled in the campaign so far up to and including cycle j , and $d_{Pareto}^{(j)}$ reports an average over the d_i values of those probes constituting the Pareto frontier at the end of round j . We present in Fig. 6.3c scatter plots showing the location of the sampled probes in each cycle of the campaign within the sensitivity-selectivity design space and the advancement of the Pareto frontier. The trends in Fig. 6.3 indicate convergence of our active learning search by approximately cycle 6, beyond which we do not observe any further improvements in $d_{all}^{(j)}$ or $d_{Pareto}^{(j)}$, or further advancement of the Pareto frontier. This suggests that

the search has identified the top performing molecules within the molecular design space and that we may not expect the discovery of significantly superior candidates with additional cycles of the search.

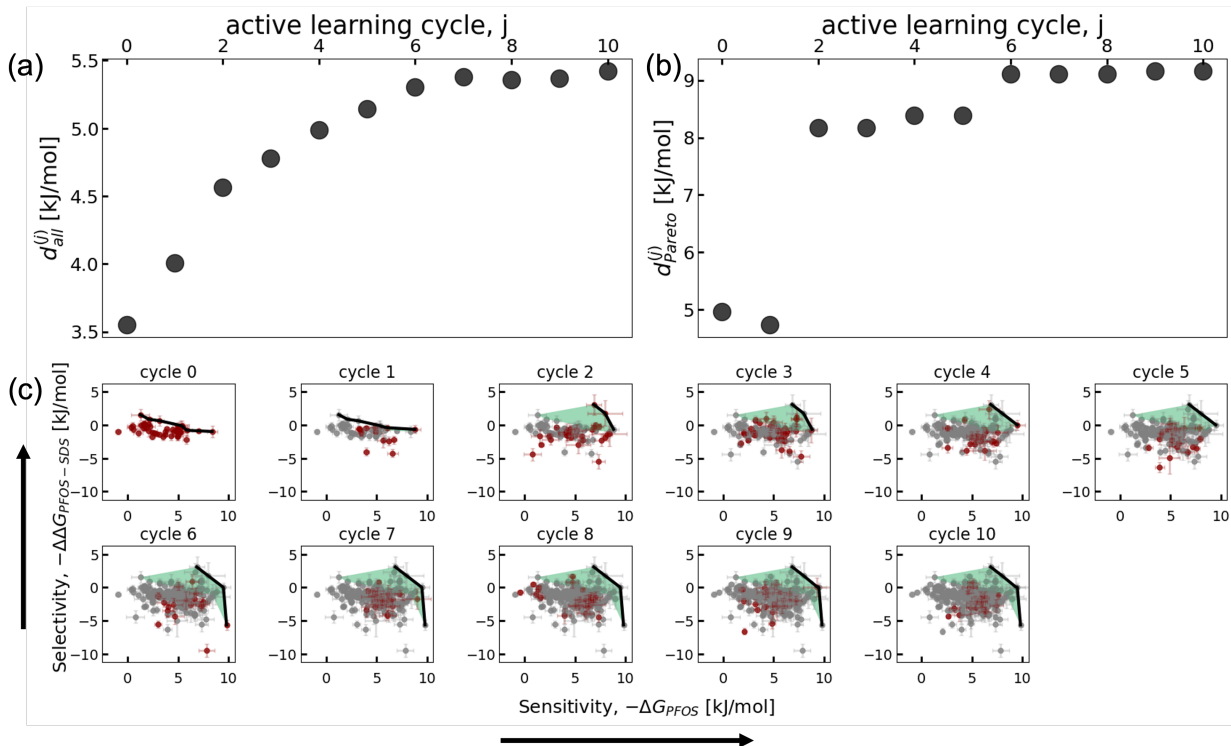


Figure 6.3: Progress of the active learning campaign in advancing the sensitivity-selectivity frontier of PFOs probes within each active learning cycle. (a) Mean distance $d_{all}^{(j)}$ reporting the cumulative average over the d_i values of all n probes sampled so far up to and including cycle j . (b) Pareto distance $d_{Pareto}^{(j)}$ reporting an average over the d_i values of those probes constituting the Pareto frontier at the end of round j , including contributions to the frontier from all prior rounds. (c) Scatter plots of the sampled probes within the sensitivity-selectivity design space over the course of the active learning campaign. The sensitivity and selectivity increase in the positive x- and y-directions, respectively. The probes sampled in the present round are indicated by red markers and those accumulated from all previous rounds by grey markers. The Pareto frontier is indicated by the black line. The green shaded region highlights the shift in the Pareto frontier relative to cycle 0. The molecules in cycle 0 contains the 45 molecules used to train the initial GPR model.

We present in Fig. 6.4 violin plots illustrating the cumulative sensitivity and selectivity distributions of all probes sampled over the course of the active learning campaign. In terms of sensitivity, we see round-on-round improvements in the top candidate identified up to cycle 6, beyond which

no further advances are observed. In terms of selectivity, we see improvements only through cycle 2, where the top candidate is identified. The initial 45 probes (cycle 0) have a broad range of sensitivities – (-0.9 ± 0.3) kJ/mol to (8.4 ± 0.5) kJ/mol (Fig. 6.4a) – but their selectivities span a relatively narrow range – (-2.2 ± 0.6) kJ/mol to (1.5 ± 0.9) kJ/mol (Fig. 6.4b). A fully chlorinated probe $C_{12}Cl_{26}$ with 12 backbone carbons has the highest sensitivity to PFOS among all the studied probes in cycle 0. However, this probe has lower selectivity to PFOS than a probe $C_3H_7NH_2$ that has three backbone carbons and one primary amine head group, which has the highest selectivity among all the probes in cycle 0. We discovered three more probes with improved sensitivity in cycles 1, 4, and 6. These include a semi-brominated probe $C_7Br_{15}C_5H_{10}P(CH_3)_2$ with 12 backbone carbons and a tertiary phosphine head group, a completely chlorinated probe $C_{11}Cl_{23}N(CH_3)_2$ with 11 backbone carbons and a tertiary amine head group, and a completely brominated probe $C_{11}Br_{23}P(CH_3)_2$ with 11 backbone carbons and a tertiary phosphine head group. A semi-brominated probe $C_5H_{11}C_7Br_{14}N(CH_3)_2$ with twelve backbone carbons and a tertiary amine head group has the highest selectivity among all the probes discovered in the active learning cycles.

To facilitate contact with experimental measurements, we compute the equilibrium binding constant K_b from our calculated free energy profiles by numerically evaluating the following integral Li and Voth [2021], Raniolo and Limongelli [2020],

$$K_b = C^0 \int_0^{r_{\text{bound}}} 4\pi r^2 \exp(-PMF(r)/k_B T) dr, \quad (6.4)$$

where $C^0 = 1/1661 \text{ \AA}^{-3}$ is the standard state concentration at 1 mol/L, and the cutoff of $r_{\text{bound}} = 1$ nm delimits the bound and unbound states, motivated by the 1 nm cutoff of the van der Waals and real-space electrostatic interactions implemented in our simulations.

How are the top performing probes identified in our screen predicted to perform in practice? With regards to sensitivity, adopting a target PFOS concentration of $4 \text{ ng/L} \approx 8 \times 10^{-12} \text{ mol/L}$ mandated

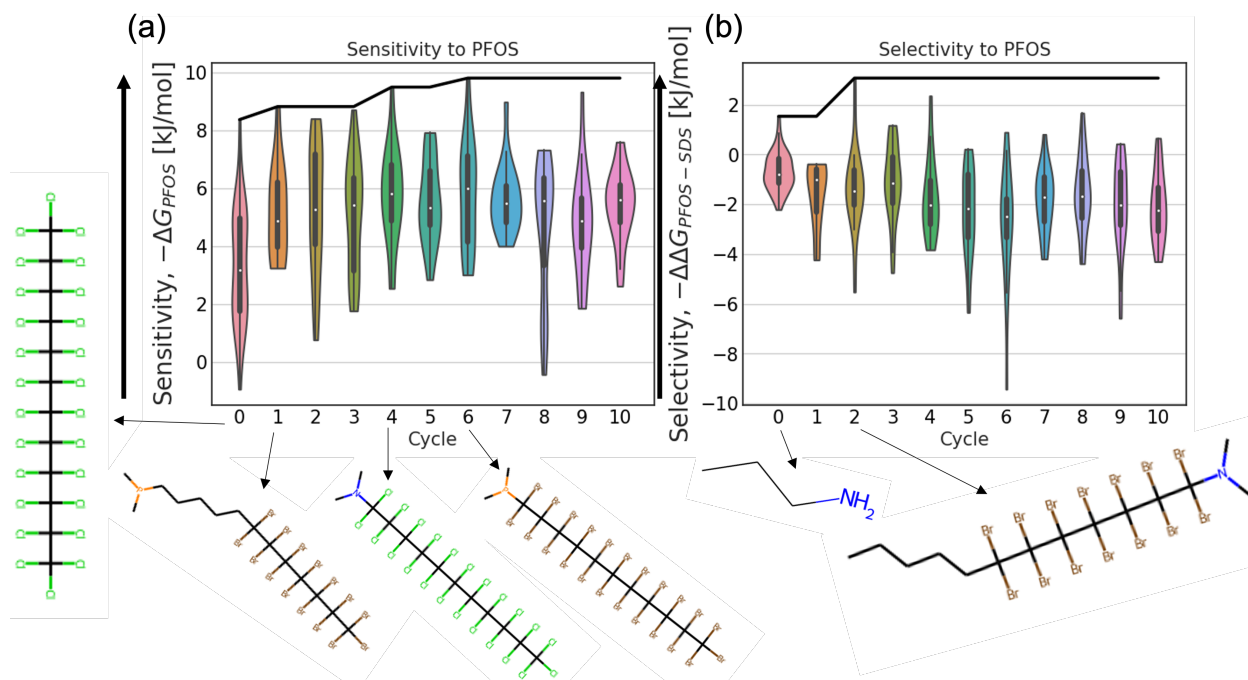


Figure 6.4: Change in the distribution of sensitivity and selectivity of probes to PFOS with each active learning cycle. Violin plots of the (a) sensitivity ($-\Delta G_{PFOS}$) and (b) selectivity ($-\Delta\Delta G_{PFOS-SDS}$) as a function of the active learning cycle. Chemical structures of molecules with maximum sensitivity and selectivity among all the probes studied up to a given cycle are shown at the cycles in which they were discovered. The black solid line shows the change in the maximum sensitivity or selectivity of probes to PFOS with each active learning cycle. Both sensitivity and selectivity increase with increase in positive direction of y-axis.

by EPA guidelines for safe drinking water U.S. Environmental Protection Agency [2023] and assuming a 10-fold higher initial free PFOS concentration and negligible interferents present, we can invert the relationship $K_b^{PFOS} = C^0 C_{\text{probe-PFOS}} / C_{\text{probe}} C_{PFOS}$ to estimate that the most sensitive probe discovered in our screen with $K_b^{PFOS} = (177.4 \pm 12.7) \text{ M}^{-1}$ would require to be present at a concentration of $C_{\text{probe}} \approx 0.05 \text{ mol/L}$. It is conceivable that such local probe concentrations may be achievable by surface immobilization within a sensing device, but substantially lower concentrations may be expected for previously reported high-affinity molecular host-guest systems such as β -cyclodextrin ($K_b \approx 10^4$ - 10^5 M^{-1} , $C_{\text{probe}} \approx 0.1$ - 1 mmol/L) Weiss-Errico and O'Shea [2017b], Zheng et al. [2019] or guanidinocalix[5]arenes ($K_b \approx 10^7 \text{ M}^{-1}$, $C_{\text{probe}} \approx 1 \text{ } \mu\text{mol/L}$) Zheng et al. [2019]. Assuming that our active learning search has identified the top performing candidates

within the molecular design space, these results indicate that this class of linear molecules does not contain highly sensitive molecular probes for PFOS detection. With regards to selectivity, assuming equimolar initial free concentrations of PFOS and SDS of 8×10^{-11} mol/L, the most selective probe in our screen with $K_b^{\text{PFOS}} = (62.7 \pm 15.3) \text{ M}^{-1}$, $K_b^{\text{SDS}} = (13.6 \pm 3.8) \text{ M}^{-1}$, and $K_b^{\text{PFOS}}/K_b^{\text{SDS}} = (4.6 \pm 1.7)$, would require a concentration of $C_{\text{probe}} \approx 0.1$ mol/L to achieve the EPA target PFOS concentration and would produce a bound PFOS to SDS relative enrichment of $C_{\text{probe-PFOS}}/C_{\text{probe-SDS}} \approx 1.4$. This suggests that the linear molecular candidate space does not contain highly selective molecular probes, although we observe that even modest selectivities can be exploited within multiplexed sensing schemes Alzate-Carvajal et al. [2023].

6.4.2 *Identification of chemical substructures that are the principal determinants of sensitivity and selectivity*

Having efficiently navigated the search space using our active learning protocol to discover the candidates populating the sensitivity-selectivity Pareto frontier, we now seek to infer the chemical features underpinning the calculated sensitivity and selectivity of the molecular probes to PFOS. This understanding can both help rationalize the trends observed within our active learning search and help guide and inform the subsequent design and exploration of augmented libraries engineered to be enriched in molecules with promising chemical features. To do so, we trained LASSO (Least Absolute Shrinkage and Selection Operator) regression models as interpretable “glass box” predictive models to identify the chemical substructures within the molecular probes that are the principal determinants of PFOS sensitivity and selectivity. We present in Fig. 6.5 the LASSO regression weights w associated with each chemical substructure. The magnitude of these weights can be interpreted, under the linear LASSO model, as the relative importance of that substructure in determining sensitivity or selectivity. Negative weights indicate molecular substructures that tend to promote elevated sensitivity ($-\Delta G_{\text{PFOS}}^i$) or selectivity ($-\Delta \Delta G_{\text{PFOS-SDS}}$), whereas positive weights indicate substructures that tend to be detrimental to sensitivity or selectivity.

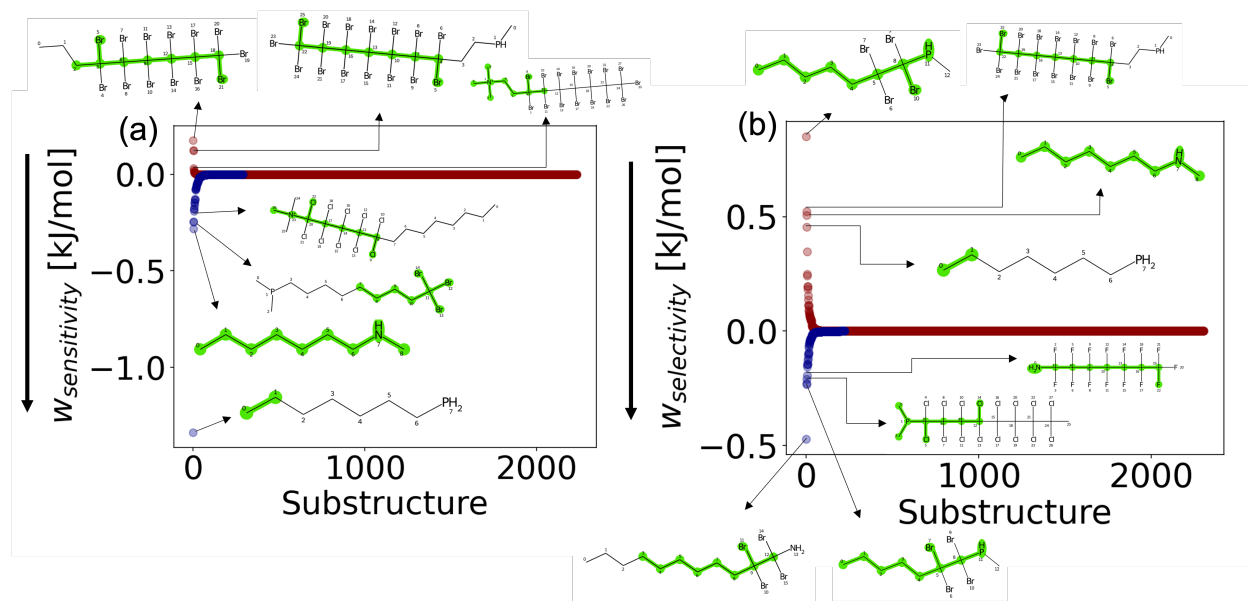


Figure 6.5: Chemical substructures within the 252 molecular probes considered within our active learning search identified by LASSO regression to be the leading determinants of PFOS (a) sensitivity and (b) selectivity. Large LASSO weights w identify substructures that are the leading determinants of sensitivity or selectivity. The chemical substructures possessing the largest LASSO weights are illustrated in green on representative probe molecules containing these substructures. The vertical arrows on the y-axes indicate the direction of favorable LASSO weights.

Analysis of the sensitivity LASSO model (Fig. 6.5a), indicates that the substructure with the most negative LASSO weight, and therefore the chemical motif that most promotes high PFOS sensitivity, is the C—C bond. The next most three most favorable substructures are a hydrogenated probe with seven backbone carbons and a secondary amine head group, a five backbone carbon substructure with one brominated methyl group, and a chlorinated substructure with an amine based head group. The top 20 substructures that promote favorable sensitivity include hydrogenated and halogenated substructures containing amine and phosphine head groups. Taken together, these results indicate that the chain length (i.e., number of C—C bonds) and the presence of chlorinated or fluorinated heads or tails, and brominated head group appear to elevate PFOS sensitivity. Turning to the positive LASSO weights, we identify a rather small number of significant chemical substructures that are detrimental to sensitivity, including two brominated substructures with seven backbone carbons and a semi-brominated substructure with quaternary amine head group. The remaining top seven

substructures that promote unfavorable sensitivity include substructures with chlorinated methyl groups and primary phosphine head groups. These results appear to indicate that bromination of the tail and the inclusion of phosphine head groups are detrimental to probe sensitivity.

Analysis of the selectivity LASSO model (Fig. 6.5b), reveals a semi-brominated substructure with eight branched backbone carbons to be the most favorable in promoting selectivity of probes for the PFOS target over the SDS interferent. This is followed by a similar semi-brominated substructure but seven branched backbone carbons and a phosphine based head group. Chlorinated, fluorinated substructures with phosphine and amine based head groups are the next most favorable substructures. From the top 20 substructures that promote favorable selectivity, we identify the importance of halogenation and head groups in improving selectivity of a probe. In contrast to sensitivity, we identify a relatively large number of substructures unfavorable to selectivity. These include a semi-brominated substructure with seven branched backbone carbons and phosphine head group, a branched brominated substructure with seven backbone carbons, a hydrogenated substructure with seven backbone carbons and a secondary amine head group, and the C–C bond. This indicates that bromination and increasing chain length can be unfavorable to promoting selectivity of the molecular probes.

Considering now both sensitivity and selectivity, we observe that the third most unfavorable substructure to selectivity is the same as the second most favorable substructure to sensitivity. Furthermore, the fourth most unfavorable substructure to selectivity – the C–C bond – is the same as the most favorable substructure to sensitivity. This analysis suggests the competing nature of both chain length and certain substructures for the simultaneous optimization of both PFOS sensitivity and selectivity, and exposes an inherent challenge for this two-dimensional optimization by controlling the presence of particular chemical groups.

6.4.3 Correlation analysis of important identified chemical substructures

Our substructure analysis exposed C–C groups, halogenation, and head groups as chemical substructures as important determinants of sensitivity and selectivity. Informed by this analysis, we now perform a correlation analysis to quantify the role of these features upon sensitivity ($-\Delta G_{PFOS}$) and selectivity ($-\Delta\Delta G_{PFOS-SDS}$). In Fig. 6.6, we plot the sensitivity and selectivity dependence upon chain length for fully hydrogenated and fully fluorinated probes. For hydrogenated probes, we observe ~ 0.4 - 1.3 kJ/mol increase on average in sensitivity with each additional C–C bond (Fig. 6.6a). Fluorinated probes appear to be marginally more sensitive than their hydrogenated counterparts, but the corresponding values do not generally lie outside of standard errors. Contrariwise, there is a degradation in selectivity with increasing probe length of ~ 1 - 2 kJ/mol upon elongating the hydrogenated and fluorinated probes from four to twelve C–C bonds (Fig. 6.6b), but this trend is not monotonic. In all cases, the selectivity for SDS is higher than or equal to that for PFOS, indicating that these probes are not selective towards the desired target molecule. Again, there tend not to be statistically meaningful differences in selectivity between fully fluorinated and hydrogenated probes of the same length.

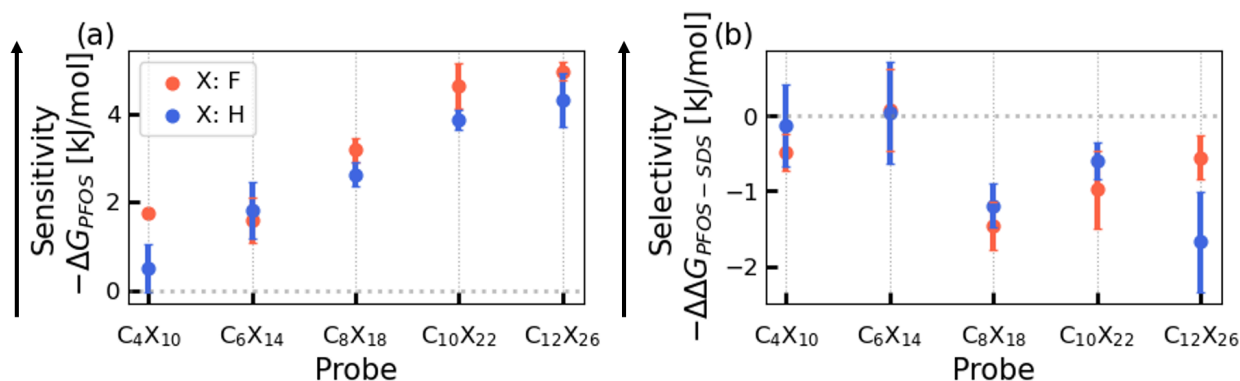


Figure 6.6: Effect of probe length upon PFOS (a) sensitivity and (b) selectivity for fully hydrogenated (blue) and fully fluorinated (red) probes. Markers and error bars represent means and standard errors estimated by three-fold block averaging over a 900 ns production run. Standard propagation of errors is employed to obtain the standard errors on selectivity. The vertical arrows on the y-axes indicate the direction of favorable sensitivity and selectivity.

Although the PFOS target has a fluorinated tail whereas the SDS interferent has a hydrogenated

tail, our results indicate both fluorinated and hydrogenated probes have largely similar sensitivities and selectivities. This indicates the relatively weak role of fluorophilic interactions in modulating probe performance, and is consistent with the DFT-SAPT based study by Tsuzuki and Uchamaru [2020], which indicate that the dispersion forces that dominate intermolecular interactions are nearly identical between $\text{CF}_4\text{-CF}_4$, $\text{CF}_4\text{-CH}_4$, and $\text{CH}_4\text{-CH}_4$. It is only the electrostatic interactions that arise due to electronegativity effects that are higher in unlike pairs relative to like pairs, but their net contributions are smaller than dispersion forces. Tsuzuki and Uchamaru [2020] Furthermore, the interactions between the probes and PFOS and SDS can be largely entropic with a significant role played by the solvent. For example, fluorination of hydrocarbons can increase hydrophobicity Wilhelm et al. [1977] and alter the hydration shell structure of fluorinated methyl groups Robalo et al. [2019]. Complementary work shows that these changes are nonmonotonic in methyl groups with increasing fluorination Robalo et al. [2020]. In the context of the present work, we find similarly nonmonotonic trends in interaction strengths as a function of degree of fluorination.

In Fig. 6.7 we present the effect of different amine head groups on sensitivity and selectivity of fully hydrogenated and fluorinated probes of various lengths. Specifically, we consider probe molecules C_4X_{10} , C_8X_{18} , and $\text{C}_{12}\text{X}_{26}$, where $\text{X} = \text{H}$ or F , and consider replacing the the terminal CH_3 group with a primary NH_2 , secondary $\text{NH}(\text{CH}_3)$, tertiary $\text{N}(\text{CH}_3)_2$, and charged quaternary $\text{N}(\text{CH}_3)_3^+$ amines. Overall, we see relatively muted and non-monotonic trends in the influence of head group upon sensitivity and selectivity, although there is evidence for a generally favorable influence of substituting in a primary, secondary, or tertiary amine upon the sensitivity and selectivity at the C_8 probe length. There are also no clear trends in the sensitivity and selectivity difference of fully hydrogenated or fully fluorinated probes, however, addition or removal of halogen atoms and amine or phosphine head groups produces non-monotonic changes in both sensitivity and selectivity with changes to sensitivity of up to 2.5 kJ/mol. The addition of charge to the probe via the charged quaternary amine did not lead to statistically meaningful improvements in sensitivity or selectivity.

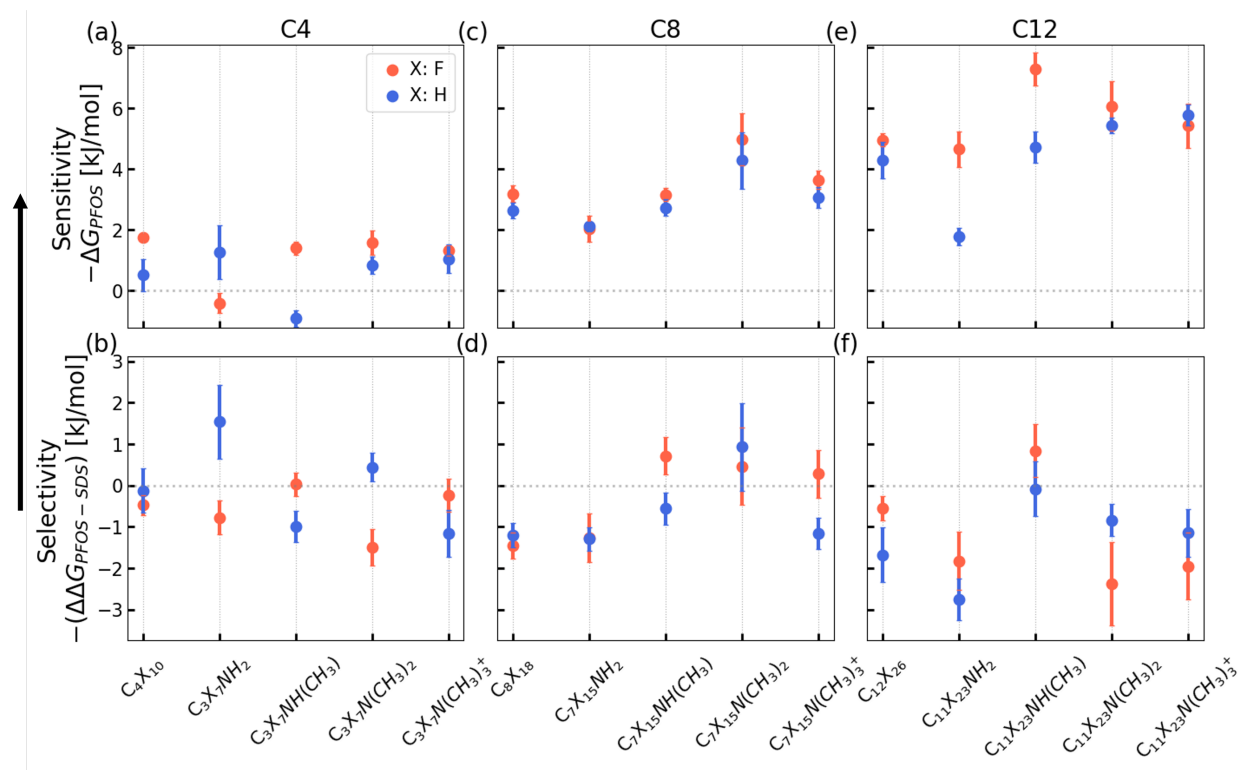


Figure 6.7: Effect of amine head group upon PFOS sensitivity and selectivity for fully hydrogenated (blue) and fully fluorinated (red) probes with a backbone length of (a) four, (b) eight, and (c) 12 single bonds. Markers and error bars represent means and standard errors estimated by three-fold block averaging over a 900 ns production run. Standard propagation of errors is employed to obtain the standard errors on selectivity. The vertical arrows on the y-axes indicate the direction of favorable sensitivity and selectivity.

6.5 Conclusions

The detection and elimination of PFAS “forever chemicals” within environmental water courses represents a pressing challenge for the design of highly sensitive and selective molecular probes. In this work, we employed a computational active learning pipeline combining enhanced sampling calculations, deep representational learning, Gaussian process regression, and multi-objective Bayesian optimization to efficiently screen molecular candidate libraries for highly sensitive and selective molecular probes. We chose to focus on a design space comprising 3850 linear hydrogenated and halogenated molecules of length three to fourteen carbons with and without an amine or phosphine based head group. The design of the library was motivated by the ready synthetic accessibility of

these candidate probes, their similar nature to PFAS, and capacity to exploit hydrophobic binding to the aliphatic tail and electrostatic binding to the head. After conducting 10 rounds of our active learning screen during which we simulated a total of 252 probe molecules ($\sim 6\%$ of the 3850 candidate design space) at a cost of $\sim 25,000$ GPU-h and $\sim 5,000$ CPU-h, we identified five candidate molecules lying on the sensitivity-selectivity Pareto frontier with sensitivities spanning $(-\Delta G_{PFOS}) = 6.9\text{-}9.8$ kJ/mol and selectivities spanning $(-\Delta\Delta G_{PFOS-SDS}) = (-5.6)\text{-}3.1$ kJ/mol. The most sensitive probe identified in our screen – a $\text{C}_{11}\text{Br}_{23}\text{P}(\text{CH}_3)_2$ molecule containing 11 backbone brominated carbons and a tertiary phosphine head group – possesses a binding constant of $K_b^{\text{PFOS}} = (177.4 \pm 12.7) \text{ M}^{-1}$ and the most selective probe – a semibrominated molecule $\text{C}_5\text{H}_{11}\text{C}_7\text{Br}_{14}\text{N}(\text{CH}_3)_2$ containing 12 backbone carbons and a tertiary amine head group possesses a binding constant ratio of $K_b^{\text{PFOS}}/K_b^{\text{SDS}} = (4.6 \pm 1.7)$. An analysis of the influence of particular chemical motifs within the probes that are the primary determinants of the observed sensitivity and selectivity expose delicate and relatively weak trends in probe length, halogenation, and head group in modulating their behaviors.

The relative paucity of highly performant linear probe molecules identified by our ML-enabled workflow militates for an expansion of the molecular search space to molecular libraries containing a richer chemical diversity of candidate molecules including branched and cyclic hydrocarbons, as well as more exotic molecules such as cyclodextrin-based probes that have garnered some attention and success in PFAS sequestration Xiao et al. [2017], Ching et al. [2020], Wang et al. [2021b], Ching et al. [2023]. Further enlargements of the design space make the computational screening platform established in this work even more valuable in performing an efficient filtration of the design space to identify the most promising molecular probes to be explored experimentally and also in excavating molecular-level understanding of probe performance that can be coupled with experimental intuition to help inform rational molecular design. The performance of our active learning pipeline in achieving convergence after considering only 6% of the molecular space provides support for its capacity to efficiently navigate and traverse large search spaces and quickly

home in on the most promising candidates.

While we focused on only one interferent in this work, it may be desirable to find probes that are selective to a given target PFAS in the presence of various types of interferents for their practical applications. To this end, one may take the top-performing probes identified in the primary screen into a lower-throughput secondary screen against a broader panel of potential interferents to better evaluate the breadth of its selectivity. In the future, it is of interest to also extend our approach for discovering molecular probes for some of the other commonly known PFAS variants United States Environmental Protection Agency [2023] such as (per) perfluorooctanoic acid (PFOA), perfluorohexane sulfonate (PFHxS), and 6:2 fluorotelomer sulfonate (6:2FTS). Similarly, it may be of interest to evaluate the performance of the top probes over a wider range of operating temperatures and pH conditions. We would also like to couple our computational search to hybrid computational/experimental active learning campaigns, in which we conduct asynchronous but simultaneous computational and experimental screens in order to integrate high-throughput computation and low-throughput experimentation to minimize the experimental costs in time and labor to identify the top performing probes for applications in PFAS sensors Shmilovich et al. [2022a].

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. DGE-2022023. This work was supported in part with funding by the University of Chicago Data Science Institute (DSI). 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. We gratefully acknowledge computing time on Frontera hosted by the Texas Advanced Computing Center (TACC) at The University of Texas at

Austin. Work by S.B.D. and Y.W. at Argonne National Laboratory was supported by the AMEWS EFRC funded by DOE, Office of Science, BES under contract DE-AC02-06CH11357.

CHAPTER 7

ESTIMATING BINDING FREE ENERGIES OF PERFLUOROOCTANESULFONIC ACID (PFOS) WITH CYCLODEXTRIN PROBES

7.1 Abstract

Per- and polyfluoroalkyl substances (PFAS), often referred to as “forever chemicals”, are ubiquitous water contaminants, posing serious risk to human well-being. With U.S. Environmental Protection Agency intake limits on PFAS molecules set at or below parts per trillion for many variants, we have limited means to detect and remove PFAS at those concentrations. Cyclodextrin-based probes have proven to be promising candidates for the development of sensing platforms due to their tunable cavity sizes and ease of functionalization; however, the optimal design of these structures for PFAS sequestration is still an open question. In this work, we examine the binding interactions of α -, β - and γ -cyclodextrin with perfluorooctane sulfonic acid (PFOS) and potential interferents trichloroacetic acid (TCAA) and sodium dodecyl sulfate (SDS) using enhanced sampling simulations. Our calculations show that PFOS binds preferentially with β -cyclodextrin with a binding constant of $2.43 \pm 0.02 \times 10^6 \text{ M}^{-1}$ - an order of magnitude greater than that of SDS and 5 orders of magnitude greater than for TCAA. This work validates the efficacy of β -cyclodextrin-based probes for PFAS detection and informs an experimental effort to build an ultrasensitive detection platform for PFOS. Moreover, the favorable binding qualities of β -cyclodextrin and molecular tunability allow for the rational design of next-generation sensing platforms.

7.2 Introduction

Per- and polyfluoroalkyl substances (PFAS) are a group of thousands of synthetic chemicals Wallington et al. [2021], U.S. Environmental Protection Agency [2022e] with highly desirable design properties from hydrophobicity to heat resistance, resulting in their widespread usage Baker and Knappe [2022], Glüge et al. [2020], Xia et al. [2022], Sim et al. [2021]. Due to the very strong (~ 440 kJ/mol Wang et al. [2021b]) carbon-fluorine covalent bonds that characterize these molecules, PFAS have lifetimes on the order of thousands of years Ghisi et al. [2019], De Silva et al. [2021], Cousins et al. [2022], Kwiatkowski et al. [2020]. Consequently, PFAS molecules have been building up in the environment for decades, permeating the water supply from agriculture to tap water. PFAS molecules have been found to increase rates of cancers, cause birth defects, and result in hormone imbalances Fenton et al. [2021], making them a target for regulatory agencies such as the United States Environmental Protection Agency (EPA) Brennan et al. [2021]

Here, we focus on molecular level detection of one such PFAS molecule, perfluorooctanesulfonic acid (PFOS). While EPA guidelines suggest a maximum contaminant level of PFOS of 4 ng/L in drinking water U.S. Environmental Protection Agency [2023], this is at the detection limit for solid-phase extraction (SPE) and liquid chromatography/tandem mass spectrometry (LC-MS/MS) Wang et al. [2021b], U.S. Environmental Protection Agency [2022d] techniques used by the EPA. Moreover, these methods are not rapidly deployable or scalable due to high costs, the requirement for skilled personnel and time consuming processes Wang et al. [2021b]. The design of rapid portable sensors for PFOS requires the identification of molecules that can selectively bind to PFOS in the face of the many interferents in water

In this work, we study cyclodextrin-based probes, a promising class of molecules whose tunable cavity sizes and ease of functionalization support efforts to selectively sequester PFAS. We compute binding energies for PFAS variant PFOS with α -, β - and γ - cyclodextrins via enhanced sampling molecular dynamics calculations. In Sect.7.3, we discuss the setup of our molecular dynamics

calculations. In Sect. 7.3.1, we discuss the all atom molecular dynamics methods used to study these molecules. In Sect. 7.3.2, we introduce funnel metadynamics used to accelerate these calculations. In Sect. 7.3.3, we discuss the estimation of binding constants from our enhanced sampling calculations and compute free energy corrections. In 7.3.3, we present the resulting binding energy calculations for target PFOS and interferents trichloroacetic acid (TCAA) and sodium dodecyl sulfate (SDS) with α -, β - and γ - cyclodextrins.

7.3 Methods

Several classical molecular dynamics (MD) simulation studies of β -cyclodextrin with Per- and Polyfluoroalkyl Substances such as PFBA (perfluorobutanoic acid) and PFOA (perfluorooctanoic acid) have been conducted Tunioli et al. [2023], Choudhary et al. [2022], however, to the best of our knowledge, studies on β -cyclodextrin-PFOS interactions are largely limited to DFT calculations for evaluating their binding interaction energies and mapping of electrostatic potentials Wang et al. [2021a]. Moreover, no simulation work has been published comparing the effects of cavity sizes on binding affinity to PFAS molecules. We conducted all-atom classical MD simulations to gain molecular-level insight into the structure and thermodynamics of the bound complex between the PFOS analyte and TCAA or SDS interferent with the α -, β -, and γ -cyclodextrin probes. Simulations were conducted using the GROMACS 2021.6 simulation suite Abraham et al. [2015] employing the General Amber Force Field (GAFF2) He et al. [2020] and the TIP3P water model Jorgensen et al. [1983a]. To improve sampling of complex formation, we combined funnel metadynamics Limongelli et al. [2013], Rizzi et al. [2021] with well-tempered parallel bias metadynamics (PBMetaD) Pfaendtner and Bonomi [2015] implemented within PLUMED 2.8.1 Bonomi et al. [2009]. We use these calculations to recover an unbiased estimate of the potential of mean force, $PMF(r,z)$, describing the relative free energy of various configurations of the probe-target system parameterized by the radial separation r and axial separation z of the molecular centers of mass within a coordinate frame centered on the cyclodextrin probe. These PMFs enable

us to identify both the thermodynamically preferred binding locations for the the PFOS analyte and TCAA interferent into the three cyclodextrin probes and also make numerical predictions of the standard state equilibrium binding constants K_b for comparison with a field-effect transistor sensors. In Subsection 7.3.1, we discuss all-atom molecular dynamics simulations of target PFOS and interferents TCAA and SDS with α -, β - and γ -cyclodextrin probes. In Subsection 7.3.2, we describe enhanced sampling using funneled well-tempered parallel bias metadynamics. In Subsection 7.3.3, we estimate equilibrium binding free energies and constants.

7.3.1 All-atom molecular dynamics simulations of PFOS, SDS and TCAA interactions with cyclodextrin probes

We study the interactions of the PFOS analyte and SDS and TCAA interferent molecules with α -, β -, and γ -cyclodextrin probes using all-atom molecular dynamics simulations conducted in GROMACS 2021.6 Abraham et al. [2015]. The initial structures of PFOS and TCAA target molecules were generated using Avogadro version 1.2.0 Hanwell et al. [2012a] in PDB format. The PFOS, SDS and TCAA molecules were generated in anionic states U.S. Environmental Protection Agency [2022a], Meegoda et al. [2020], Wang et al. [2021b] corresponding to the pH 8 conditions explored in this work. The restrained electrostatic potential (RESP) method Bayly et al. [1993] was used to compute partial charges with Gaussian 16RevA.03 Frisch et al. [2016] software and General Amber Force Field (GAFF2) He et al. [2020] force field parameters. System topologies were generated in GROMACS Abraham et al. [2015] format using the acpype Sousa da Silva and Vranken [2012] library. The GAFF2 compatible q4md-CD force field Cézard et al. [2011], Dupradeau et al. [2010] combining GLYCAM04 Kirschner and Woods [2001b], Basma et al. [2001], Kirschner and Woods [2001a] and Amber99SB Cornell et al. [1995] was used to model the glucose units in the α -, β -, and γ -cyclodextrin probes. The structure and q4md-CD force field files in GROMACS Abraham et al. [2015] format of the α -, β -, and γ -cyclodextrin probes were obtained from Zhang et al. Zhang et al. [2012], van der Spoel [2017]. Calculations were performed in

cubic boxes of at least $6 \times 6 \times 6 \text{ nm}^3$ with periodic boundary conditions implemented in all three dimensions. The linear dimensions of the box were sufficiently large for the target-probe pair to reach the effectively non-interacting regime in the far field. The PFOS, SDS or TCAA molecule being simulated was initialized in the center of the cyclodextrin cavity for each calculation. Each system was solvated with TIP3P water molecules Jorgensen et al. [1983a] to a density of 1.0 g/cm^3 of water and PFOS systems were charge neutralized by the addition of a single sodium counterion to balance the formal (-1) net charge carried by PFOS. The total number of atoms in the simulations lie in the range of $\sim 12,000$ to $\sim 35,000$. Forces arising due to the Lennard-Jones potential were switched smoothly to zero at 1 nm. Electrostatic interactions were treated using the particle mesh Ewald Darden et al. [1993] method with a 1 nm real-space cutoffs and a reciprocal grid spacing optimized during runtime. The neighbor list was generated every 10 steps using the Verlet method Verlet [1967]. The LINCS algorithm Hess et al. [1997] was used to update the hydrogen atomic positions relative to their bonded non-hydrogen atoms in the system. High-energy overlaps in the initial system configuration were eliminated by steepest descent energy minimization using the low memory Broyden-Fletcher-Goldfarb-Shanno approach Saputro and Widyaningsih [2017] employing a maximum force tolerance of $10 \text{ kJ mol}^{-1} \text{ nm}^{-1}$ on any single atom. Initial velocities corresponding to $T = 300 \text{ K}$ were drawn from the Maxwell-Boltzmann distribution. A harmonic position restraint with force constant of $k = 500 \text{ kJ/mol.nm}^2$ was applied during these equilibration runs to the positions of the target or interferent-probe complex to prevent drift from their initial positions within the cyclodextrin cavity. The energy minimized systems were equilibrated for 1 ns in a NVT ensemble at temperature $T = 300 \text{ K}$ with a stochastic velocity rescaling thermostat Bussi et al. [2007]. Next, a 1 ns NPT ensemble equilibration with a target $T = 300 \text{ K}$ and a target pressure $P = 1 \text{ bar}$ was performed employing a stochastic velocity rescaling thermostat Bussi et al. [2007] and a Berendsen barostat Berendsen et al. [1984]. Equations of motion were integrated with a 2 fs time step using the leapfrog algorithm Hockney et al. [1974].

7.3.2 *Enhanced sampling using funnel well-tempered parallel bias metadynamics*

Enhanced sampling techniques are required to accelerate sampling of the interactions between the α -, β -, and γ -cyclodextrin probes and the PFOS analyte or TCAA or SDS interferent and assure good exploration of the thermally-relevant configurational phase space. We employed parallel bias metadynamics (PBMetaD), a technique in which the molecular dynamics simulation is subjected to biasing potentials along pre-selected collective variables (CVs) to accelerate transitions over free energy barriers in these coordinates and drive good sampling of configurational space Pfaendtner and Bonomi [2015]. Unbiased free energy landscapes can be extracted from the converged PBMetaD trajectories using thermodynamic reweighting of the configurations according to the applied converged biasing potential to recover the Boltzmann distribution in the thermodynamic ensemble under which the simulation was conducted Zhao et al. [2020]. We define our PBMetaD CVs within a coordinate system defined by the cyclodextrin probe. Since we are interested in the configurational phase space sampled by the target molecule relative to the probe (or vice versa), it is convenient to use a coordinate system centered on one of the molecules rather than the absolute coordinate system of the simulation box to mod out rigid translational and rotational displacements that alter the absolute position of the two molecules but not relative to one another. We define this coordinate system as illustrated in Fig. 7.1. The origin point is defined as the center of mass (COM) of all carbon and oxygen atoms in the narrow-side plane (primary face) of the D-glucose units of cyclodextrin probes. The XY plane is defined as this plane constructed from the carbon and oxygen atoms in the narrow-side plane of each 6 sided D-glucose units that constitute the cyclodextrin probes. The Z axis is defined via the normal vector to this plane passing through the origin of the coordinate system defined at the center of mass (COM) of all carbon and oxygen atoms in the cyclodextrin rings and with the positive sense of the z -axis pointing from the broader to the narrower end of the cyclodextrin molecule. Due to the high level of angular symmetry, our PBMetaD CVs do not require definition of the X and Y axes, so it is sufficient to define just the XY plane and the Z axis.

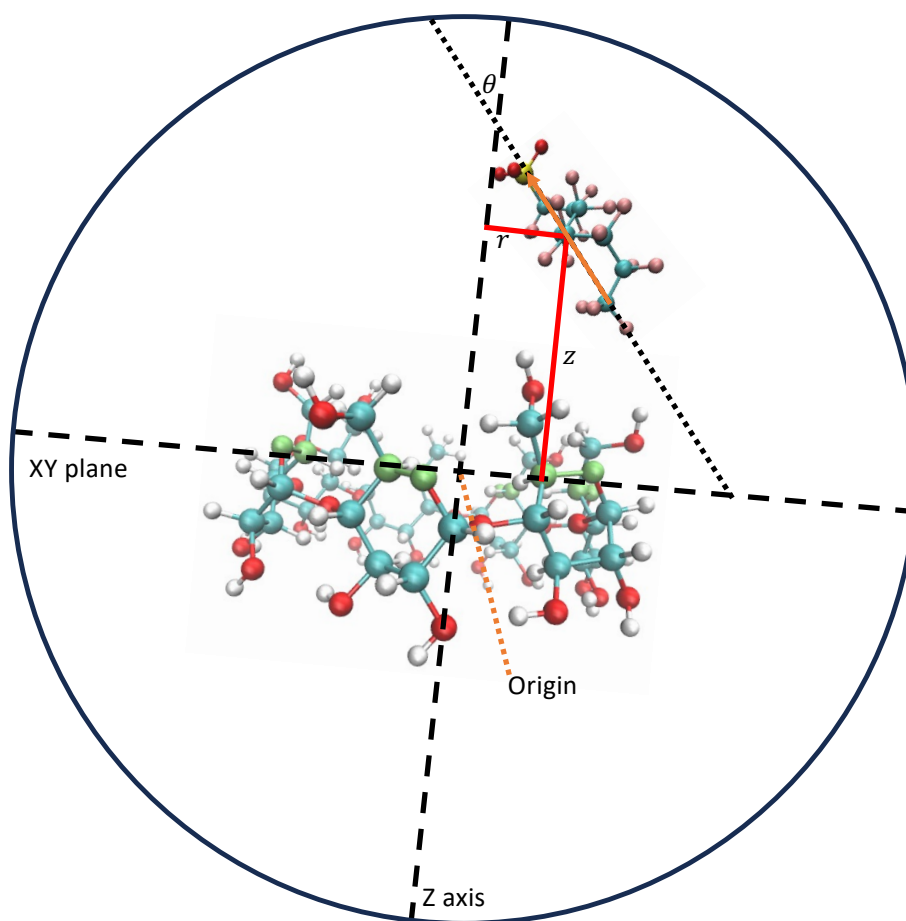


Figure 7.1: Illustration of the cyclodextrin-centric coordinate system and PBMetaD CVs within this reference frame for the α -cyclodextrin and PFOS system. The XY plane is defined by the carbon and oxygen atoms that are colored in green in the narrow-side plane of each 6 sided D-glucose units in each cyclodextrin probe and the Z axis is defined as the normal vector to that plane. The three PBMetaD CVs correspond to (i) z : the Z coordinate of the COM of the target molecule, (ii) r : the radial distance of the COM of the target molecule away from the Z axis in the XY plane, and (iii) θ : the planar angle between a vector drawn between the tail and head of the target molecule and the Z axis. Atoms are color coded; Carbon, hydrogen, oxygen atoms are colored in cyan, white, and red, respectively. Something along these lines will be nice. Molecular renderings were produced with VMD Humphrey et al. [1996]

We now define the three CVs used to drive our PBMetaD calculations (i) z : the Z coordinate of the COM of the target molecule, (ii) r : the radial distance of the COM of the target molecule away from the Z axis in the XY plane, and (iii) θ : the planar angle between a vector drawn between the head and tail of the target molecule and the Z axis. Physically, z controls the vertical separation between

the cyclodextrin probe and the target molecule, while r controls the orthogonal distance of the target molecule from the axis of the cyclodextrin ring. In selecting as our CV the radial distance r – as opposed to its decomposition into the x and y components, where $r^2 = x^2 + y^2$ – we have implicitly modeled out the approximate cylindrical symmetry of the cyclodextrin ring in the XY plane, and chosen r to drive sampling in a single lumped CV that just controls the radial component of the intermolecular COM separation. Finally, θ characterizes whether the target molecule approaches the cyclodextrin ring head-first or tail-first, and driving sampling in this CV helps sample both of these possibilities as the target molecule approaches the narrow ($z > 0$) and broad ($z < 0$) sides of the cyclodextrin ring. Since the target molecule can exist as a guest within the core of the cyclodextrin host, this CV is particularly useful in driving sampling of the different docked orientations.

Having defined the PBMetaD CVs, we then conduct the enhanced sampling calculations employing a well-tempered metadynamics procedure Barducci et al. [2008] in which we use a bias factor of $\gamma = 20$, an initial Gaussian height of $W = 1.2$ kJ/mol, and a Gaussian width of $\sigma = 0.2$ nm for z and r and 0.2 radians for θ . We deposit Gaussians with a pace corresponding to once every 500 time steps. Furthermore, to focus PBMetaD sampling on the region of configurational space of most interest where the target molecule is close to the cyclodextrin probe, we utilized funnel metadynamics Limongelli et al. [2013] following a procedure largely similar to that detailed in Rizzi et al. [2021]. The principle of the approach is that within the funnel region defined within the cyclodextrin-centric coordinate system, the target molecule does not feel any additional restraining forces and the metadynamics simulation proceeds as normal, whereas outside the funnel region, the target molecule experiences restraining forces that drive it back into the funnel. This focuses sampling on the region of space within the funnel to improve computational efficiency. In this work, we elect to use a spherical funnel of radius 2 nm centered on the origin of the cyclodextrin-centric coordinate system (Fig.7.1) and applied a harmonic restraining potential at the funnel walls with a force constant of $k = 5,000,000$ kJ/mol.nm². As demonstrated in Fig.7.2, the volume of the funnel was sufficiently small to achieve good sampling of the full configurational phase space within the

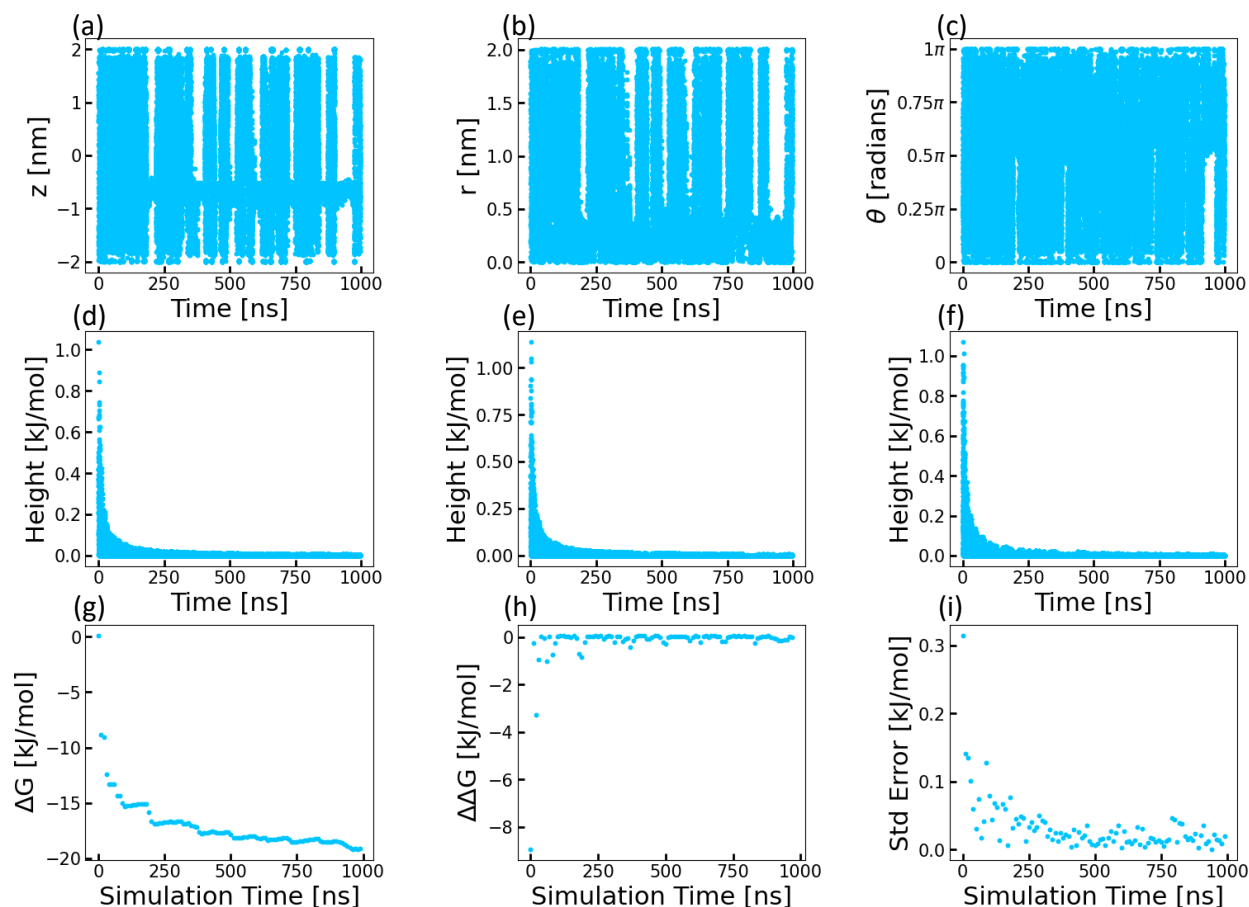


Figure 7.2: Illustration of convergence of the funnel well-tempered parallel bias metadynamics calculations for the α -cyclodextrin – PFOS system. (a-c) CVs z , r and θ (from left to right) as a function of simulation time. (d-f) Bias heights deposited as a function of time for CVs z , r and θ (from left to right). (g-i) ΔG s and related quantities computed from the simulation by using trajectory of length equal to the reported simulation time on x-axis. Points are computed at 10 ns intervals of simulation length. (g) Observed ΔG computed over the entire trajectory up until the reported simulation time. (h) Change in ΔG computed for each marginal 10 ns of simulation. (i) Standard error computed from 3 fold block averaging of ΔG over the trajectory of length equal to the reported simulation time on x-axis.

funnel using PBMetaD, and sufficiently large that the target molecule was sufficiently far from the probe at the funnel wall that it entered the effectively non-interacting regime such that the free energy landscape plateaued and no longer changed as a function of COM distance between the molecules.

We conducted the funnel well-tempered PBMetaD simulations in the NPT ensemble at $T = 300$

K and $P = 1$ bar using a Nosé-Hoover thermostat Evans and Holian [1985] and Parrinello-Rahman barostat Parrinello and Rahman [1981] using GROMACS 2021.6 Abraham et al. [2015] patched with PLUMED-2.8.1 Bonomi et al. [2009]. The simulation parameters were identical to those detailed above for system initialization, with the exception that integration of the equations of motion were performed with a 2 fs time step for PFOS and a 1 fs time step for TCAA. Empirically, we found the TCAA system required a shorter timestep to maintain numerical stability of the numerical integration. Snapshots were saved for analysis at a period of 2 ps. Following best metadynamics practices Bussi and Laio [2020], simulation convergence was validated by (i) monitoring for diffusive sampling in the three CVs, (ii) measuring the fall-off in the heights of the deposited Gaussian biases in each CV deposited over time, and (iii) assessing convergence of the calculated ΔG s of interaction for each simulated system using a block average estimate of uncertainty. Convergence was defined to have been reached when (i) qualitatively the transitions in CV-time plot are diffusive (Fig. 7.2) (ii) deposited bias height falls off to less than 1/100th of initial height (Fig. 7.2 b), and (iii) ΔG , uncertainties in ΔG and change in ΔG as a function of simulation length remain static with increased simulation length. (Fig. 7.2 c). In practice, we ran calculations for each of the nine systems – α -, β -, and γ -cyclodextrin probes with PFOS, SDS and TCAA targets – for a total of 1 μ s (500,000,000 steps) and we observed convergence in all three criteria within approximately 500 ns (Fig.7.2) The production MD simulation with enhanced sampling simulation took $\sim$ 4-11 days (between 90-250 ns/day) per system depending on the system size on a shared supercomputing node utilizing either $4 \times$ NVIDIA Quadro RTX 5000 GPUs and $2 \times$ Intel Xeon E5-2620 v4 Broadwell CPUs or $1 \times$ V100 GPU and $20 \times$ CPU cores of Intel Skylake processor.

7.3.3 *Estimation of equilibrium binding free energies ΔG and binding constants*

$$K_b$$

We now extract from the funnel well-tempered PBMetaD simulations for each system estimates of the equilibrium binding free energies and binding constants. We compile biased histograms

$\tilde{p}(r, z)$ in the radial and vertical CVs used to drive our PBMetaD calculations over the converged portion of our simulation trajectories. Since the bias continues to (slowly) evolve over the converged portion, to obtain the unbiased histograms $p(r, z)$, up to an unknown normalizing constant, we are obliged to reweight each frame according to the instantaneous bias applied in the CVs and funnel restraints using the reweighting method detailed in Zhao et al. [2020]. The unknown normalizing constant reflects our ignorance of the absolute free energy scale or, equivalently, the partition function, although the probability ratio of various (r, z) configurations can be quantitatively compared in a relative sense. As an aside, although the 2D r - z projection turns out to be a useful projection of the free energy surface for this system, we note that we are at liberty to accumulate unbiased histograms in this manner in any CVs of our choosing provided we properly reweight as a function of the PBMetaD CVs (and funnel restraints) in which the accelerating biases were applied. In this particular case, therefore, we conducted enhanced sampling in $\{r, z, \theta\}$ but choose to project our free energy landscapes into $\{r, z\}$ and marginalize out θ such that we integrate over the various different probe orientations sampled by our PBMetaD calculations at each (r, z) configuration.

Having obtained our estimates of the unbiased but unnormalized probability distributions through our histogram approximation, we then convert this into a potential of mean force (PMF) using the standard statistical mechanical relation $PMF(r, z) = -k_B T \log p(r, z) + C$, where C is an arbitrary additive constant reflecting our ignorance of the absolute free energy scale. Since we will only appeal to relative free energies (i.e., free energy differences or probability ratios) we are at liberty to specify C for numerical convenience and different choices of C do not affect our results.

In Fig. 7.4, we present the $PMF(r, z)$ for the systems to resolve the preferred binding location of each two target molecules with each of the three cyclodextrin probes, and choose to set C such that the mean PMF of the non-interacting region has a free energy of zero for visualization convenience. For the purposes of the binding constant calculations described below, it is convenient to adopt a convention wherein the PMF takes on a value of zero in the far field where the target and probe

are distantly separated at large r and/or z corresponding to the effectively non-interacting state (i.e., $PMF(r \rightarrow \infty, z \rightarrow \infty)=0$). In practice, we do this by projecting the PMF into the COM separation $s = \sqrt{r^2 + z^2}$ between the target and the probe, subtracting out the entropic correction Abraham et al. [2015], Páll et al. [2014], Neumann [1980], and computing the mean value of $PMF(s)$ in the plateau region $1 \leq (s = \sqrt{r^2 + z^2}) \leq 3.5$. In performing this calculation, we exclude the region near the funnel wall to avoid edge effects associated with the large restraining potentials in these regions of configurational space. We present in Fig. 7.3 the resulting $PMF(s)$ free energy landscapes for each combination of PFOS and TCAA target-probe systems. These plots confirm that the box size was sufficiently large in each case that the non-interacting far field region was sampled in the calculations corresponding to the plateau region where the free energy no longer changes with increasing separation. As described below, this observation is important for the purposes of reliable estimation of binding constants.

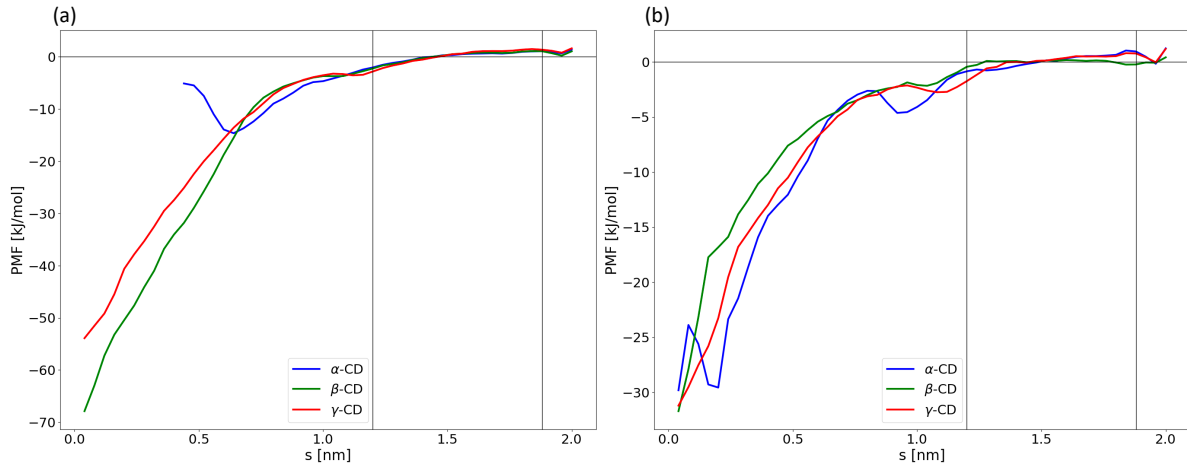


Figure 7.3: Univariate free energy landscapes $PMF(s)$ projected into the COM separation of the target and probe $s = \sqrt{r^2 + z^2}$ illustrating that we reach the effectively non-interacting regime in the far field where the free energy ceases to be a function of s . Landscapes are adjusted by the entropic correction accounting for the trivial increase in configurational phase space volume associated with increasing s and which physically corresponds to subtracting out the $PMF(s)$ corresponding to non-interacting point particles. (a) Describes $PMF(s)$ for PFOS. (b) Describes $PMF(s)$ for TCAA.

We estimate the free energy difference ΔG between the bound and unbound states using a statistical mechanical formalism based on the ratio of the partition functions in each of the two states Dasetty

et al. [2023], Aldeghe et al. [2018], General [2010]. Specifically, we define

$$\Delta G = -k_B T \ln \left(\frac{Z_b}{Z_u} \right), \quad (7.1)$$

where the bound and unbound state partition functions are,

$$Z_b = \int_{\Gamma_b} \exp(-\beta U(\mathbf{x})) d\mathbf{x} \quad (7.2)$$

$$Z_u = \int_{\Gamma_u} \exp(-\beta U(\mathbf{x})) d\mathbf{x}, \quad (7.3)$$

where $U(\mathbf{x})$ is the potential energy of the system as a function of the full dimensional configurational phase space coordinates $\mathbf{x}$, Γ_b and Γ_u respectively delimit the configurational phase space volume occupied of the bound and unbound states, and $\beta = 1/k_B T$.

We choose to specify the extent of the bound state as a cylinder of radius 1.2 nm and linear extent 2 nm centered around the COM of the cyclodextrin probe. The limits of this cylinder were motivated by the $PMF(s)$ flattening as we enter the non interacting region, shown in Fig. S3 and a cylindrical, as opposed to spherical, geometry was selected based on the shape of the cyclodextrin molecules. Alternative definitions for the bound state can result in slightly different values for ΔG , and we quantify these differences below. Importantly, this cylinder is contained fully within the spherical funnel region used in the funnel well-tempered PBMetaD calculations and so we need not be concerned about any issues of poor sampling outside of the funnel region. The bound state partition function then becomes,

$$Z_b = 2\pi \int_0^{r_{\text{bound}}} \int_{-z_{\text{bound}}}^{z_{\text{bound}}} r \exp(-\beta(PMF(r, z) - PMF(r \rightarrow \infty, z \rightarrow \infty))) dr dz \quad (7.4)$$

$$= 2\pi \int_0^{r_{\text{bound}}} \int_{-z_{\text{bound}}}^{z_{\text{bound}}} r \exp(-\beta PMF(r, z)) dr dz, \quad (7.5)$$

where $r_{\text{bound}} = z_{\text{bound}} = 1.2$ nm. In the first line, we note that the potential energy in the full configurational phase space coordinates $U(\mathbf{x})$ has been replaced by the potential of mean force $PMF(r, z)$ in the Boltzmann factor due to marginalization over all of the other degrees of freedom in constructing the cylindrical integral, and we have performed the angular integral since there is no dependence on θ in $PMF(r, z)$. In the second line, we note that under our specification of the arbitrary additive constant in the potential of mean force, we have specified that $PMF(r \rightarrow \infty, z \rightarrow \infty) = 0$. We note that asserting this limit is numerically convenient but the final results of our ΔG calculation do not depend on this choice due to cancellation in the numerator and denominator of the ratio in Eqn. 7.1.

The extent of the unbound state is naturally defined as the volume of the simulation box not occupied by the bound state. For a sufficiently large simulation box and rapid convergence of $PMF(r, z)$ to the $PMF(r \rightarrow \infty, z \rightarrow \infty)$ limit, it is a reasonable approximation to take $PMF(r, z) = PMF(r \rightarrow \infty, z \rightarrow \infty) = 0$ and perform the integral over the full volume of the simulation box. In the present case, assuming the restraining potentials in our metadynamics funnel to be sufficiently strong, we may approximate them as hard walls and conduct the integration over the volume of the spherical funnel,

$$Z_u \approx \int_0^{r_{\text{funnel}}} 4\pi r^2 \exp(-\beta PMF(\infty)) dr \quad (7.6)$$

$$= \int_0^{r_{\text{funnel}}} 4\pi r^2 dr \quad (7.7)$$

$$= \frac{4}{3} \pi r_{\text{funnel}}^3 \quad (7.8)$$

$$= V_{\text{funnel}}, \quad (7.9)$$

where $r_{\text{funnel}} = 2$ nm is the radius of the spherical funnel and $V_{\text{funnel}} = \frac{32}{3} \pi \text{ nm}^3$ is the volume of the funnel.

Combining Eqns. 7.1, 7.5, and 7.9, we arrive at our final expression for ΔG ,

$$\Delta G = -k_B T \ln \left[\frac{2\pi}{V_{\text{funnel}}} \int_0^{r_{\text{bound}}} \int_{-z_{\text{bound}}}^{z_{\text{bound}}} r \exp(-\beta PMF(r, z)) dr dz \right]. \quad (7.10)$$

We recognize that the concentration of each species in the simulation box is $C = C_{\text{probe}} = C_{\text{target}} = C_{\text{probe-target}} = 1/V_{\text{funnel}}$ and that in general C differs from the standard concentration of $C^\Theta = 1 \text{ M}$ corresponding to a standard volume of $V^\Theta = 1/C^\Theta = 1661 \text{ \AA}^3$ Hermans and Wang [1997]. We are obliged, therefore, to apply a concentration correction to estimate the binding free energy at the standard concentration Dasetty et al. [2023],

$$\Delta G^\Theta = \Delta G + k_B T \ln \left(\frac{C}{C^\Theta} \right) \quad (7.11)$$

$$= -k_B T \ln \left[C^\Theta 2\pi \int_0^{r_{\text{bound}}} \int_{-z_{\text{bound}}}^{z_{\text{bound}}} r \exp(-\beta PMF(r, z)) dr dz \right]. \quad (7.12)$$

Comparing this expression to the relation Dasetty et al. [2023],

$$\Delta G^\Theta = -k_B T \ln K_b, \quad (7.13)$$

we can immediately express the binding constant K_b as,

$$K_b = C^\Theta 2\pi \int_0^{r_{\text{bound}}} \int_{-z_{\text{bound}}}^{z_{\text{bound}}} r \exp(-\beta PMF(r, z)) dr dz. \quad (7.14)$$

Results

7.3.4 Evaluation of binding energies

We present in Fig. 7.4 the results of our enhanced sampling all-atom molecular dynamics calculations to map out the free energy landscape for binding of the PFOS analyte and TCAA interferent with the α -, β -, and γ -cyclodextrin probes. We represent this landscape through the potential of

mean force $PMF(r, z)$ describing the relative free energy of different (r, z) configurations of the probe-target system in the radial separation r and axial separation z of the molecular centers of mass within a coordinate frame centered on the cyclodextrin probe (Fig. 7.4a). In Figs. 7.4b-g we present the PMFs for the six combinations of the two target molecules (PFOS analyte and TCAA interferent) and three probes (α -, β -, and γ -cyclodextrin) along with molecular snapshots of the thermodynamically preferred bound configuration. We also report in Table 7.1 the values of the standard state binding constant K_b calculated for each of the six systems along with our experimentally measured values from QCM. In all cases, we observe good agreement between the computational and experimental K_b values, lending confidence to the computational results.

First considering the α -cyclodextrin system (Figs. 7.4b,e), the white region in the vicinity of $(r = 0, z = 0)$ indicates that the ~ 0.53 nm diameter α -cyclodextrin cavity is too narrow to accommodate either molecule within the probe. Driven primarily by the hydrophobic nature of the two target molecules, the most probable location (i.e., lowest free energy) for the target molecules relative to the probe is at $(r \approx 0, z \approx -0.5)$ for the PFOS analyte and $(r \approx 0, z \approx -0.2)$ for the TCAA interferent. This indicates that both target molecules preferentially bind to the broader (negative z) opening to the probe in favor of the narrower (positive z) opening. The smaller TCAA molecule is able to insert further into the cyclodextrin cavity than the relatively larger PFOS but this belies the two orders of magnitude weaker binding constant for TCAA relative to PFOS. The PMFs also reveal the existence of a secondary, less preferred, binding site for TCAA at the narrow end of the α -cyclodextrin probe, although this does not exist for the larger PFOS. For both TCAA and PFOS, we observe weak binding interactions on the exterior of the probe evinced by the weak free energy well around the exterior of the interior forbidden white region.

Turning to β -cyclodextrin (Figs. 7.4c,f), the wider ~ 0.65 nm diameter probe is sufficiently large to accommodate both TCAA and PFOS within the cavity and in both cases the thermodynamically preferred state is for the target guest to reside within the cavity of the probe host. The value of our

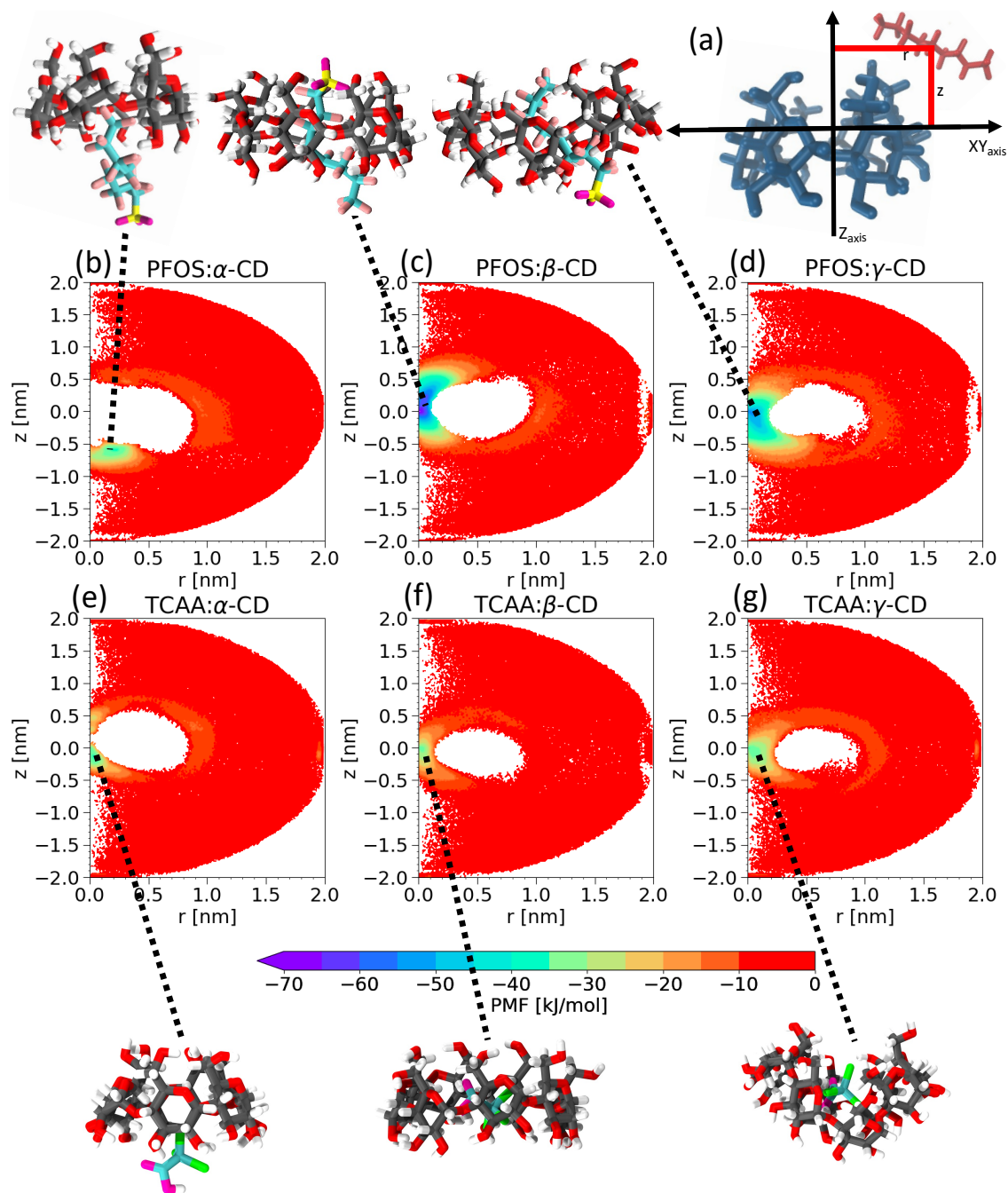


Figure 7.4: Molecular simulations of PFOS analyte and TCAA interferent binding with α -, β -, and γ -cyclodextrin probes. (a) Illustration of the cyclodextrin-centric coordinate frame. Potential of mean force (PMF) landscapes $PMF(r, z)$ are constructed in the radial separation r and axial separation z of the molecular centers of mass. PMFs for (a) α -cyclodextrin–PFOS (b) β -cyclodextrin–PFOS (c) γ -cyclodextrin–PFOS (d) α -cyclodextrin–TCAA (e) β -cyclodextrin–TCAA (f) γ -cyclodextrin–TCAA. For visualization convenience, the arbitrary additive constant for each PMF reflecting our

ignorance of the absolute free energy scale is specified such that the average free energy of the non-interacting region ($r > 1.2, z > 1.2$) is set to zero. The location ($r=0$ nm, $z = 0$ nm) corresponds to a configuration in which the target molecule resides within the pore of the probe and the centers of mass are coincident. The white region in the center of each plot corresponds to the prohibited region due to excluded volume overlaps of the target molecule and the cylindrical wall of the probe. The white region around the edge of each plot corresponds to the far-field region that was unsampled under our funnel metadynamics approach that focused computational sampling on the near-field region. In all cases the narrow end of the cyclodextrin probe is oriented towards positive z values and the broader end towards negative z values. Molecular snapshots corresponding to the thermodynamically preferred bound configuration are presented for each system. All renderings were constructed using VMD Humphrey et al. [1996].

computed PFOS binding constant of $K_{b,simulation}^{PFOS} = 2.43 \pm 0.02 \times 10^6$ is in good agreement with field-effect transistor sensors readings suggesting strongest adsorption of PFAS on β -CD surfaces and weak adsorption for TCAA. It is also consistent with previous experimental reports that lie in the range $K_b = 10^5$ - 10^6 Weiss-Errico and O'Shea [2017a], Chen et al. [2023], Weiss-Errico and O'Shea [2019], and is approximately six orders of magnitude larger than that for TCAA. As illustrated in the representative molecular snapshots of the thermodynamically preferred bound configurations, the relative preference for PFOS can likely be understood as the bulkier PFOS providing stronger van der Waals interactions with the interior probe wall than the relatively smaller TCAA. This trend also appears to underpin the relatively larger K_b for TCAA with the smaller α -cyclodextrin probe relative to the larger β -cyclodextrin cavity due to better contact between the TCAA and the interior probe wall in the smaller α -cyclodextrin system. Further analysis of the PFOS- β -cyclodextrin bound complex indicates that the fluorinated backbone tends to occupy the cavity leaving the charged head exposed into solvent. This suggests that there may be an opportunity to further enhance the PFOS binding constant by engineering the lip of the β -cyclodextrin probe to promote favorable electrostatic interactions with the PFOS head.

The γ -cyclodextrin probe possesses the widest cavity diameter of ~ 0.83 nm, which can easily accommodate both TCAA and PFOS within the cavity as the thermodynamically preferred state of the complex (Figs. 7.4d,g). The wider cavity gives the bound target molecules more configurational

Probe	$K_b^{PFOS} (M^{-1})$	$K_b^{TCAA} (M^{-1})$	K_b^{PFOS}/K_b^{TCAA}
α -cyclodextrin	$2.13 \pm 0.01 \times 10^3$	$3.06 \pm 0.02 \times 10^1$	$6.99 \pm 0.05 \times 10^1$
β -cyclodextrin	$2.43 \pm 0.02 \times 10^6$	$4.78 \pm 0.02 \times 10^0$	$5.08 \pm 0.05 \times 10^5$
γ -cyclodextrin	$1.09 \pm 0.02 \times 10^5$	$3.13 \pm 0.03 \times 10^1$	$3.46 \pm 0.03 \times 10^3$

Table 7.1: Binding constants K_b for the PFOS analyte and TCAA interferent with the α -, β -, and γ -cyclodextrin probes calculated using enhanced sampling all-atom classical MD simulations. The selectivity of each cyclodextrin probe for PFOS relative to TCAA is quantified by the ratio of the binding constants K_b^{PFOS}/K_b^{TCAA} .

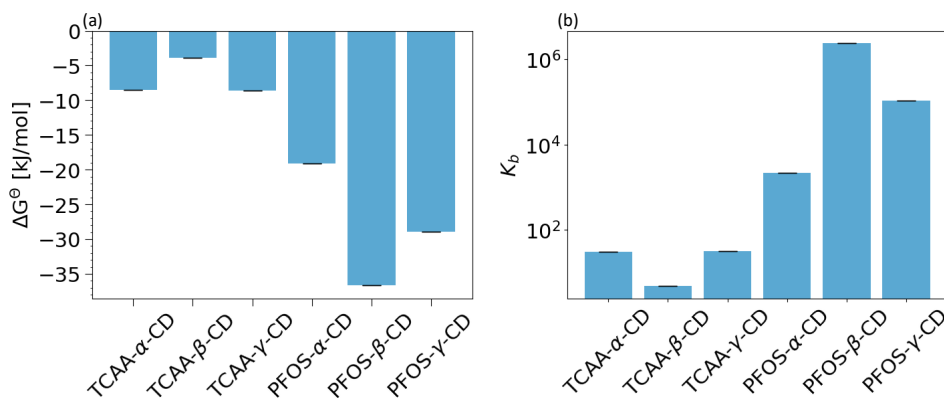


Figure 7.5: Calculated values of the standard state (a) binding free energy and (b) binding constant for α -, β -, and γ -cyclodextrin probes with the PFOS and TCAA target molecules.

freedom within the cavity leading to a greater diversity of bound states. As illustrated by the representative bound configuration, the PFOS chain no longer contacts both walls of the probe while the charged hydrophilic head of the chain is often still extended into the solvent. The less snug fit of the PFOS chain leads to weaker dispersion interactions and an order of magnitude smaller binding constant for PFOS to γ -cyclodextrin relative to β -cyclodextrin. Conversely, the TCAA interferent has a weaker stronger binding constant to γ -cyclodextrin relative to β -cyclodextrin. The widening of the cyclodextrin cavity appears to increase the number of conformational states accessible to the TCAA within that region by accommodating the relatively bulky carbon-oxygen bonds. We present a summary of the binding constants extracted from all simulations of TCAA and PFOS with each cyclodextrin probe as well as relative binding affinities in Table 7.1.

Finally, to demonstrate the robustness of parameter choices for the bound region, we report the calculated values of $\Delta G^\ominus$ and K_b from our funnel well-tempered PBMetaD calculations for the binding of the PFOS analyte or TCAA interferent with the α -, β -, and γ -cyclodextrin probes in Tables 7.2 and 7.3. Numerical quadrature of the 2D integral was performed 2D Simpson’s rule over $PMF(r, z)$. Uncertainties were estimated by three-fold block averaging over the converged portion of the simulation trajectories. The numerical values of $\Delta G^\ominus$ and K_b can depend on the definition of the bound state. We report the calculated values of $\Delta G^\ominus$ and K_b under different choices for the bound state. These calculations reveal that there is some minor sensitivity to the choice of bound state phase space volume within noise, and the trends in these quantities are preserved under all choices for the bound state.

$K_b [M^{-1}]$	$s_{bound} \leq 0.8 \text{ nm}$	$s_{bound} \leq 1.0 \text{ nm}$	$s_{bound} \leq 1.2 \text{ nm}$	$s_{bound} \leq 1.4 \text{ nm}$	$s_{bound} \leq 1.6 \text{ nm}$
TCAA-α-CD	$9.91 \pm 0.16 \times 10^0$	$2.01 \pm 0.01 \times 10^0$	$2.73 \pm 0.01 \times 10^0$	$3.82 \pm 0.03 \times 10^0$	$4.74 \pm 0.07 \times 10^0$
PFOS-α-CD	$1.29 \pm 0.05 \times 10^3$	$1.50 \pm 0.06 \times 10^3$	$1.71 \pm 0.02 \times 10^3$	$1.96 \pm 0.01 \times 10^3$	$2.11 \pm 0.07 \times 10^3$
TCAA-β-CD	$1.86 \pm 0.04 \times 10^0$	$3.74 \pm 0.01 \times 10^0$	$7.26 \pm 0.02 \times 10^0$	$1.21 \pm 0.06 \times 10^0$	$1.83 \pm 0.06 \times 10^0$
PFOS-β-CD	$1.82 \pm 0.04 \times 10^6$	$1.92 \pm 0.06 \times 10^6$	$2.16 \pm 0.04 \times 10^6$	$2.41 \pm 0.03 \times 10^6$	$2.37 \pm 0.03 \times 10^6$
TCAA-γ-CD	$1.07 \pm 0.06 \times 10^1$	$1.45 \pm 0.02 \times 10^1$	$2.49 \pm 0.02 \times 10^1$	$3.49 \pm 0.04 \times 10^1$	$5.23 \pm 0.03 \times 10^1$
PFOS-γ-CD	$3.97 \pm 0.03 \times 10^4$	$4.21 \pm 0.02 \times 10^4$	$4.63 \pm 0.04 \times 10^4$	$5.11 \pm 0.05 \times 10^4$	$5.31 \pm 0.02 \times 10^4$

Table 7.2: Binding constant K_b computed under cutoff regimes from $0.8 \text{ nm} \leq s_{bound} \leq 1.6 \text{ nm}$.

$\Delta G^\ominus [\text{kJ/mol}]$	$s_{bound} \leq 0.8 \text{ nm}$	$s_{bound} \leq 1.0 \text{ nm}$	$s_{bound} \leq 1.2 \text{ nm}$	$s_{bound} \leq 1.4 \text{ nm}$	$s_{bound} \leq 1.6 \text{ nm}$
TCAA-α-CD	-5.72 ± 0.04	-7.48 ± 0.02	-8.25 ± 0.01	-9.08 ± 0.02	-9.62 ± 0.00
PFOS-α-CD	-17.87 ± 0.09	-18.22 ± 0.10	-18.57 ± 0.04	-18.89 ± 0.01	-19.10 ± 0.08
TCAA-β-CD	-1.55 ± 0.06	-3.29 ± 0.01	-4.94 ± 0.01	-6.22 ± 0.03	-7.26 ± 0.01
PFOS-β-CD	-35.95 ± 0.05	-36.09 ± 0.01	-36.38 ± 0.04	-36.65 ± 0.03	-36.61 ± 0.02
TCAA-γ-CD	-5.91 ± 0.01	-6.68 ± 0.04	-8.01 ± 0.02	-8.86 ± 0.03	-9.87 ± 0.02
PFOS-γ-CD	-26.40 ± 0.21	-26.55 ± 0.14	-26.80 ± 0.02	-27.05 ± 0.03	-27.13 ± 0.09

Table 7.3: Binding free energies $\Delta G^\ominus$ computed under cutoff regimes from $0.8 \text{ nm} \leq s_{bound} \leq 1.6 \text{ nm}$.

In Fig 7.6, we present a comparison of the sensitivities and selectivities of β -CD with PFOS, SDS and TCAA. We choose this probe as it is the most sensitive to PFOS. Fig 7.6(a-c) indicates that β -CD experiences its strongest binding to all analytes when they occupy the center cavity. However, PFOS enjoys both the deepest free energy of binding and largest overall binding constant with β -CD. SDS has a wider free energy basin, admitting a larger variety of bound states than PFOS, however

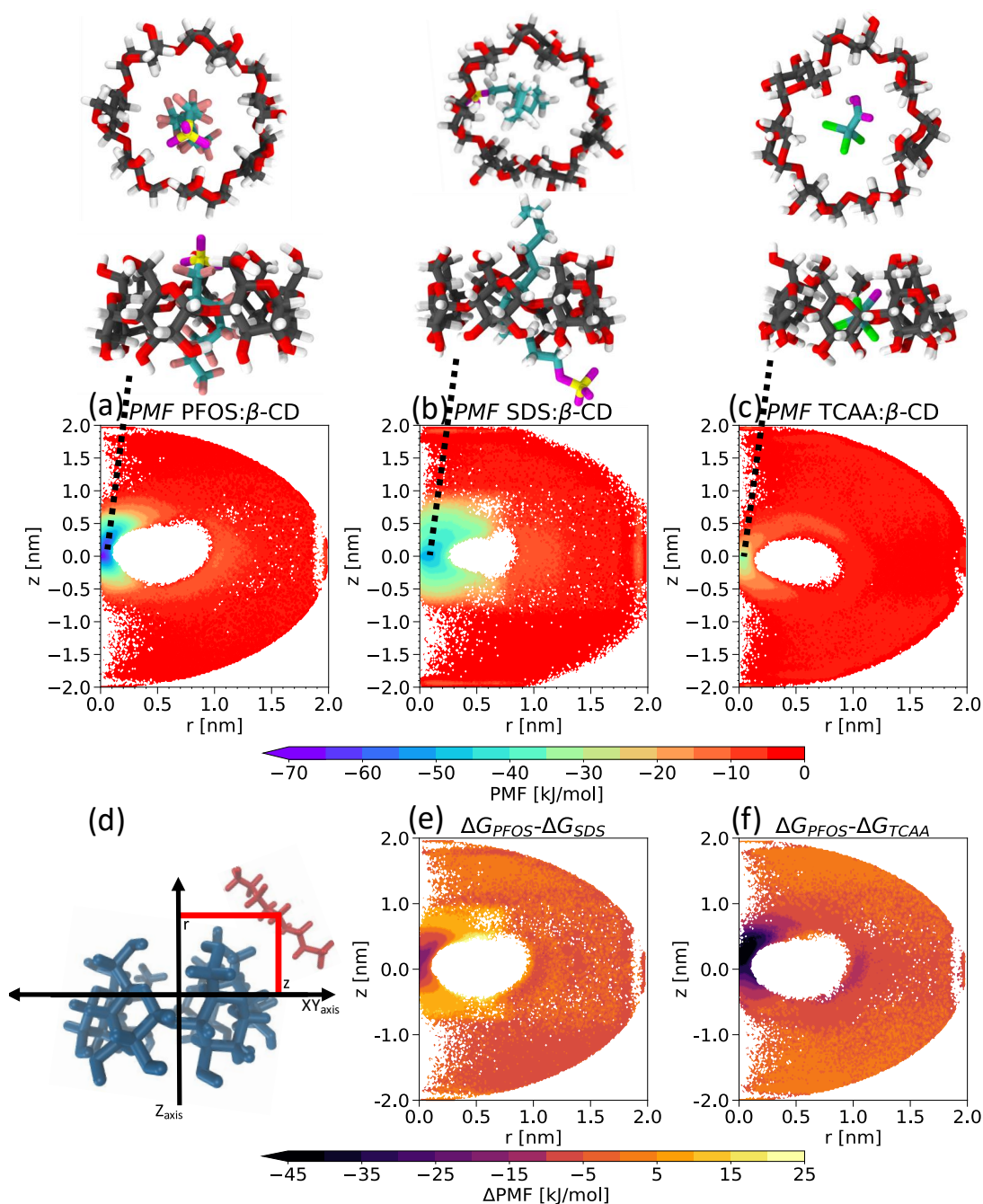


Figure 7.6: PMF landscapes and selectivity of β -CD with PFOS, SDS and TCAA. (a) PMF landscape of β -CD with PFOS. (b) PMF landscape of β -CD with SDS. (c) PMF landscape of β -CD with TCAA. Renderings of each system in its strongest bound state are presented above (a-c). (d) diagram describing construction of r and z collective variables as a function of probe and analyte positioning. (e) Spatially resolved selectivity of β -CD with PFOS relative to SDS. (f) Spatially resolved selectivity of β -CD with PFOS relative to TCAA.

with a lower mean free energy. TCAA experiences only very weak binding. Fig 7.6(e) demonstrates that while PFOS has a deeper maximum δG of interaction with β -CD than SDS has, there are less accessible states on the $[r, z]$ plane that experience strong bonding with β -CD. Although both PFOS and SDS are straight chain hydrocarbons, the electrostatic interactions between the fluorinated backbone of the PFOS probe are more restrictive than the hydrogen backbone of SDS. This accounts for the stronger free energies of interaction relative to SDS but larger exclusion zone on the $[r, z]$ plane relative to SDS. Fig 7.6(f) demonstrates that β -CD preferentially binds to PFOS relative to TCAA. These spatially resolved sensitivities and selectivities indicate the β -CD will preferentially bind to PFOS relative to both SDS and TCAA, however, interaction energies are similar for PFOS and SDS.

7.4 Conclusions

The all-atom molecular simulations provide molecular-level insight into the structural and thermodynamic behaviors of the target-probe complex binding. The computed binding constants are in good agreement with both our and prior experimental measurements, lending confidence to our computational results. Our calculations suggest that the superior sensitivity and specificity of β -cyclodextrin for PFOS over TCAA can be understood as a good size match of the β -cyclodextrin cavity for the PFOS molecule that results in favorable dispersion interactions of the hydrophobic chain within the cavity while enabling the hydrophilic head group to be solvent exposed.

Our results also suggest potential routes for rational engineering of the rim of the β -cyclodextrin to promote favorable electrostatic interactions with the PFOS head and further elevate the binding constant. The favorable selectivity of β -CD for PFOS against all tested interferents, particularly in the largest interaction energy states indicates that β -CD is a good vehicle to build upon. In future work, we intend to extend our high throughput screening approach to molecular design by engineering the rims of these cyclodextrin probes to increase sensitivity to and selectivity for PFOS

in the face of interferent molecules.

7.5 Acknowledgments

This material is based upon the work supported by the National Science Foundation under Grant No. DGE-2022023. This work was supported in part with funding by the University of Chicago Data Science Institute (DSI). 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. We gratefully acknowledge computing time on Frontera hosted by the Texas Advanced Computing Center (TACC) at The University of Texas at Austin. Work by S.B.D. and Y.W. at Argonne National Laboratory was supported by the AMEWS EFRC funded by DOE, Office of Science, BES under contract DE-AC02-06CH11357.

CHAPTER 8

CONCLUSIONS

8.1 Summary of Contributions

This work provides *a* solution to the problem of reconstructing full-dimensional dynamics from time series for arbitrary dynamical systems. Building on the confluence of the groundbreaking work of Takens, Ruelle, Eckmann and others in the embedding theorems and advances in universal functional approximation through statistical learning, this dissertation addresses several key challenges. First, we empirically demonstrate that full-dimensional dynamics of complex systems can be reconstructed from univariate time series, even under deterministic and stochastic forcing characteristic for real world experiments. Second, we developed a software package that implements TAR on any time series and introduced new constructions of Takens' Theorem to maximize the information embedded in latent space representations of dynamics. Third, we extended this framework to problems lacking an explicit differential equation formulation and demonstrated its value in a social science context. While the universal functional approximators employed may not always represent the optimal computational model for any given mapping problem, the robust and interoperable TAR mechanism allows for each step to be retrofitted with any model. The mathematics that governing latent space embeddings from data, observable selection, and delay vector structuring are likely to outlast any specific choice of dimensionality reduction and nonlinear mapping algorithms implemented in this dissertation. We hope that these models and the TAR pipeline become one in a class of reconstruction techniques that treat the geometry of time series as a fundamental facet of any dynamics problem to be used in concert with machine learning. This will allow for fast, efficient and mathematically rigorous representations of the dynamics of any complex system in latent space.

8.2 A review of ideas that did not work and future directions

The market can remain wrong for longer than I can remain solvent - John Maynard Keynes

It is often difficult to distinguish between ideas that did not work and potential future directions. Below, I provide a non-exhaustive list of ideas I tested and explored over the course of the past six years Which, in my opinion, constitute interesting but incomplete studies. Time and PhDs are finite, so while I have faith in the core concepts here presented, none are fully complete. It is my hope that, with time and as the intersection of machine learning and embedding theorems evolves, some of these ideas continue to develop and take root.

8.2.1 *Generative Techniques in Latent Space and Latent Space Liouville*

Propogators

After the publication of the first TAR paper Topel and Ferguson [2020], we explored the use of generative adversarial networks (GANs) as an alternative mapping to ANNs, with the added benefit of generating novel trajectories in latent space Sidky et al. [2020a]. While the unpredictable nature of GANs and their relatively expensive training, compared with ANNs, made them impractical for backmapping in STAR, such generative techniques could still be useful in generating synthetic latent space dynamics. I experimented with various data structuring methods to enforce constraints to the solutions suggested by the GAN, including a second implementation of Takens' Theorem in latent space, attempting to guide the GAN to learn the properties of physically correct trajectories. While these efforts were largely unsuccessful, this spurred an idea that could ensure holonomic constraints were respected even in latent space. There are likely more sophisticated statistical learning approaches that could produce improved latent space trajectories, however, molecular systems are also subject to physical constraints providing another avenue to improve modeling.

Several dimensionality reduction techniques have been introduced that can construct latent space

manifolds with a consistent metric Perrault-Joncas and Meilă [2011], Chadebec et al. [2023], Bertalan et al. [2020]. This property is highly desirable for physical systems, as it allows us to enforce geometric holonomic constraints that exist in high-dimensional space in the reduced-dimensional latent space. The idea is as follows: i) Liouville’s theorem guarantees that any legal trajectory must preserve area in phase space, ii) techniques like MML and RHVAE can produce latent space manifolds that preserve Riemannian metric iii) any generative technique producing a novel trajectory in latent space can ensure the physical viability of that trajectory by preserving area of infinitesimal regions propagated in the latent space.

The idea of symplectic integrators is not new; Ronald Ruth Forest and Ruth [1990] proposed algorithms capable of performing symplectic integration in high-dimensional spaces. The major innovation here is the ability to apply this in latent space, allowing molecular dynamics simulations to evolve in tens of thousands of dimensions using just a few. While theoretically simple, the main challenge lies in computing areas in phase space. As it turns out, for moderately high-dimensional systems (5+ dimensions), this remains a largely unsolved problem that requires relatively expensive bounding calculations. However, many protein systems are low-dimensional, and volumes can still be computed. Although I had to abandon this work in favor of more viable applications of the TAR framework, the idea remains sound and merits further investigation.

8.2.2 *Botany Meets Dynamics*

Both Hawking and Oppenheimer mused that particle physics had become akin to botany or zoology, focused on the classification of the unending parade of new particles. Though these may have been jabs at particle physics, the practical merits of classification are quite clear and often come with scientific utility.

The Protein Data Bank (PDB) contains hundreds of thousands of protein structures. However, as we discussed in Chapters 2 and 3, structure is only a small part of the problem. TAR offers a

window into the dynamics of arbitrary systems from time series data, producing manifolds that are diffeomorphically related to the dynamics of the full-dimensional system. Consequently, both molecular dynamics simulations and experimental time series data can be used to extract and classify the dynamics of proteins. Significant efforts in the classification of dynamical systems, particularly by physicists like William Gilpin, have focused on toy or low-dimensional systems. By applying TAR to large numbers of proteins or molecules and classifying their dynamics, we could generate new way of understanding and grouping these systems. For instance, proteins in the same family may not have the same function, and proteins from different families could share functions. Could we group proteins that share function dynamically in phase space? TAR enables us to do just that from either experimental observations like smFRET or MD simulation, offering insights into protein engineering and design.

This idea emerged from a much more fundamental problem. After conducting some 500 simulations to produce the data in Chapter 6 and performing a LASSO regression on the design properties of our candidates, we found that the only monotonic relation was that larger molecular probes had larger free energies of association with both PFAS and SDS. This result is wholly unsurprising from an entropic perspective, but it reveals that even the design space of simple, straight chain hydrocarbons resists being characterized without from their dynamics. This raised the question: do we even know how the free energy landscape of methane changes when we add a second or third carbon? MD simulations can provide some insight, but we lack hard and fast rules to predict how dynamics will change in these simple cases, let alone in more complex ones. TAR offers a unique, universal method to extract the dynamics from experiments or simulations, allowing for the classification of these phase spaces. In doing so, it presents the opportunity to learn transformations phase space when modifying systems, such as adding single carbons to a molecule.

Moreover, while we can construct phase transition diagrams for systems undergoing changes such as an increase in pressure or temperature, getting at the changes to the accessible phase spaces

of these systems is a far more difficult problem. TAR allows us to understand these transitions, from either simulations or experiments. For instance, smFRET measurements run taken at varying temperatures could inform us about how accessible phase spaces change. In combination with something like a protein dynamics data bank, it could allow us to infer changes to dynamics in proteins that are out of sample, but that are dynamically similar to ones proteins for which we have already studied these temperature-induced changes.

All of these proposals require a substantial simulation or experimental costs. However, the time and compute cost of running TAR is trivial relative to both of those costs, especially with the TAR package now available in Python. If a data bank of these dynamics could be constructed, it would significantly democratize access to protein dynamics data, similar to what PDB has done for structural data.

8.2.3 *Econometrics*

In this dissertation, I presented market microstructure time series reconstructed with TAR. These system all evolve on short time scales, with trades happening on the scale of milliseconds for very liquid securities. What about things with lower data resolution such as monthly jobs numbers or company quarterly reports? Our work on currencies (particularly on the Canadian Dollar and Chinese Yuan), commodities and interest rate rate dynamics on long time scales has revealed the strength of theoretically predicted links between dynamical drivers of the returns and the returns themselves via TAR. While it is difficult to compare the results to linearized econometric models which, while theoretically interperetable, do not describe dynamics well. Our investigations of these systems reveal secular trends evolving on the multi-month to multi-year time scale. It then stands to reason that these lower time resolution data are sufficient to extract useful dynamics in and of themselves. This opens the door to studying some of the most interesting problems in macroeconomics.

For instance, the Phillips curve, which describes the theoretical relation between unemployment and inflation, is typically estimated for individual countries and time periods in macroeconomic literature. While some variation across countries and time is expected, why does the Phillips curve fail to exhibit a consistent form, at the very least. I cannot count the number of “Is the Phillips Curve Dead?” articles I have read. This debate likely stems from the fact that the dynamics of both inflation and unemployment are unlikely to be one-dimensional, confounding correlation-based analyses. Instead of considering the scalar relations between the two variables, we can construct manifolds of unemployment and inflation. While there may still be a monotonic relationship between the manifolds, true insight lies in understanding the relationship between the underlying dynamics behind these variables, rather than measuring isolated monthly or country-specific values. TAR provides a way to learn these manifolds directly from unemployment and inflation data. This allows us to examine the universality of the Phillips curve. Similar to our study of the key drivers of Canadian Dollar prices, we can study the mutual information between the leading eigenvectors spanning the inflation and unemployment manifolds.

This same logic can be extended to other macroeconomic models such as the Solow Model to construct evolving attractors describing economic growth in an equation-free setting. The form-free nature of TAR eliminates the need to introduce mercurial X-factors like productivity, allowing for a more direct understanding of the dynamical effects of changes in macroeconomic variables on economic growth.

TAR models currently lack explicit interpretability of leading eigenvectors, especially in complex systems. However, as we have demonstrated with currency systems, the judicious application of economic reasoning to these dynamics means that we can identify core aspects leading behind dynamics. My hope is that this class of geometry-based models, supported by statistical learning, finds a home in fields such as econometrics and biostatistics — disciplines rich in data but lacking robust mathematical descriptions of the complex phenomena they study. I plan to continue these

economic and biological investigations focused on precisely these types of questions.

8.2.4 *Topology, Topography and Holonomic Constraints*

There is no such thing as a free lunch. If we attempt to reconstruct an output with a higher Shannon entropy than its input, this extra information must come from somewhere. Statistical methods seek to bridge this information gap by studying large datasets describing that mapping. However, holonomic constraints can achieve this *without* the informational overhead of learning directly from the data. I am particularly excited about the applications of symmetry groups in MD, network structures in brain connectomics, and restrictions on simultaneous currency exchange rates, all of which can provide these types of holonomic constraints. This is a move towards mathematically rigorous bounding for learning problems around complex systems that lack clear-cut constraints such as Liouville's Theorem for classical physics. The topology of the manifolds describing dynamics under these constraints will inherently reflect them, suggesting that methods like TAR can be significantly improved by imposing those constraints.

Perhaps the most important insight I have gained over the course of my PhD stems from the intersection of embedding theorems and machine learning. Takens' Theorem states that a geometrical representation of full dimensional dynamics in the form of Takens' delay embeddings, which are diffeomorphically related to the true manifold embedding the full dimensional dynamics. Through years of observing how machines learn, combined with insights from information theory—such as the information bottleneck—I have come to understand that a central point in machine learning-based modeling has to be learning a representation of information that is diffeomorphically related to the quantity we wish to predict. Ambiguity in these relationships, in the form of non-injective or surjective mappings, introduces error into the modeling of deterministic phenomena. It is therefore paramount for the development of statistical learning for such systems to find ways to enforce constraints on a topological level. If we cannot do so directly with physical insights, TAR offers an alternative method to do so from data.

Chapter 5 highlights why, even if we can learn a manifold with the same topology as the manifold of full-dimensional dynamics, topographic deformations can still affect reconstruction quality. While this issue is explored in this work both empirically and mathematically, open questions remain regarding the optimization of learning techniques and data representations. It is my contention that a key to the future of machine learning lies in understanding topology and topography underlying the data we study.

What is the relationship between dynamics learned by different statistical learning techniques from the same data? Can we predict the optimal statistical learning tool for a given problem? Are there optimal metric spaces to describe dynamics or that facilitate learning? How can we physically constrain arbitrary dynamics to reduce solution space sizes? These questions are too important to be left unexplored.

8.3 Contradictory Advice

It is rare that life offers corner solutions. We live on a Pareto frontier and we are constantly balancing our time and efforts. Below I present a few of the main tradeoffs I have tried to balance over the course of my PhD, to varying degrees of efficacy. It is easy to go too far to one side in each of these so I hope that the reader will take this as an opportunity to reflect.

1. Think before you do. This will save you a lot of time (and compute).
2. Do not think too much otherwise you will not do.
3. Get to know your field and its literature well. Figure out what the interesting problems are before you start solving one that is perhaps not so interesting.
4. Do not stop yourself from contributing because there is more to learn. Sometimes the literature is missing something very important and you may do well pursuing it.

5. Make time for exercise and socializing each day to avoid burnout. It is beneficial to combine the two.
6. Use your limited time in grad school so spend it learning as much as possible. Classes can be good but I have found that finding the right mentors, people interested in your research problems, is critical. Always keep learning and find good people to learn with.
7. Be intellectually open and diverse. Many innovations come from collaboration and your colleagues are experts in fields you will struggle to grasp. You will learn a lot from them.
8. Learn a specialty well. Deep knowledge of a particular problem is required to make a deep contribution.
9. Only publish high quality work with a clean story supported by data and a scientifically rigorous explanation of that data.
10. No paper is a complete solution to a problem so do not stop yourself from completing a project because unanswered questions remain. There are always more unanswered questions.

REFERENCES

- Max Aalto and Nakul Verma. Metric learning on manifolds. [arXiv:1902.01738](https://arxiv.org/abs/1902.01738), page <https://doi.org/10.48550/arXiv.1902.01738>, 2019.
- 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.
- Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019. URL <https://arxiv.org/abs/1803.08375>.
- Agency for Toxic Substances and Disease Registry, United States Department of Health and Human Services. Per- and polyfluoroalkyl substances (pfas) and your health. <https://www.atsdr.cdc.gov/pfas/health-effects/index.html>, 2022. (accessed September 2022).
- Matteo Aldeghi, Joseph P Bluck, and Philip C Biggin. Absolute alchemical free energy calculations for ligand binding: A beginner’s guide. In *Computational Drug Discovery and Design*, pages 199–232. Springer, 2018.
- M P Allen and D J Tildesley. *Computer Simulations of Liquids*. Oxford University Press, 1989.
- Allonnia. Allonnia’s protein breakthrough brings commercial pfas sensor a step closer. <https://allonnia.com/wp-content/uploads/2022/08/Aug22-GWI-SWW-Allonnia.pdf>, 2022. (accessed September 2022).
- Natalia Alzate-Carvajal, Jaewoo Park, Ilhem Bargaoui, Ranjana Rautela, Zachary J Comeau, Lukas Scarfe, Jean-Michel Menard, Seth B Darling, Benoît H Lessard, and Adina Luican-Mayer. Arrays of functionalized graphene chemiresistors for selective sensing of volatile organic compounds. *ACS Appl. Electron. Mater.*, 5(3):1514–1520, 2023.
- Andrea Amadei, Antonius Linssen, and Herman JC Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425, 1993.
- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. In Doina Precup and Yeewhye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 1, pages 298–321, 2017.
- Ludwig Arnold. *Random Dynamical Systems*. Springer-Verlag, 1998.
- V. I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer-Verlag, 2nd edition, 1989.
- Wesley B. Asher, Peter Geggier, Michael D. Holsey, Grant T. Gilmore, Avik K. Pati, Jozsef Meszaros, Daniel S. Terry, Signe Mathiasen, Megan J. Kaliszewski, Mitchell D. McCauley, Alekhya Govindaraju, Zhou Zhou, Kaleeckal G. Harikumar, Khuloud Jaqaman, Laurence J. Miller, Adam W. Smith, Scott C. Blanchard, and Jonathan A. Javitch. Single-molecule fret imaging

- of gpcr dimers in living cells. Nature Methods, 18(4):397–405, Apr 2021. ISSN 1548-7105. doi:10.1038/s41592-021-01081-y. URL <https://doi.org/10.1038/s41592-021-01081-y>.
- Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. J. Mach. Learn. Res., 3(Nov):397–422, 2002.
- Erin S. Baker and Detlef R. U. Knappe. Per- and polyfluoroalkyl substances (pfas)—contaminants of emerging concern. Anal. Bioanal. Chem., 414(3):1187–1188, Jan 2022. ISSN 1618-2650. doi:10.1007/s00216-021-03811-9. URL <https://doi.org/10.1007/s00216-021-03811-9>.
- Albert-László Barabási. Linked: The New Science of Networks. Perseus Publishing, 2002.
- Krzysztof Barański, Yonatan Gutman, and Adam Śpiewak. A probabilistic takens theorem. Nonlinearity, 33(9):4940, jul 2020. doi:10.1088/1361-6544/ab8fb8. URL <https://doi.org/10.1088/1361-6544/ab8fb8>.
- Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. Phys. Rev. Lett., 100:020603, Jan 2008. doi:10.1103/PhysRevLett.100.020603. URL <https://link.aps.org/doi/10.1103/PhysRevLett.100.020603>.
- Maral Basma, S Sundara, Dilek Çalgan, Tereza Vernali, and Robert J Woods. Solvated ensemble averaging in the calculation of partial atomic charges. J. Comput. Chem., 22(11):1125–1137, 2001.
- Christopher I Bayly, Piotr Cieplak, Wendy Cornell, and Peter A Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. J. Phys. Chem., 97(40):10269–10280, 1993.
- Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15(6):1373–1396, 2003.
- H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. J. Chem. Phys., 81(8):3684–3690, 1984. doi:10.1063/1.448118. URL <https://doi.org/10.1063/1.448118>.
- H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, and J. Hermans. Interaction models for water in relation to protein hydration. In ‘Intermolecular Forces’, ed. B. Pullman, Reidel, Dordrecht, pages 331–342, 1981.
- Tyrus Berry and Suddhasattwa Das. Learning theory for dynamical systems. SIAM Journal on Applied Dynamical Systems, 22(3):2082–2122, 2023. doi:10.1137/22M1516865. URL <https://doi.org/10.1137/22M1516865>.

- Tom Bertalan, Felix Dietrich, and Ioannis G. Kevrekidis. Transformations between deep neural networks. CoRR, abs/2007.05646, 2020. URL <https://arxiv.org/abs/2007.05646>.
- Himaghna Bhattacharjee and Dionisios G Vlachos. Thermochemical data fusion using graph representation learning. J. Chem. Inf. Model., 60(10):4673–4683, 2020.
- G. D. Birkhoff. Proof of the ergodic theorem. Proceedings of the National Academy of Sciences, 17:656–660, 1931.
- Matthew P. Blakeley, Samar S. Hasnain, and Svetlana V. Antonyuk. Sub-atomic resolution x-ray crystallography and neutron crystallography: Promise, challenges and potential. IUCrJ, 2(4): 464–474, June 2015. doi:10.1107/s2052252515011239. URL <https://doi.org/10.1107/s2052252515011239>.
- Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provati, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A. Broglia, and Michele Parrinello. Plumed: A portable plugin for free-energy calculations with molecular dynamics. Comput. Phys. Commun., 180(10):1961–1972, 2009. ISSN 0010-4655. doi:<https://doi.org/10.1016/j.cpc.2009.05.011>. URL <https://www.sciencedirect.com/science/article/pii/S001046550900157X>.
- Jean-Philippe Bouchaud, Julius Bonart, Jonathan Donier, and Martin Gould. Trades, Quotes and Prices: Financial Markets Under the Microscope. Cambridge University Press, 2018.
- K. J. Bowers, D. E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, and D. E. Shaw. Scalable algorithms for molecular dynamics simulations on commodity clusters. In SC ’06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, pages 43–43, 2006.
- Nicole Marie Brennan, Abigail Teresa Evans, Meredith Kate Fritz, Stephanie Allison Peak, and Haley Elizabeth von Holst. Trends in the regulation of per- and polyfluoroalkyl substances (pfas): A scoping review. Int. J. Environ. Res. Public Health, 18(20):10900, Oct 2021. ISSN 1660-4601. doi:10.3390/ijerph182010900. URL <http://dx.doi.org/10.3390/ijerph182010900>.
- Michael Brin and Garrett Stuck. Introduction to Dynamical Systems. Cambridge University Press, 2002.
- D S Broomhead and Gregory P King. Extracting qualitative dynamics from experimental data. Physica D: Nonlinear Phenomena, 20(2–3):217–236, 1986. ISSN 0167-2789. doi:[http://dx.doi.org/10.1016/0167-2789\(86\)90031-X](http://dx.doi.org/10.1016/0167-2789(86)90031-X). URL <http://www.sciencedirect.com/science/article/pii/016727898690031X>.
- Charles D. Brummitt, Kyu-Min Lee, and K.-I. Goh. Multiplexity-facilitated cascades in networks. Phys. Rev. E, 85:045102, Apr 2012. doi:10.1103/PhysRevE.85.045102. URL <https://link.aps.org/doi/10.1103/PhysRevE.85.045102>.

- Robert C Buck, James Franklin, Urs Berger, Jason M Conder, Ian T Cousins, Pim De Voogt, Allan Astrup Jensen, Kurunthachalam Kannan, Scott A Mabury, and Stefan PJ van Leeuwen. Perfluoroalkyl and polyfluoroalkyl substances in the environment: terminology, classification, and origins. Integr. Environ. Assess. Manage., 7(4):513–541, 2011.
- Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. Nature Reviews Physics, 2(4):200–212, 2020.
- Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. J. Chem. Phys., 126(1):014101, 2007. doi:10.1063/1.2408420. URL <https://doi.org/10.1063/1.2408420>.
- J. C. Butcher. On runge-kutta processes of high order. Journal of the Australian Mathematical Society, 4(2):179–194, 1964. doi:10.1017/S1446788700023387.
- John Y. Campbell, Andrew W. Lo, A. Craig MacKinlay, and Robert F. Whitelaw. The econometrics of financial markets. Macroeconomic Dynamics, 2(4):559–562, 1998. doi:10.1017/S1365100598009092.
- L Cao, Alistair Mees, and Kevin Judd. Dynamics from multivariate time series. Physica D: Nonlinear Phenomena, 121(1–2):75–88, 1998. URL <http://www.sciencedirect.com/science/article/pii/S0167278998001511>.
- Liangyue Cao. Practical method for determining the minimum embedding dimension of a scalar time series. Physica D: Nonlinear Phenomena, 110(1–2):43–50, 1997. ISSN 0167-2789. doi:[http://dx.doi.org/10.1016/S0167-2789\(97\)00118-8](http://dx.doi.org/10.1016/S0167-2789(97)00118-8). URL <http://www.sciencedirect.com/science/article/pii/S0167278997001188>.
- Christine Cézard, Xavier Trivelli, Frédéric Aubry, Florence Djedaïni-Pilard, and François-Yves Dupradeau. Molecular dynamics studies of native and substituted cyclodextrins in different media: 1. charge derivation and force field performances. Phys. Chem. Chem. Phys., 13:15103–15121, 2011. doi:10.1039/C1CP20854C. URL <http://dx.doi.org/10.1039/C1CP20854C>.
- Clément Chadebec, Elina Thibeau-Sutre, Ninon Burgos, and Stéphanie Allasonnière. Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3):2879–2896, March 2023. ISSN 1939-3539. doi:10.1109/tpami.2022.3185773. URL <http://dx.doi.org/10.1109/TPAMI.2022.3185773>.
- Jerry C. Chang and Sandra J. Rosenthal. Real-Time Quantum Dot Tracking of Single Proteins, pages 51–62. Humana Press, Totowa, NJ, 2011. ISBN 978-1-61779-052-2. doi:10.1007/978-1-61779-052-2_4. URL https://doi.org/10.1007/978-1-61779-052-2_4.
- Ludovic Charleux, Emile Roux, Thibault Goyallon, Giovanni Feverati, and Pierre Nagorny. Lotka-volterra equations: Scientific python – a collection of science-oriented python examples docu-

- mentation. https://scientific-python.readthedocs.io/en/latest/notebooks_rst/3_Ordinary_Differential_Equations/02_Examples/Lotka_Volterra_model.html, 2024.
- Wei Chen and Andrew L. Ferguson. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *Journal of Computational Chemistry*, 39(25):2079–2102, 2018. doi:<https://doi.org/10.1002/jcc.25520>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.25520>.
- Wei Chen, Hythem Sidky, and Andrew L Ferguson. Nonlinear discovery of slow molecular modes using state-free reversible vampnets. *The Journal of Chemical Physics*, 150(21):214114, 2019.
- Zhonghang Chen, Yi-Lin Lu, Liang Wang, Jun Xu, Jing Zhang, Xiufang Xu, Peng Cheng, Sihai Yang, and Wei Shi. Efficient recognition and removal of persistent organic pollutants by a bifunctional molecular material. *Journal of the American Chemical Society*, 145(1):260–267, 2023. doi:10.1021/jacs.2c09866. URL <https://doi.org/10.1021/jacs.2c09866>. PMID: 36538618.
- Jie Cheng, Elefteria Psillakis, MR Hoffmann, and AJ Colussi. Acid dissociation versus molecular association of perfluoroalkyl oxoacids: environmental implications. *J. Phys. Chem. A*, 113(29): 8152–8156, 2009.
- Yifan Cheng. Single-particle cryo-EM at crystallographic resolution. *Cell*, 161(3):450–457, April 2015. doi:10.1016/j.cell.2015.03.049. URL <https://doi.org/10.1016/j.cell.2015.03.049>.
- Yu H Cheng, Dushyant Barpaga, Jennifer A Soltis, V Shutthanandan, Roli Kargupta, Kee Sung Han, B Peter McGrail, Radha Kishan Motkuri, Sagnik Basuray, and Sayandev Chatterjee. Metal–organic framework-based microfluidic impedance sensor platform for ultrasensitive detection of perfluorooctanesulfonate. *ACS Appl. Mater. Interfaces*, 12(9):10503–10514, 2020.
- Eliodoro Chiavazzo, Roberto Covino, Ronald R. Coifman, C. William Gear, Anastasia S. Georgiou, Gerhard Hummer, and Ioannis G. Kevrekidis. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, 114(28):E5494–E5503, 2017. doi:10.1073/pnas.1621481114. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1621481114>.
- Matthew Carter Childers and Valerie Daggett. Insights from molecular dynamics simulations for computational protein design. *Molecular Systems Design & Engineering*, 2:9–33, 2017. doi:10.1039/C6ME00083E. URL <http://dx.doi.org/10.1039/C6ME00083E>.
- Casey Ching, Max J Klemes, Brittany Trang, William R Dichtel, and Damian E Helbling. β -cyclodextrin polymers with different cross-linkers and ion-exchange resins exhibit variable adsorption of anionic, zwitterionic, and nonionic pfass. *Environ. Sci. Technol.*, 54(19):12693–12702, 2020.

- Casey Ching, Zhi-Wei Lin, William R Dichtel, and Damian E Helbling. Evaluating the performance of novel cyclodextrin polymer granules to remove perfluoroalkyl acids (pfaas) from water. ACS ES&T Engg., 3(5):661–670, 2023.
- Aditya Choudhary, Dengpan Dong, Marina Tsianou, Paschalis Alexandridis, and Dmitry Bedrov. Adsorption mechanism of perfluorooctanoate on cyclodextrin-based polymers: Probing the synergy of electrostatic and hydrophobic interactions with molecular dynamics simulations. ACS Materials Letters, 4(5):853–859, 2022. doi:10.1021/acsmaterialslett.2c00168. URL <https://doi.org/10.1021/acsmaterialslett.2c00168>.
- City of Chicago, Department of Water Management - Bureau of Water Supply. 2023 q1-q2 comprehensive chemical analysis. https://www.chicago.gov/city/en/depts/water/supp_info/water_quality_resultsandreports/comprehensive_chemicalanalysis.html, 2023. (accessed August 2023).
- Peter K. Clark. A subordinated stochastic process model with finite variance for speculative prices. Econometrica, 41(1):135–155, 1973. ISSN 00129682, 14680262. URL <http://www.jstor.org/stable/1913889>.
- R R Coifman, S Lafon, A B Lee, M Maggioni, B Nadler, F Warner, and S W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proceedings of the National Academy of Sciences of the United States of America, 102(21):7426–7431, May 2005a. ISSN 0027-8424. doi:10.1073/pnas.0500334102. URL <http://www.pnas.org/content/102/21/7426.abstract>.
- Ronald R Coifman and Stéphane Lafon. Diffusion maps. Applied and Computational Harmonic Analysis, 21(1):5–30, 2006.
- Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proceedings of the National Academy of Sciences of the United States of America, 102(21):7426–7431, 2005b.
- Ronald R Coifman, Ioannis G Kevrekidis, Stéphane Lafon, Mauro Maggioni, and Boaz Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. Multiscale Modeling & Simulation, 7(2):842–864, 2008a.
- Ronald R Coifman, Yoel Shkolnisky, Fred J Sigworth, and Amit Singer. Graph Laplacian tomography from unknown random projections. IEEE Transactions on Image Processing, 17(10):1891–1899, 2008b.
- CompTox Chemicals Dashboard, U.S. Environmental Protection Agency. Pfas: Listed in oecd global database. <https://comptox.epa.gov/dashboard/chemical-lists/pfasoecd>, 2022. (accessed September 2022).

- Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc., 117(19):5179–5197, 1995.
- I. P. Cornfeld and Y. G. Fomin, S. V. and Sinai. Ergodic Theory. Springer-Verlag, 1982.
- Ian T Cousins, Jana H Johansson, Matthew E Salter, Bo Sha, and Martin Scherlinger. Outside the safe operating space of a new planetary boundary for per-and polyfluoroalkyl substances (pfas). Environ. Sci. Technol., pages 11172–11179, 2022.
- Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory, 13:21–27, 1967.
- Dennis D. Cox and Susan John. Sdo: A statistical method for global optimization. In Multidisciplinary Design Optimization: State-of-the-Art, pages 315–329, 1997.
- G.M Crippen. Note rapid calculation of coordinates from distance matrices. Journal of Computational Physics, 26(3):449 – 452, 1978. ISSN 0021-9991. doi:[https://doi.org/10.1016/0021-9991\(78\)90081-5](https://doi.org/10.1016/0021-9991(78)90081-5). URL <http://www.sciencedirect.com/science/article/pii/0021999178900815>.
- Daniel J Cross and R Gilmore. Differential embedding of the Lorenz attractor. Physical Review E, 81(6):066220, 2010.
- Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An $n \cdot \log(n)$ method for ewald sums in large systems. J. Chem. Phys., 98(12):10089–10092, 1993. doi:10.1063/1.464397. URL <https://doi.org/10.1063/1.464397>.
- 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. Proceedings of the National Academy of Sciences of the United States of America, 103(26): 9885–9890, 2006.
- Siva Dasetty, Igor Coropceanu, Joshua Portner, Jiyuan Li, Juan J de Pablo, Dmitri Talapin, and Andrew L Ferguson. Active learning of polarizable nanoparticle phase diagrams for the guided design of triggerable self-assembling superlattices. Mol. Syst. Des. Eng., 7(4):350–363, 2022.
- Siva Dasetty, Maximilian Topel, Yifeng Oliver Tang, Yuqin Wang, Eric Jonas, Seth B. Darling, Junhong Chen, and Andrew L. Ferguson. Data-driven discovery of linear molecular probes with optimal selective affinity for pfas in water. Journal of Chemical & Engineering Data, 0(0):null, 2023. doi:10.1021/acs.jced.3c00404. URL <https://doi.org/10.1021/acs.jced.3c00404>.
- Vinicius Martins de Oliveira, Ruibin Liu, and Jana Shen. Constant ph molecular dynamics simulations: current status and recent applications. Curr. Opin. Struct. Biol., 77:102498, 2022.

Amila O De Silva, James M Armitage, Thomas A Bruton, Clifton Dassuncao, Wendy Heiger-Bernays, Xindi C Hu, Anna Kärrman, Barry Kelly, Carla Ng, Anna Robuck, Mei Sun, Thomas F. Webster, and Elsie M. Sunderland. Pfas exposure pathways for humans and wildlife: a synthesis of current knowledge and key gaps in understanding. Environ. Toxicol. Chem., 40(3):631–657, 2021.

Alexander P Demchenko. Photobleaching of organic fluorophores: Quantitative characterization, mechanisms, protection. Methods and Applications in Fluorescence, 8(2):022001, feb 2020. doi:10.1088/2050-6120/ab7365. URL <https://doi.org/10.1088/2050-6120/ab7365>.

Volker L. Deringer, Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian process regression for materials and molecules. Chem. Rev., 121(16): 10073–10141, 2021. doi:10.1021/acs.chemrev.1c00022. URL <https://doi.org/10.1021/acs.chemrev.1c00022>. PMID: 34398616.

Ethan R. Deyle and George Sugihara. Generalized theorems for nonlinear state space reconstruction. PLOS ONE, 6(3):1–8, 03 2011. doi:10.1371/journal.pone.0018295. URL <https://doi.org/10.1371/journal.pone.0018295>.

Mykola Dimura, Thomas-Otavio Peulen, Hugo Sanabria, Dmitro Rodnin, Katherina Hemmen, Christian A Hanke, Claus AM Seidel, and Holger Gohlke. Automated and optimally fret-assisted structural modeling. Nature Communications, 11(1):1–14, 2020.

Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: Essential theory, algorithms, and applications. IEEE Signal Processing Magazine, 32(6):12–30, Nov 2015. ISSN 1053-5888. doi:10.1109/msp.2015.2398954. URL <http://dx.doi.org/10.1109/MSP.2015.2398954>.

J. R. Dorfman. Kolmogorov–Sinai entropy, page 118–128. Cambridge Lecture Notes in Physics. Cambridge University Press, 1999.

Carmeline J. Dsilva, Ronen Talmon, Neta Rabin, Ronald R. Coifman, and Ioannis G. Kevrekidis. Nonlinear intrinsic variables and state reconstruction in multiscale simulations. The Journal of Chemical Physics, 139(18):184109, 2013. doi:10.1063/1.4828457. URL <https://doi.org/10.1063/1.4828457>.

François-Yves Dupradeau, Adrien Pigache, Thomas Zaffran, Corentin Savineau, Rodolphe Lelong, Nicolas Grivel, Dimitri Lelong, Wilfried Rosanski, and Piotr Cieplak. The r.e.d. tools: advances in resp and esp charge derivation and force field library building. Phys. Chem. Chem. Phys., 12:7821–7839, 2010. doi:10.1039/C0CP00111B. URL <http://dx.doi.org/10.1039/C0CP00111B>.

David Easley and Jon Kleinberg. Information cascades. In Networks, Crowds, and Markets: Reasoning about a Highly Connected World, chapter 16, pages 483–508. Cambridge University Press, 2010. URL <https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch16.pdf>.

- J. P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. Rev. Mod. Phys., 57: 617–656, Jul 1985. doi:10.1103/RevModPhys.57.617. URL <https://link.aps.org/doi/10.1103/RevModPhys.57.617>.
- J. P. Eckmann, S. Oliffson Kamphorst, D. Ruelle, and S. Ciliberto. Liapunov exponents from time series. Phys. Rev. A, 34:4971–4979, Dec 1986. doi:10.1103/PhysRevA.34.4971. URL <https://link.aps.org/doi/10.1103/PhysRevA.34.4971>.
- Sebastian Ehlert, Marcel Stahn, Sebastian Spicher, and Stefan Grimme. Robust and efficient implicit solvation model for fast semiempirical methods. J. Chem. Theory and Comput., 17(7): 4250–4261, 2021.
- Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh Ewald method. Journal of Chemical Physics, 103(19):8577–8593, 1995.
- D. J. Evans and B. L. Holian. The nose–hoover thermostat. J. Chem. Phys., 83(8):4069–4074, 1985. doi:10.1063/1.449071. URL <https://doi.org/10.1063/1.449071>.
- Eugene F. Fama. Random walks in stock market prices. Financial Analysts Journal, 51(1):75–80, 1995. doi:10.2469/faj.v51.n1.1861. URL <https://doi.org/10.2469/faj.v51.n1.1861>.
- Suzanne E Fenton, Alan Ducatman, Alan Boobis, Jamie C DeWitt, Christopher Lau, Carla Ng, James S Smith, and Stephen M Roberts. Per-and polyfluoroalkyl substance toxicity and human health review: Current state of knowledge and strategies for informing future research. Environ. Toxicol. Chem., 40(3):606–630, 2021.
- Andrew L Ferguson. Machine learning and data science in soft materials engineering. Journal of Physics: Condensed Matter, 30(4):043002, dec 2017. doi:10.1088/1361-648x/aa98bd. URL <https://doi.org/10.1088%2F1361-648x%2Faa98bd>.
- Andrew L Ferguson, Athanassios Z Panagiotopoulos, Pablo G Debenedetti, and Ioannis G Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. Proceedings of the National Academy of Sciences of the United States of America, 107(31):13597–13602, 2010.
- Andrew L Ferguson, Athanassios Z Panagiotopoulos, Ioannis G Kevrekidis, and Pablo G Debenedetti. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. Chemical Physics Letters, 509:1–11, 2011.
- E Megan Flynn, Jeffrey A Hanson, Tom Alber, and Haw Yang. Dynamic active-site protection by the m. tuberculosis protein tyrosine phosphatase ptpb lid domain. Journal of the American Chemical Society, 132(13):4772–4780, 2010.
- Etienne Forest and Ronald D. Ruth. Fourth-order symplectic integration. Physica D: Nonlinear

- Phenomena, 43(1):105–117, May 1990. doi:10.1016/0167-2789(90)90019-L. URL [https://doi.org/10.1016/0167-2789\(90\)90019-L](https://doi.org/10.1016/0167-2789(90)90019-L).
- A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. Phys. Rev. A, 33:1134, 1986a.
- Andrew M. Fraser and Harry L. Swinney. Independent coordinates for strange attractors from mutual information. Physical Review A, 33:1134–1140, Feb 1986b. doi:10.1103/PhysRevA.33.1134. URL <http://link.aps.org/doi/10.1103/PhysRevA.33.1134>.
- Daan Frenkel and Berend Smit. Understanding Molecular Simulation: From algorithms to applications, volume 1. Academic press, 2002.
- M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian~16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.
- Yann Gambin and Ashok A. Deniz. Multicolor single-molecule fret to explore protein folding and binding. Molecular BioSystems, 6:1540–1547, 2010. doi:10.1039/C003024D. URL <http://dx.doi.org/10.1039/C003024D>.
- Angel E García. Large-amplitude nonlinear motions in proteins. Physical Review Letters, 68(17):2696, 1992.
- Ignacio J General. A note on the standard state’s binding free energy. Journal of Chemical Theory and Computation, 6(8):2520–2524, 2010.
- Thomas Gensch, Martin Böhmer, and Pedro F. Aramendía. Single molecule blinking and photo-bleaching separated by wide-field fluorescence microscopy. The Journal of Physical Chemistry A, 109(30):6652–6658, 2005. doi:10.1021/jp0510847. URL <https://doi.org/10.1021/jp0510847>.
- Rossella Ghisi, Teofilo Vamerali, and Sergio Manzetti. Accumulation of perfluorinated alkyl substances (pfas) in agricultural plants: A review. Environ. Res., 169:326–341, 2019.
- Juliane Glüge, Martin Scherlinger, Ian T Cousins, Jamie C DeWitt, Gretta Goldenman, Dorte Herzke, Rainer Lohmann, Carla A Ng, Xenia Trier, and Zhanyun Wang. An overview of the uses of

- per-and polyfluoroalkyl substances (pfas). Environ. Sci.: Processes Impacts, 22(12):2345–2373, 2020.
- Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent. Sci., 4(2):268–276, 2018. ISSN 2374-7943, 2374-7951. doi:10.1021/acscentsci.7b00572. URL <https://pubs.acs.org/doi/10.1021/acscentsci.7b00572>.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2015.
- Grebogi, Ott, and Yorke. Critical exponents for crisis-induced intermittency. Physical Review Letters, 48:1507–1510, 1983.
- Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 30: 5768–5778, 2017.
- Taekjip Ha and Philip Tinnefeld. Photophysics of fluorescence probes for single molecule biophysics and super-resolution imaging. Annual Review of Physical Chemistry, 63:595–617, 2012.
- Yazan Haddad, Vojtech Adam, and Zbynek Heger. Rotamer dynamics: Analysis of rotamers in molecular dynamics simulations of proteins. Biophysical Journal, 116(11):2062–2072, 2019. ISSN 0006-3495. doi:<https://doi.org/10.1016/j.bpj.2019.04.017>. URL <https://www.sciencedirect.com/science/article/pii/S0006349519303352>.
- Hermann Haken. Synergetics: An Introduction. Springer Berlin Heidelberg, 1983. ISBN 978-3-642-88340-8. doi:10.1007/978-3-642-88338-5. URL <https://doi.org/10.1007/978-3-642-88338-5>.
- Jiequn Han, Linfeng Zhang, Roberto Car, and Weinan E. Deep potential: A general representation of a many-body potential energy surface. Communications in Computational Physics, 23(3), 2018.
- Wei Han and Klaus Schulten. Fibril elongation by ab17–42: Kinetic network analysis of hybrid-resolution molecular dynamics simulations. Journal of the American Chemical Society, 136(35): 12450–12460, 2014. doi:10.1021/ja507002p. URL <https://doi.org/10.1021/ja507002p>.
- Peter Reinhard Hansen and Asger Lunde. Realized variance and market microstructure noise. Journal of Business & Economic Statistics, 24(2):127–161, 2006. doi:10.1198/073500106000000071. URL <https://doi.org/10.1198/073500106000000071>.
- Marcus D Hanwell, Donald E Curtis, David C Lonie, Tim Vandermeersch, Eva Zurek, and Geof-

- frey R Hutchison. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminf.*, 4(1):1–17, 2012a.
- Marcus D. Hanwell, Donald E. Curtis, David C. Lonie, Tim Vandermeersch, Eva Zurek, and Geoffrey R. Hutchison. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 4(1):17, 2012b. doi:10.1186/1758-2946-4-17. URL <https://doi.org/10.1186/1758-2946-4-17>.
- Boris Hasselblatt and Anatolij Borisovic Katok. A first course in dynamics: With a panorama of recent developments. Cambridge University Press, 2010.
- Mohamad H. Hassoun, Nathan Intrator, Susan McKay, and Wolfgang Christian. Fundamentals of artificial neural networks. *Computers in Physics*, 10(2):137–137, 1996. doi:10.1063/1.4822376. URL <https://aip.scitation.org/doi/abs/10.1063/1.4822376>.
- Michael Hauser and Asok Ray. Principles of riemannian geometry in neural networks. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30 (NIPS 2017), pages 2808–2817, 2017.
- Xibing He, Viet H. Man, Wei Yang, Tai-Sung Lee, and Junmei Wang. A fast and high-quality charge model for the next generation general AMBER force field. *The Journal of Chemical Physics*, 153(11):114502, 09 2020. ISSN 0021-9606. doi:10.1063/5.0019056. URL <https://doi.org/10.1063/5.0019056>.
- Rainer Hegger, Alexandros Altis, Phuong H Nguyen, and Gerhard Stock. How complex is the dynamics of peptide folding? *Physical Review Letters*, 98(2):028102, 2007.
- Björn Hellenkamp, Sonja Schmid, Olga Doroshenko, Oleg Opanasyuk, Ralf Kühnemuth, Soheila Rezaei Adariani, Benjamin Ambrose, Mikayel Aznauryan, Anders Barth, Victoria Birkedal, Mark E. Bowen, Hongtao Chen, Thorben Cordes, Tobias Eilert, Carel Fijen, Christian Gebhardt, Markus Götz, Giorgos Gouridis, Enrico Gratton, Taekjip Ha, Pengyu Hao, Christian A. Hanke, Andreas Hartmann, Jelle Hendrix, Lasse L. Hildebrandt, Verena Hirschfeld, Johannes Hohlbein, Boyang Hua, Christian G. Hübner, Eleni Kallis, Achillefs N. Kapanidis, Jae-Yeol Kim, Georg Krainer, Don C. Lamb, Nam Ki Lee, Edward A. Lemke, Brié Levesque, Marcia Levitus, James J. McCann, Nikolaus Naredi-Rainer, Daniel Nettels, Thuy Ngo, Ruoyi Qiu, Nicole C. Robb, Carlheinz Röcker, Hugo Sanabria, Michael Schlierf, Tim Schröder, Benjamin Schuler, Henning Seidel, Lisa Streit, Johann Thurn, Philip Tinnefeld, Swati Tyagi, Niels Vandenberg, Andrés Manuel Vera, Keith R. Weninger, Bettina Wünsch, Inna S. Yanez-Orozco, Jens Michaelis, Claus A. M. Seidel, Timothy D. Craggs, and Thorsten Hugel. Precision and accuracy of single-molecule fret measurements—a multi-laboratory benchmark study. *Nature Methods*, 15(9):669–676, Sep 2018. ISSN 1548-7105. doi:10.1038/s41592-018-0085-0. URL <https://doi.org/10.1038/s41592-018-0085-0>.
- Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450

- (7172):964–972, 2007. doi:10.1038/nature06522. URL <https://doi.org/10.1038/nature06522>.
- Jan Hermans and Lu Wang. Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding. application to a complex of benzene and mutant t4 lysozyme. Journal of the American Chemical Society, 119(11):2707–2714, 1997. doi:10.1021/ja963568+. URL <https://doi.org/10.1021/ja963568+>.
- Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. J. Comput. Chem., 18(12):1463–1472, 1997. doi:[https://doi.org/10.1002/\(SICI\)1096-987X\(199709\)18:12<1463::AID-JCC4>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H). URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291096-987X%28199709%2918%3A12%3C1463%3A%3AAID-JCC4%3E3.0.CO%3B2-H>.
- Morris Hirsch. Differential Topology, volume 33. Springer-Verlag, 01 1994. ISBN 978-1-4684-9451-8. doi:10.1007/978-1-4684-9449-5.
- Morris W. Hirsch, Robert L. Devaney, and Stephen Smale. Differential Equations, Dynamical Systems, and an Introduction to Chaos. Academic Press, 2nd edition, 2003.
- Roger W Hockney and James W Eastwood. Computer Simulation Using Particles. CRC Press, 2010.
- 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>.
- Scott A. Hollingsworth and Ron O. Dror. Molecular dynamics simulation for all. Neuron, 99(6):1129–1143, Sep 2018. ISSN 0896-6273. doi:10.1016/j.neuron.2018.08.011. URL <https://doi.org/10.1016/j.neuron.2018.08.011>.
- Shinya Honda, Kazuhiko Yamasaki, Yoshito Sawada, and Hisayuki Morii. 10 residue folded peptide designed by segment statistics. Structure (London, England : 1993), 12(8):1507—1518, August 2004. ISSN 0969-2126. doi:10.1016/j.str.2004.05.022. URL <https://doi.org/10.1016/j.str.2004.05.022>.
- William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. Physical Review A, 31:1695–1697, Mar 1985. doi:10.1103/PhysRevA.31.1695. URL <http://link.aps.org/doi/10.1103/PhysRevA.31.1695>.
- Fang Huang, Giovanni Settanni, and Alan R. Fersht. Fluorescence resonance energy transfer analysis of the folding pathway of Engrailed Homeodomain. Protein Engineering, Design and Selection, 21(3):131–146, 01 2008. ISSN 1741-0126. doi:10.1093/protein/gzm069. URL <https://doi.org/10.1093/protein/gzm069>.

- William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. J. Mol. Graphics, 14(1):33–38, 1996. doi:[https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5). URL <https://www.sciencedirect.com/science/article/pii/0263785596000185>.
- H. E. Hurst. The problem of long-term storage in reservoirs. International Association of Scientific Hydrology. Bulletin, 1(3):13–27, 1956. doi:10.1080/02626665609493644.
- H. E. Hurst. A suggested statistical model of some time series which occur in nature. Nature, 180: 494, 1957. doi:10.1038/180494a0.
- Toshiko Ichiye and Martin Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. Proteins: Structure, Function, and Bioinformatics, 11(3):205–217, 1991. doi:<https://doi.org/10.1002/prot.340110305>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340110305>.
- John J. Irwin and Brian K. Shoichet. ZINC—a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model., 45(1):177–182, 2005 Jan-Feb. ISSN 1549-9596. doi:10.1021/ci049714+.
- John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. J. Chem. Inf. Model., 52(7):1757–1768, 2012. ISSN 1549-9596. doi:10.1021/ci3001277. URL <https://doi.org/10.1021/ci3001277>.
- John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. J. Chem. Inf. Model., 60(12): 6065–6073, 2020. ISSN 1549-9596. doi:10.1021/acs.jcim.0c00675. URL <https://doi.org/10.1021/acs.jcim.0c00675>.
- Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065):20150202, April 2016. doi:10.1098/rsta.2015.0202. URL <https://doi.org/10.1098/rsta.2015.0202>.
- 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, 1983a.
- William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. Journal of Chemical Physics, 79(2):926–935, 1983b.
- Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 32(5):922–923, 1976.

- Holger Kantz and Thomas Schreiber. Nonlinear Time Series Analysis. Cambridge University Press, 2004.
- Holger Kantz and Thomas Schreiber. Nonlinear Time Series Analysis. Cambridge, 2nd edition, 2005.
- Matthew B Kennel, Reggie Brown, and Henry DI Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. Physical Review A, 45(6):3403, 1992.
- Ioannis G. Kevrekidis, C. William Gear, and Gerhard Hummer. Equationfree Modeling For Complex Systems, pages 1453–1475. Springer Netherlands, Dordrecht, 2005. ISBN 978-1-4020-3286-8. doi:10.1007/978-1-4020-3286-8_72. URL https://doi.org/10.1007/978-1-4020-3286-8_72.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.
- Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. <http://arxiv.org/abs/1312.6114>, 2022.
- Karl N Kirschner and Robert J Woods. Quantum mechanical study of the nonbonded forces in water- methanol complexes. J. Phys. Chem. A, 105(16):4150–4155, 2001a.
- Karl N Kirschner and Robert J Woods. Solvent interactions determine carbohydrate conformation. Proc. Natl. Acad. Sci. U. S. A., 98(19):10541–10545, 2001b.
- Tuomas P. J. Knowles, Michele Vendruscolo, and Christopher M. Dobson. The amyloid state and its association with protein misfolding diseases. Nature Reviews Molecular Cell Biology, 15(6): 384–396, Jun 2014. ISSN 1471-0080. doi:10.1038/nrm3810. URL <https://doi.org/10.1038/nrm3810>.
- Shruti Koulgi, Uddhaves Sonavane, and Rajendra Joshi. Insights into the folding pathway of the engrailed homeodomain protein using replica exchange molecular dynamics simulations. Journal of Molecular Graphics and Modelling, 29(3):481–491, 2010. ISSN 1093-3263. doi:<https://doi.org/10.1016/j.jmgm.2010.09.007>. URL <https://www.sciencedirect.com/science/article/pii/S1093326310001427>.
- Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. Mach. Learn.: Sci. Technol., 1(4):045024, 2020.
- Jan Kubelka, William A. Eaton, and James Hofrichter. Experimental tests of villin subdomain folding simulations. Journal of Molecular Biology, 329(4):625–630, 2003. ISSN 0022-2836. doi:[https://doi.org/10.1016/S0022-2836\(03\)00519-9](https://doi.org/10.1016/S0022-2836(03)00519-9). URL <https://www.sciencedirect.com/science/article/pii/S0022283603005199>.

- Jan Kubelka, Thang K. Chiu, David R. Davies, William A. Eaton, and James Hofrichter. Sub-microsecond protein folding. *Journal of Molecular Biology*, 359(3):546–553, 2006. ISSN 0022-2836. doi:<https://doi.org/10.1016/j.jmb.2006.03.034>. URL <https://www.sciencedirect.com/science/article/pii/S0022283606003706>.
- M. Kurplus and J. A. McCammon. Dynamics of proteins: Elements and function. *Annual Review of Biochemistry*, 52(Volume 52, 1983):263–300, 1983. ISSN 1545-4509. doi:<https://doi.org/10.1146/annurev.bi.52.070183.001403>. URL <https://www.annualreviews.org/content/journals/10.1146/annurev.bi.52.070183.001403>.
- Carol F Kwiatkowski, David Q Andrews, Linda S Birnbaum, Thomas A Bruton, Jamie C DeWitt, Detlef RU Knappe, Maricel V Maffini, Mark F Miller, Katherine E Pelch, Anna Reade, Anna Soehl, Xenia Trier, and Marta Venier. Scientific basis for managing pfas as a chemical class. *Environ. Sci. Technol. Lett.*, 7(8):532–543, 2020.
- CR Laing, TA Frewen, and IG Kevrekidis. Coarse-grained dynamics of an activity bump in a neural field model. *Nonlinearity*, 20(9):2127, 2007.
- David J Lampert, Michael A Frisch, and Gerald E Speitel Jr. Removal of perfluorooctanoic acid and perfluorooctane sulfonate from wastewater by ion exchange. *Pract. Period. Hazard., Toxic, Radioact. Waste Manage.*, 11(1):60–68, 2007.
- Greg Landrum, Paolo Tosco, Brian Kelley, sriniker, Ric, gedec, Riccardo Vianello, NadineSchneider, Andrew Dalke, Dan N, Eisuke Kawashima, Brian Cole, Samo Turk, Matt Swain, Alexander-Savelyev, David Cosgrove, Alain Vaucher, Maciej Wójcikowski, Gareth Jones, Daniel Probst, guillaume godin, Vincent F. Scalfani, Axel Pahl, Francois Berenger, JLVarjo, strets123, JP, Do-liathGavid, Gianluca Sforza, and Jan Holst Jensen. rdkit/rdkit: 2022_09_5 (q3 2022) release, March 2023. URL <https://doi.org/10.5281/zenodo.7671152>.
- L Lankiewicz, J Malicka, and W Wicz. Fluorescence resonance energy transfer in studies of inter-chromophoric distances in biomolecules. *Acta biochimica Polonica*, 44(3):477–489, 1997. ISSN 0001-527X. URL <http://europepmc.org/abstract/MED/9511959>.
- Hongxing Lei, Chun Wu, Haiguang Liu, and Yong Duan. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 104(12):4925–4930, 2007. doi:10.1073/pnas.0608432104. URL <https://www.pnas.org/doi/abs/10.1073/pnas.0608432104>.
- C Letellier and G Gouesbet. Topological characterization of reconstructed attractors modding out symmetries. *Journal de Physique II*, 6(11):1615–1638, 1996.
- C Letellier, J Maquet, L Le Sceller, G Gouesbet, and LA Aguirre. On the non-equivalence of observables in phase-space reconstructions from recorded time series. *Journal of Physics A: Mathematical and General*, 31(39):7913, 1998.

- Chenghan Li and Gregory A Voth. Accurate and transferable reactive molecular dynamics models from constrained density functional theory. *J. Phys. Chem. B*, 125(37):10471–10480, 2021.
- Chun-Guang Li, Jun Guo, Guang Chen, Xiang-Fei Nie, and Zhen Yang. A version of isomap with explicit mapping. In *IEEE International Conference on Machine Learning and Cybernetics*, pages 3201–3206, 2006.
- Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences*, 110(16):6358–6363, 2013. doi:10.1073/pnas.1303186110. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1303186110>.
- Claudia Lindim, J Van Gils, and Ian T Cousins. Europe-wide estuarine export and surface water concentrations of pfos and pfoa. *Water research*, 103:124–132, 2016.
- K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334:517– 520, 2011. doi:10.1126/science.1208351.
- Yang Liu, Jeehae Park, Karin A. Dahmen, Yann R. Chemla, and Taekjip Ha. A comparative study of multivariate and univariate hidden markov modelings in time-binned single-molecule fret data analysis. *The Journal of Physical Chemistry B*, 114(16):5386–5403, 2010. doi:10.1021/jp9057669. URL <https://doi.org/10.1021/jp9057669>.
- Andrew W. Long and Andrew L. Ferguson. Landmark diffusion maps (l-dmaps): Accelerated manifold learning out-of-sample extension. *Applied and Computational Harmonic Analysis*, 47(1):190–211, 2019a. ISSN 1063-5203. doi:<https://doi.org/10.1016/j.acha.2017.08.004>. URL <https://www.sciencedirect.com/science/article/pii/S1063520317300921>.
- Andrew W Long and Andrew L Ferguson. Landmark diffusion maps (L-dmaps): Accelerated manifold learning out-of-sample extension. *Applied and Computational Harmonic Analysis*, 47(1):190–211, 2019b.
- Peter Y. Lu, Joan Ariño Bernad, and Marin Soljacić. Discovering sparse interpretable dynamics from partial observations. *Communications Physics*, 5(1):206, 2022. doi:10.1038/s42005-022-00987-z. URL <https://doi.org/10.1038/s42005-022-00987-z>.
- Ananth Madhavan. Market microstructure: A survey. *Journal of Financial Markets*, 3(3):205–258, 2000. ISSN 1386-4181. doi:[https://doi.org/10.1016/S1386-4181\(00\)00007-0](https://doi.org/10.1016/S1386-4181(00)00007-0). URL <https://www.sciencedirect.com/science/article/pii/S1386418100000070>.
- Douglas Magde, Gail E. Rojas, and Paul G. Seybold. Solvent dependence of the fluorescence lifetimes of xanthene dyes. *Photochemistry and Photobiology*, 70(5):737–744, 1999. doi:<https://doi.org/10.1111/j.1751-1097.1999.tb08277.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-1097.1999.tb08277.x>.

- Aleksandr V Marenich, Christopher J Cramer, and Donald G Truhlar. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J. Phys. Chem. B, 113(18):6378–6396, 2009.
- Aleksandr V Marenich, Christopher J Cramer, and Donald G Truhlar. Generalized born solvation model sm12. J. Chem. Theory and Comput., 9(1):609–620, 2013.
- Marcus G Martin and J Ilja Siepmann. Transferable potentials for phase equilibria. 1. united-atom description of n-alkanes. Journal of Physical Chemistry B, 102(14):2569–2577, 1998.
- Robert M. May. Simple mathematical models with very complicated dynamics. Nature, 261(5560):459–467, 1976.
- Ugo Mayor, Christopher M. Johnson, Valerie Daggett, and Alan R. Fersht. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. Proceedings of the National Academy of Sciences of the United States of America, 97(25):13518–13522, 2000. doi:10.1073/pnas.250473497. URL <https://www.pnas.org/doi/abs/10.1073/pnas.250473497>.
- Jay N Meegoda, Jitendra A Kewalramani, Brian Li, and Richard W Marsh. A review of the applications, environmental release, and remediation technologies of per-and polyfluoroalkyl substances. Int. J. Environ. Res. Public Health, 17(21):8117, 2020.
- Melanie Mitchell. Complexity: A Guided Tour. Oxford University Press, 2009.
- Shriyaa Mittal and Diwakar Shukla. Maximizing kinetic information gain of markov state models for optimal design of spectroscopy experiments. The Journal of Physical Chemistry B, 122(48):10793–10805, 2018. doi:10.1021/acs.jpcc.8b07076.
- Bernadette Mohr, Kirill Shmilovich, Isabel S Kleinwächter, Dirk Schneider, Andrew L Ferguson, and Tristan Bereau. Data-driven discovery of cardiolipin-selective small molecules by computational active learning. Chem. Sci., 13(16):4498–4511, 2022.
- AA Moosavi-Movahedi, M Gharanfoli, K Nazari, M Shamsipur, J Chamani, B Hemmateenejad, M Alavi, A Shokrollahi, M Habibi-Rezaei, C Sorenson, and N Sheibani. A distinct intermediate of rnase a is induced by sodium dodecyl sulfate at its pka. Colloids Surf., B, 43(3-4):150–157, 2005.
- B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. In Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference (Neural Information Processing), pages 955–962. The MIT Press, 2006a. ISBN 0262232537.
- Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps,

- spectral clustering and reaction coordinates of dynamical systems. Applied and Computational Harmonic Analysis, 21(1):113–127, 2006b.
- Richard M. Neumann. Entropic approach to brownian movement. Am. J. Phys., 48(5):354–357, 1980. doi:10.1119/1.12095. URL <https://doi.org/10.1119/1.12095>.
- Eyal Nir, Xavier Michalet, Kambiz M Hamadani, Ted A Laurence, Daniel Neuhauser, Yevgeniy Kovchegov, and Shimon Weiss. Shot-noise limited single-molecule fret histograms: Comparison between theory and experiments. The Journal of Physical Chemistry B, 110(44):22103–22124, 2006.
- Emmy Noether. Invariante variationsprobleme. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse 1918, pages 235–257, 1918.
- Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. Journal of Chemical Physics, 81(1):511–519, 1984.
- Maureen O’Hara. Market microstructure theory. Blackwell Publishers, 1995. ISBN 1557864438.
- Hiroyuki Oikawa, Takumi Takahashi, Supawich Kamonprasertsuk, and Satoshi Takahashi. Microsecond resolved single-molecule fret time series measurements based on the line confocal optical system combined with hybrid photodetectors. Physical Chemistry Chemical Physics, 20:3277–3285, 2018. doi:10.1039/C7CP06268K. URL <http://dx.doi.org/10.1039/C7CP06268K>.
- V. I. Oseledets. A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems. Trans. Moscow Math. Soc., 19:197–231, 1968.
- NH Packard, JP Crutchfield, JD Farmer, and RS Shaw. Geometry from a time series. Physical Review Letters, 45(9):712–716, 1980. URL <http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.45.712>.
- Szilárd Páll, Mark J. Abraham, Carsten Kutzner, Berk Hess, and Erik Lindahl. Tackling exascale software challenges in molecular dynamics simulations with gromacs. In Stefano Markidis and Erwin Laure, editors, Solving Software Challenges for Exascale, volume 8759, pages 3–27. Springer International Publishing, Cham, 2014. ISBN 978-3-319-15975-1. doi:<https://doi.org/10.1007/978-3-319-15976-8>.
- Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. A flexible framework for multi-objective bayesian optimization using random scalarizations. In Uncertainty in Artificial Intelligence, pages 766–776. PMLR, 2020.
- M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. J. Appl. Phys., 52(12):7182–7190, December 1981. doi:10.1063/1.328693.

- Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. Journal of Applied Physics, 52(12):7182–7190, 1981.
- Fabian Pedregosa, Gael 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 Edouard Duchesnay. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res., 12:2825–2830, 2011. ISSN 15324435. URL <http://proxy.uchicago.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=70109929&site=eds-live&scope=site>.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, 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, 2018. URL <https://arxiv.org/abs/1201.0490>.
- Dominique Perrault-Joncas and Marina Meilă. Metric learning on manifolds. Journal of Machine Learning Research, 0(0):0–00, 2011. URL <https://sites.stat.washington.edu/mmp/Papers/RMetric.pdf>.
- Ya B. Pesin. Characteristic lyapunov exponents and smooth ergodic theory. Russian Mathematical Surveys, 32(4):55–114, August 1977. ISSN 0036-0279. doi:10.1070/RM1977v032n04ABEH001639.
- Erez Peterfreund, Ofir Lindenbaum, Felix Dietrich, Tom Bertalan, Matan Gavish, Ioannis G. Kevrekidis, and Ronald R. Coifman. Local conformal autoencoder for standardized data coordinates. Proceedings of the National Academy of Sciences of the United States of America, 117(49):30918–30927, 2020. doi:10.1073/pnas.2014627117. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2014627117>.
- Jim Pfandtnr and Massimiliano Bonomi. Efficient sampling of high-dimensional free-energy landscapes with parallel bias metadynamics. J. Chem. Theory Comput., 11(11):5062–5067, 2015. doi:10.1021/acs.jctc.5b00846. URL <https://doi.org/10.1021/acs.jctc.5b00846>. PMID: 26574304.
- Carey Phelps, Brett Israels, Davis Jose, Morgan C. Marsh, Peter H. von Hippel, and Andrew H. Marcus. Using microsecond single-molecule fret to determine the assembly pathways of t4 ssdna binding protein onto model dna replication forks. Proceedings of the National Academy of Sciences of the United States of America, 114(18):E3612–E3621, 2017. doi:10.1073/pnas.1619819114. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1619819114>.
- Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. How robust are protein folding simulations with respect to force field parameterization? Biophysical Journal, 100(9):L47–9, May 2011. ISSN 0006-3495. doi:10.1016/j.bpj.2011.03.051. URL <https://europepmc.org/articles/PMC3149239>.

- Fabio Pietrucci. Novel enhanced sampling strategies for transitions between ordered and disordered structures. Handbook of Materials Modeling: Methods: Theory and Modeling, pages 597–619, 2020.
- Menahem Pirchi, Roman Tsukanov, Rashid Khamis, Toma E Tomov, Yaron Berger, Dinesh Chandra Khara, Hadas Volkov, Gilad Haran, and Eyal Nir. Photon-by-photon hidden markov model analysis for microsecond single-molecule fret kinetics. The Journal of Physical Chemistry B, 120 51:13065–13075, 2016.
- Stefano Raniolo and Vittorio Limongelli. Ligand binding free-energy calculations with funnel metadynamics. Nat. Protoc., 15(9):2837–2866, 2020.
- Carl Edward Rasmussen and Chiristopher K. Williams. Gaussian Processes for Machine Learning. Adaptive computation and machine learning. MIT Press, Cambridge, MA, 2005. ISBN ISBN 0-262-18253-X. URL <https://www.GaussianProcess.org/gpml>.
- Ines Reinartz, Marie Weiel, and Alexander Schug. Fret dyes significantly affect saxs intensities of proteins. Israel Journal of Chemistry, 60(7):725–734, 2020. doi:<https://doi.org/10.1002/ijch.202000007>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/ijch.202000007>.
- H.-J Reinhardt. Analysis of Approximation Methods for Differential and Integral Equations. Springer, 1985.
- Ricardo D. Righetto, Nikhil Biyani, Julia Kowal, Mohamed Chami, and Henning Stahlberg. Retrieving high-resolution information from disordered 2d crystals by single-particle cryo-EM. Nature Communications, 10(1), April 2019. doi:10.1038/s41467-019-09661-5. URL <https://doi.org/10.1038/s41467-019-09661-5>.
- Valerio Rizzi, Luigi Bonati, Narjes Ansari, and Michele Parrinello. The role of water in host-guest interaction. Nat. Commun., 12(1):93, 2021.
- João R. Robalo, Louis M. Streacker, Denilson Mendes de Oliveira, Petra Imhof, Dor Ben-Amotz, and Ana Vila Verde. Hydrophobic but water-friendly: Favorable water–perfluoromethyl interactions promote hydration shell defects. J. Am. Chem. Soc., 141(40):15856–15868, 2019. doi:10.1021/jacs.9b06862. URL <https://doi.org/10.1021/jacs.9b06862>. PMID: 31503489.
- João R. Robalo, Denilson Mendes de Oliveira, Petra Imhof, Dor Ben-Amotz, and Ana Vila Verde. Quantifying how step-wise fluorination tunes local solute hydrophobicity, hydration shell thermodynamics and the quantum mechanical contributions of solute–water interactions. Phys. Chem. Chem. Phys., 22:22997–23008, 2020. doi:10.1039/D0CP04205F. URL <http://dx.doi.org/10.1039/D0CP04205F>.
- Mary A Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of

- reaction coordinates via locally scaled diffusion map. Journal of Chemical Physics, 134(12): 124116, 2011.
- Michael T. Rosenstein, James J. Collins, and Carlo J. De Luca. A practical method for calculating largest lyapunov exponents from small data sets. Physica D: Nonlinear Phenomena, 65(1): 117–134, 1993. ISSN 0167-2789. doi:[https://doi.org/10.1016/0167-2789\(93\)90009-P](https://doi.org/10.1016/0167-2789(93)90009-P). URL <https://www.sciencedirect.com/science/article/pii/016727899390009P>.
- Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500):2323–2326, 2000a. doi:10.1126/science.290.5500.2323. URL <https://www.science.org/doi/abs/10.1126/science.290.5500.2323>.
- Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500):2323–2326, 2000b.
- Rahul Roy, Sungchul Hohng, and Taekjip Ha. A practical guide to single-molecule FRET. Nature Methods, 5(6):507–516, 06 2008. URL <http://dx.doi.org/10.1038/nmeth.1208>.
- Walter Rudin. Principles of Mathematical Analysis. McGraw-Hill, New York, 3rd edition, 1976. ISBN 9780070542358.
- David Ruelle. An inequality for the entropy of differentiable maps. Boletim da Sociedade Brasileira de Matemática - Bulletin/Brazilian Mathematical Society, 9(1):83–87, 1978. doi:10.1007/BF02584795. URL <https://doi.org/10.1007/BF02584795>.
- David Ruelle. Ergodic theory of differentiable dynamical systems. Publications Mathématiques de l’IHÉS, 50:27–58, 1979.
- Clemens-Bässem Salem, Evelyn Ploetz, and Don C. Lamb. Chapter 2 - probing dynamics in single molecules. In Carey K. Johnson, editor, Spectroscopy and Dynamics of Single Molecules, Developments in Physical & Theoretical Chemistry, pages 71–115. Elsevier, 2019. ISBN 978-0-12-816463-1. doi:<https://doi.org/10.1016/B978-0-12-816463-1.00002-X>. URL <https://www.sciencedirect.com/science/article/pii/B978012816463100002X>.
- 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 (ICTAI), 2004, pages 576–584. IEEE, 2004.
- Dewi Retno Sari Saputro and Purnami Widyaningsih. Limited memory broyden-fletcher-goldfarb-shanno (l-bfgs) method for the parameter estimation on geographically weighted ordinal logistic regression model (gwolr). AIP Conference Proceedings, 1868(1):040009, 2017. doi:10.1063/1.4995124. URL <https://aip.scitation.org/doi/abs/10.1063/1.4995124>.
- Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. Journal of Statistical Physics, 65 (3-4):579–616, 1991.

- Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009. doi:10.1038/nature08227. URL <https://doi.org/10.1038/nature08227>.
- B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 2002.
- Matthias Scholz, Martin Fraunholz, and Joachim Selbig. Nonlinear principal component analysis: Neural network models and applications. In Alexander N. Gorban, Balázs Kégl, Donald C. Wunsch, and Andrei Y. Zinovyev, editors, *Principal Manifolds for Data Visualization and Dimension Reduction*, volume 58 of *Lecture Notes in Computational Science and Engineering*, pages 44–67. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-73749-0. doi:10.1007/978-3-540-73750-6_2. URL http://dx.doi.org/10.1007/978-3-540-73750-6_2.
- Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10, 1966.
- Benjamin Schuler. Single-molecule fret of protein structure and dynamics - a primer. *Journal of Nanobiotechnology*, 11(1):S2, Dec 2013. ISSN 1477-3155. doi:10.1186/1477-3155-11-S1-S2. URL <https://doi.org/10.1186/1477-3155-11-S1-S2>.
- Alexander W. Schüttelkopf and Daan M. F. van Aalten. PRODRG: A tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica Section D*, 60(8):1355–1363, Aug 2004. doi:10.1107/S0907444904011679. URL <http://dx.doi.org/10.1107/S0907444904011679>.
- Christopher Schölzel. Nonlinear measures for dynamical systems, May 2020. URL <https://doi.org/10.5281/zenodo.3814723>.
- Wharton Research Data Services. Executed trades data for vanguard s&p 500 etf (voo). Dataset retrieved from Wharton Research Data Services (WRDS), January 2024. All executed sales transactions during trading hours in January 2024.
- Patrick Shaffer, Omar Valsson, and Michele Parrinello. Enhanced, targeted sampling of high-dimensional free-energy landscapes using variationally enhanced sampling, with an application to chignolin. *Proceedings of the National Academy of Sciences of the United States of America*, 113(5):1150–1155, 2016. doi:10.1073/pnas.1519712113. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1519712113>.
- Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proc. IEEE*, 104(1):148–175, 2015. doi:10.1109/JPROC.2015.2494218.

- Yibing Shan, John L. Klepeis, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Gaussian split ewald: A fast ewald mesh method for molecular simulation. The Journal of Chemical Physics, 122(5):054101, 2005. doi:10.1063/1.1839571.
- C. E. Shannon. A mathematical theory of communication. The Bell System Technical Journal, 27(3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.
- David E. Shaw, Ron O. Dror, John K. Salmon, J. P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Martin M. Deneroff, Brannon Batson, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Douglas J. Ierardi, John L. Klepeis, Jeffrey S. Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Stefano Piana, Yibing Shan, and Brian Towles. Millisecond-scale molecular dynamics simulations on anton. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC '09, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605587448. doi:10.1145/1654059.1654099. URL <https://doi.org/10.1145/1654059.1654099>.
- David E. Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, Michael P. Eastwood, Joseph A. Bank, John M. Jumper, John K. Salmon, Yibing Shan, and Willy Wriggers. Atomic-level characterization of the structural dynamics of proteins. Science, 330(6002):341–346, 2010. ISSN 0036-8075. doi:10.1126/science.1187409.
- K. Shmilovich, S.S. Panda, A. Stouffer, J. D. Tovar, and A. L. Ferguson. Hybrid computational-experimental data-driven design of self-assembling π -conjugated peptides. Digital Discovery, 1: 448–462, 2022a.
- Kirill Shmilovich, Rachael A Mansbach, Hythem Sidky, Olivia E Dunne, Sayak Subhra Panda, John D Tovar, and Andrew L Ferguson. Discovery of self-assembling π -conjugated peptides by active learning-directed coarse-grained molecular simulation. J. Phys. Chem. B, 124(19): 3873–3891, 2020.
- Kirill Shmilovich, Yifan Yao, John D Tovar, Howard E Katz, André Schleife, and Andrew L Ferguson. Computational discovery of high charge mobility self-assembling π -conjugated peptides. Mol. Syst. Des. Eng., 7(5):447–459, 2022b.
- Hythem Sidky, Wei Chen, and Andrew L. Ferguson. Molecular latent space simulators. Chemical Science, 11:9459–9467, 2020a. doi:10.1039/D0SC03635H. URL <http://dx.doi.org/10.1039/D0SC03635H>.
- Hythem Sidky, Wei Chen, and Andrew L. Ferguson. Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. Molecular Physics, 118(5):e1737742, 2020b. doi:10.1080/00268976.2020.1737742.
- Wonjin Sim, Sol Choi, Gyojin Choo, Mihee Yang, Ju-Hyun Park, and Jeong-Eun Oh. Organophosphate flame retardants and perfluoroalkyl substances in drinking water treatment plants from

- korea: Occurrence and human exposure. Int. J. Environ. Res. Public Health, 18(5):2645, 2021. ISSN 1660-4601. doi:10.3390/ijerph18052645.
- Amit Singer. A remark on global positioning from local distances. Proceedings of the National Academy of Sciences, 105(28):9507–9511, 2008. ISSN 0027-8424. doi:10.1073/pnas.0709842104. URL <https://www.pnas.org/content/105/28/9507>.
- Florian Sittel, Abhinav Jain, and Gerhard Stock. Principal component analysis of molecular dynamics: On the use of cartesian vs. internal coordinates. The Journal of Chemical Physics, 141(1):014111, 2014. doi:10.1063/1.4885338. URL <https://doi.org/10.1063/1.4885338>.
- Benjamin E Soday, Mikko Haataja, and Ioannis G Kevrekidis. Coarse-graining the dynamics of a driven interface in the presence of mobile impurities: Effective description via diffusion maps. Physical Review E, 80(3):031102, 2009.
- Alan W. Sousa da Silva and Wim F. Vranken. Acypype - antechamber python parser interface. BMC Res. Notes, 5(1):367, Jul 2012. ISSN 1756-0500. doi:10.1186/1756-0500-5-367. URL <https://doi.org/10.1186/1756-0500-5-367>.
- R. W. Southern. The Making of the Middle Ages. Faber and Faber, London, 1970.
- Michael Spivak. Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus. W. A. Benjamin, New York, 1965. ISBN 9780805390216.
- J Stark, David S Broomhead, ME Davies, and J Huke. Delay embeddings for forced systems. II. Stochastic forcing. Journal of Nonlinear Science, 13(6):519–577, 2003.
- Jaroslav Stark. Delay embeddings for forced systems. I. Deterministic forcing. Journal of Nonlinear Science, 9(3):255–332, 1999.
- author Strogatz, Steven. Nonlinear dynamics and chaos : with applications to physics, biology, chemistry, and engineering. Westview Press, 2015. URL <https://search.library.wisc.edu/catalog/9910223127702121>.
- Lubert Stryer and Richard P Haugland. Energy transfer: A spectroscopic ruler. Proceedings of the National Academy of Sciences of the United States of America, 58(2):719–726, 1967.
- Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Robust kernel regression for restoration and reconstruction of images from sparse noisy data. In 2006 International Conference on Image Processing, pages 1257–1260, 2006. doi:10.1109/ICIP.2006.312573.
- Floris Takens. Detecting strange attractors in turbulence. Dynamical Systems and Turbulence, 898: 366–381, 1981. URL <http://link.springer.com/content/pdf/10.1007/BFb0091924.pdf>.

- Yifeng Tang, Jeremiah Y Kim, Carman KM IP, Azadeh Bahmani, Qing Chen, Matthew G Rosenberger, Aaron P Esser-Kahn, and Andrew L Ferguson. Data-driven discovery of innate immunomodulators via a closed-loop system combining machine learning and high throughput screening. bioRxiv 2023.06.26.546393, 2023.
- Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319–2323, 2000. doi:10.1126/science.290.5500.2319. URL <https://www.science.org/doi/abs/10.1126/science.290.5500.2319>.
- Richard Chace Tolman. The Principles of Statistical Mechanics. Courier Corporation, 1938.
- Maximilian Topel and Andrew L. Ferguson. Reconstruction of protein structures from single-molecule time series. The Journal of Chemical Physics, 153(19):194102, 11 2020. ISSN 0021-9606. doi:10.1063/5.0024732. URL <https://doi.org/10.1063/5.0024732>.
- Maximilian Topel, Ayesha Ejaz, Allison Squires, and Andrew L. Ferguson. Learned reconstruction of protein folding trajectories from noisy single-molecule time series. Journal of Chemical Theory and Computation, 19(14):4654–4667, 2023. doi:10.1021/acs.jctc.2c00920. URL <https://doi.org/10.1021/acs.jctc.2c00920>. PMID: 36701162.
- Seiji Tsuzuki and Tadafumi Uchamaru. Magnitude of attraction in cf4-ch4 interactions: Are cf4-ch4 interactions weaker than average of cf4-cf4 and ch4-ch4 interactions? J. Fluorine Chem., 231: 109468, 2020. ISSN 0022-1139. doi:<https://doi.org/10.1016/j.jfluchem.2020.109468>. URL <https://www.sciencedirect.com/science/article/pii/S0022113920300191>.
- Francesca Tunioli, Tainah D. Marforio, Laura Favaretto, Sebastiano Mantovani, Angela Pintus, Antonio Bianchi, Alessandro Kovtun, Marco Agnes, Vincenzo Palermo, Matteo Calvaresi, Maria Luisa Navacchia, and Manuela Melucci. Chemical tailoring of β -cyclodextrin-graphene oxide for enhanced per- and polyfluoroalkyl substances (pfas) adsorption from drinking water. Chemistry – A European Journal, 29(60):e202301854, 2023. doi:<https://doi.org/10.1002/chem.202301854>. URL <https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/chem.202301854>.
- United States Environmental Protection Agency. Working list of pfas chemicals with research interest and ongoing work by epa. <https://www.epa.gov/chemical-research/working-list-pfas-chemicals-research-interest-and-ongoing-work-epa>, 2023. (accessed August 2023).
- U.S. Environmental Protection Agency. Aquatic life criteria - perfluorooctane sulfonate (pfos). <https://www.epa.gov/wqc/aquatic-life-criteria-perfluorooctane-sulfonate-pfos>, 2022a. (accessed September 2022).
- U.S. Environmental Protection Agency. Method 533: Determination of per- and polyfluoroalkyl substances in drinking water by isotope dilution anion exchange solid phase extraction and liquid

chromatography/tandem mass spectrometry. <https://www.epa.gov/dwanalyticalmethods/method-533-determination-and-polyfluoroalkyl-substances-drinking-water-isotope>, 2022b. (accessed September 2022).

U.S. Environmental Protection Agency. Method 537.1 determination of selected per- and polyfluorinated alkyl substances in drinking water by solid phase extraction and liquid chromatography/tandem mass spectrometry (lc/ms/ms). https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=348508&Lab=CESER&simpleSearch=0&showCriteria=2&searchAll=537.1&TIMSType=&dateBeginPublishedPresented=03%2F24%2F2018, 2022c. (accessed September 2022).

U.S. Environmental Protection Agency. Pfas analytical methods development and sampling research. <https://www.epa.gov/water-research/pfas-analytical-methods-development-and-sampling-research>, 2022d. (accessed September 2022).

U.S. Environmental Protection Agency. TSCA section 8(a)(7) reporting and recordkeeping requirements for perfluoroalkyl and polyfluoroalkyl substances. <https://www.federalregister.gov/documents/2021/06/28/2021-13180/tsc-section-8a7-reporting-and-recordkeeping-requirements-for-perfluoroalkyl-and-polyfluoroalkyl>, 2022e. (accessed September 2022).

U.S. Environmental Protection Agency. Proposed pfas national primary drinking water regulation. <https://www.epa.gov/sdwa/and-polyfluoroalkyl-substances-pfas>, 2023. (accessed March 2023).

David van der Spoel, 2017. URL <https://www.virtualchemistry.org/ff.php#refs>.

David van der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16): 1701–1718, 2005. ISSN 1096-987X. doi:10.1002/jcc.20291. URL <http://dx.doi.org/10.1002/jcc.20291>.

Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2): 3–28, 2014.

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. URL <https://link.aps.org/doi/10.1103/PhysRev.159.98>.

Thierry Vialar. *Complex and Chaotic Nonlinear Dynamics: Advances in Economics and Finance, Mathematics and Statistics*. Springer, 2009. ISBN 978-3-540-85977-2. URL <http://www.mathworks.com/support/books/book49040.html?category=4&language=1>.

Vincenzo Villani and José M. Zaldivar Comenges. Analysis of biomolecular chaos in aqueous solution. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica*

- Chimica Acta*), 104(3-4):290–295, July 2000. ISSN 1432-881X. doi:10.1007/s002140000121. URL <http://link.springer.com/10.1007/s002140000121>.
- David J. Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press, 2003.
- TJ Wallington, MP Sulbaek Andersen, and OJ Nielsen. The case for a more precise definition of regulated pfas. *Environ. Sci.: Processes Impacts*, 23(12):1834–1838, 2021.
- Peter Walters. *An Introduction to Ergodic Theory*. Springer-Verlag, 1982.
- Peter Walters. A dynamical proof of the multiplicative ergodic theorem. *Transactions of the American Mathematical Society*, 335(1):245–257, 1993. ISSN 00029947. URL <http://www.jstor.org/stable/2154267>.
- J. Wang and A.L. Ferguson. Nonlinear reconstruction of single-molecule free energy surfaces from univariate time series. *Phys. Rev. E.*, 93:032412, 2016.
- J. Wang and A.L. Ferguson. Recovery of protein folding funnels from single-molecule time series by delay embeddings and manifold learning. *J. Phys. Chem. B*, 122 50:11931–11952, 2018a.
- Jiang Wang and Andrew L. Ferguson. Nonlinear machine learning in simulations of soft and biological materials. *Molecular Simulation*, 44(13-14):1090–1107, 2018b. doi:10.1080/08927022.2017.1400164.
- Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E. Charron, Gianni de Fabritiis, Frank Noé, and Cecilia Clementi. Machine learning of coarse-grained molecular dynamics force fields. *ACS Central Science*, 5(5):755–767, 2019. doi:10.1021/acscentsci.8b00913. URL <https://doi.org/10.1021/acscentsci.8b00913>.
- Jianzhong Wang. *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer, 2011.
- Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004. doi:<https://doi.org/10.1002/jcc.20035>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20035>.
- Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.*, 25(2):247–260, 2006.
- Wei Wang, Haipei Shao, Shuangxi Zhou, Donghai Zhu, Xiangzhe Jiang, Gang Yu, and Shubo Deng. Rapid removal of perfluoroalkanesulfonates from water by β -cyclodextrin covalent organic frameworks. *ACS Applied Materials & Interfaces*, 13(41):48700–48708, 2021a.

- Yuin Wang, Seth B Darling, and Junhong Chen. Selectivity of per-and polyfluoroalkyl substance sensors and sorbents in water. ACS Appl. Mater. Interfaces, 13(51):60789–60814, 2021b.
- Peter J. Wangersky. Lotka-volterra population models. Annual Review of Ecology and Systematics, 9:189–218, 1978. ISSN 00664162. URL <http://www.jstor.org/stable/2096748>.
- Killian Q. Weinberger and Lawrence K. Saul. Unsupervised learning of image manifolds by semidefinite programming. International Journal of Computer Vision volume, 70:77–90, 2006.
- Mary Jo Weiss-Errico and Kevin E. O’Shea. Detailed nmr investigation of cyclodextrin-perfluorinated surfactant interactions in aqueous media. Journal of Hazardous Materials, 329: 57–65, 2017a. ISSN 0304-3894. doi:<https://doi.org/10.1016/j.jhazmat.2017.01.017>. URL <https://www.sciencedirect.com/science/article/pii/S0304389417300298>.
- Mary Jo Weiss-Errico and Kevin E O’Shea. Detailed nmr investigation of cyclodextrin-perfluorinated surfactant interactions in aqueous media. J. Hazard. Mater., 329:57–65, 2017b.
- Mary Jo Weiss-Errico and Kevin E. O’Shea. Enhanced host–guest complexation of short chain perfluoroalkyl substances with positively charged β -cyclodextrin derivatives. Journal of Inclusion Phenomena and Macrocyclic Chemistry, 95(1):111–117, 2019. doi:10.1007/s10847-019-00930-w. URL <https://doi.org/10.1007/s10847-019-00930-w>.
- R. Weron. Estimating long-range dependence: finite sample properties and confidence intervals. Physica A: Statistical Mechanics and its Applications, 312(1):285–299, 2002. doi:10.1016/S0378-4371(02)00961-5.
- WHO. Acceptability aspects: Taste, odour and appearance. In Guidelines for drinking-water quality: Fourth edition incorporating the first and second addenda. World Health Organization, Genève, Switzerland, 2022.
- Emmerich. Wilhelm, Rubin. Battino, and Robert J. Wilcock. Low-pressure solubility of gases in liquid water. Chem. Rev., 77(2):219–262, 1977. doi:10.1021/cr60306a003. URL <https://doi.org/10.1021/cr60306a003>.
- Amie Wilkinson. What are lyapunov exponents, and why are they interesting?, 2016. URL <https://arxiv.org/abs/1608.02843>.
- Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In T. Leen, T. Dietterich, and V. Tresp, editors, Advances in Neural Information Processing Systems, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf.
- Steffen Wolf, Benjamin Lickert, Simon Bray, and Gerhard Stock. Multisecond ligand dissociation dynamics from atomistic simulations. Nature Communications, 11(1):2918, Jun 2020. ISSN

- 2041-1723. doi:10.1038/s41467-020-16655-1. URL <https://doi.org/10.1038/s41467-020-16655-1>.
- Jirasak Wong-ekkabut and Mikko Karttunen. The good, the bad and the user in soft matter simulations. Biochimica et Biophysica Acta (BBA) - Biomembranes, 1858(10):2529–2538, 2016. doi:<https://doi.org/10.1016/j.bbamem.2016.02.004>.
- P G Wu and L Brand. Resonance Energy Transfer: Methods and Applications. Analytical Biochemistry, 218:1–13, 1994. ISSN 0003-2697. doi:<http://dx.doi.org/10.1006/abio.1994.1134>. URL <http://www.sciencedirect.com/science/article/pii/S0003269784711341>.
- Chunjie Xia, Miriam L. Diamond, Graham F. Peaslee, Hui Peng, Arlene Blum, Zhanyun Wang, Anna Shalin, Heather D. Whitehead, Megan Green, Heather Schwartz-Narbonne, Diwen Yang, and Marta Venier. Per- and polyfluoroalkyl substances in north american school uniforms. Environ. Sci. Technol., 56(19):13845–13857, 2022. doi:10.1021/acs.est.2c02111.
- Leilei Xiao, Yuhan Ling, Alaaeddin Alsbaiee, Chenjun Li, Damian E Helbling, and William R Dichtel. β -cyclodextrin polymer network sequesters perfluorooctanoic acid at environmentally relevant concentrations. J. Am. Chem. Soc., 139(23):7689–7692, 2017.
- Or Yair, Ronen Talmon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Reconstruction of normal forms by learning informed observation geometries from data. Proceedings of the National Academy of Sciences, 114(38):E7865–E7874, 2017. doi:10.1073/pnas.1620045114. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1620045114>.
- Hao Ye, Richard J Beamish, Sarah M Glaser, Sue C H Grant, Chih-Hao Hsieh, Laura J Richards, Jon T Schnute, and George Sugihara. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. Proceedings of the National Academy of Sciences of the United States of America, 112(13):E1569—76, March 2015. ISSN 0027-8424. doi:10.1073/pnas.1417063112. URL <https://europepmc.org/articles/PMC4386326>.
- Sheng Yin, Juan Francisco López, Jonathan J Calvillo Solís, Michael S Wong, and Dino Villagrán. Enhanced adsorption of pfoa with nano mgal2o4@ cnts: Influence of ph and dosage, and environmental conditions. J. Hazard. Mater. Adv., 9:100252, 2023.
- Ying Yin, Anton Arkhipov, and Klaus Schulten. Simulations of membrane tubulation by lattices of amphiphysin n-bar domains. Structure, 17(6):882–892, 2009. ISSN 0969-2126. doi:<https://doi.org/10.1016/j.str.2009.03.016>. URL <https://www.sciencedirect.com/science/article/pii/S0969212609001841>.
- Lai-Sang Young. What are srb measures, and which dynamical systems have them? Journal of Statistical Physics, 108(5-6):733–754, 2002.
- Gül H. Zerze, Robert B. Best, and Jeetain Mittal. Modest influence of FRET chromophores on the properties of unfolded proteins. Biophysical Journal, 107(7):1654–1660, 2014.

doi:10.1016/j.bpj.2014.07.071. URL [http://www.cell.com/biophysj/abstract/S0006-3495\(14\)00847-9](http://www.cell.com/biophysj/abstract/S0006-3495(14)00847-9).

Haiyang Zhang, Tianwei Tan, Wei Feng, and David van der Spoel. Molecular recognition in different environments: β -cyclodextrin dimer formation in organic solvents. The Journal of Physical Chemistry B, 116(42):12684–12693, 2012. doi:10.1021/jp308416p. URL <https://doi.org/10.1021/jp308416p>. PMID: 23025718.

Zhenyue Zhang and Jing Wang. Mlle: Modified locally linear embedding using multiple weights. In Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS’06, pages 1593–1600, Cambridge, MA, USA, 2006. MIT Press.

Mingfei Zhao, Janani Sampath, Sarah Alamdari, Gillian Shen, Chun-Long Chen, Christopher J Mundy, Jim Pfaendtner, and Andrew L Ferguson. Martini-compatible coarse-grained model for the mesoscale simulation of peptoids. J. Phys. Chem. B, 124(36):7745–7764, 2020.

Xi Zhao, Xu-ri Huang, and Chia-chung Sun. Molecular dynamics analysis of the engrailed homeodomain–dna recognition. Journal of Structural Biology, 155(3):426–437, 2006. ISSN 1047-8477. doi:<https://doi.org/10.1016/j.jsb.2006.03.031>. URL <https://www.sciencedirect.com/science/article/pii/S1047847706000827>.

Zhe Zheng, Huijuan Yu, Wen-Chao Geng, Xin-Yue Hu, Yu-Ying Wang, Zhihao Li, Yuefei Wang, and Dong-Sheng Guo. Guanidinocalix [5] arene for sensitive fluorescence detection and magnetic removal of perfluorinated pollutants. Nat. Commun., 10(1):5762, 2019.

Pavel I Zhuravlev, Christopher Kroboth Materese, and Garegin A Papoian. Deconstructing the native state: Energy landscapes, function, and dynamics of globular proteins. Journal of Physical Chemistry B, 113(26):8800–8812, July 2009. ISSN 1520-6106. doi:10.1021/jp810659u. URL <http://www.ncbi.nlm.nih.gov/pubmed/19453123>.

Robert Zwanzig. Nonequilibrium Statistical Mechanics. Oxford University Press, 2001.