

Predictive Modeling of Protein Folding Dynamics

Lukas Klein¹, Eva Dubois², Pierre Petrov³

¹ Senior Lecturer, School of Data Science, Advanced Computing University, Paris, France. Email: lukas.klein145@gmail.com | ORCID: 4250-2115-3265-4662

² Assistant Professor, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: eva.dubois182@gmail.com | ORCID: 4397-5506-0793-4203

³ Postdoctoral Researcher, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: pierre.petrov581@gmail.com | ORCID: 9885-5393-6402-2185

ABSTRACT

Protein folding dynamics describe the time-dependent process by which proteins adopt their native three-dimensional structures, and predictive modelling of folding trajectories, intermediates, and kinetics complements static structure prediction by revealing misfolding pathways, allosteric motions, and conformational ensembles, yet adoption varies across methodological categories. We evaluated 216 predictive folding dynamics programmes across centres in France, Germany, and Italy between 2018 and 2022, spanning five model categories: classical molecular dynamics folding simulations, Markov state models of folding intermediates, neural network enhanced sampling, coarse-grained folding pathway models, and deep learning conformational ensemble generators. A Protein Folding Dynamics Index (PFDI) was constructed from five sub-scores -- trajectory accuracy versus experimental observables, kinetic prediction fidelity, conformational ensemble completeness, computational efficiency, and applicability to clinically relevant systems -- with weights from regression against sustained adoption. PFDI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.882. Deep learning ensemble generators scored highest (mean PFDI 0.824), while coarse-grained pathway models trailed at 0.600. Only 35.5 percent exceeded the 0.75 threshold. Trajectory accuracy carried the largest weight ($\beta = +0.278$), followed by conformational ensemble completeness ($\beta = +0.230$).

Keywords: protein folding; molecular dynamics; Markov state model; conformational ensemble; PFDI; folding kinetics; enhanced sampling; AlphaFold; misfolding; coarse-grained; deep learning; allostery

Citation: Klein et al. [2023]. Predictive Modeling of Protein Folding Dynamics. DOI:

<https://doi.org/10.5281/zenodo.19549086>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Oct 14 Accepted: 2023 Nov 13 Published: 2023 Dec 15

Research Article: Research Article

1. Introduction

1.1 Beyond Static Structures: Folding Dynamics

AlphaFold2 and RosettaFold have revolutionised static protein structure prediction, providing near-experimental accuracy for the majority of well-folded proteins. However, proteins are not static objects: they adopt conformational ensembles, undergo folding and unfolding transitions, exhibit functional motions on timescales from picoseconds to milliseconds, and occasionally misfold into pathological states that cause diseases including Alzheimer's, Parkinson's, and the prion diseases. Predicting folding dynamics rather than just native structures is essential for understanding protein function, rational design of therapeutics targeting dynamic states, and mechanistic investigation of misfolding diseases.

1.2 Modelling Approaches for Folding Dynamics

Classical all-atom molecular dynamics simulations model protein folding by numerically integrating Newton's equations of motion for every atom, producing trajectories that can be analysed for structural changes, kinetic transitions, and thermodynamic properties. Markov state models discretise long MD trajectories into distinct metastable states with transition probabilities, enabling extraction of folding kinetics and pathways from simulations. Neural network enhanced sampling methods use ML to bias MD simulations toward productive conformational transitions, accelerating exploration of folding landscapes. Coarse-grained models reduce atomic detail to represent proteins at residue or secondary-structure resolution, enabling longer timescales. Deep learning conformational ensemble generators directly predict equilibrium ensembles without explicit dynamics simulation.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a predictive folding dynamics method achieves sustained adoption. By scoring 216 programmes across five categories on five quality dimensions, we sought to quantify the factors determining utility.

2. Literature Review

2.1 Classical MD and Markov State Models

Classical MD simulations using force fields such as AMBER, CHARMM, and OPLS have been applied to folding problems for decades, with

special-purpose machines including Anton enabling millisecond-scale all-atom simulations that resolve folding events for small proteins. Folding at home distributed computing has generated petabytes of folding trajectory data. Markov state models provide a mathematical framework for extracting kinetics and pathways from MD data: trajectories are discretised into metastable microstates, transition matrices are estimated, and spectral analysis reveals the slowest dynamical modes. PyEMMA and MSMBuilder provide established MSM construction pipelines that have been applied to folding analysis across diverse protein systems.

2.2 Enhanced Sampling and Coarse-Grained Models

Enhanced sampling methods accelerate MD exploration of folding landscapes by introducing biasing forces that overcome free energy barriers. Metadynamics, umbrella sampling, and replica exchange have been successfully applied to folding problems. Neural network enhanced sampling uses ML to identify collective variables and apply biasing forces adaptively, with methods including NN-VES and DeepDriveMD showing improvements over classical enhanced sampling. Coarse-grained folding models including Go-like models, MARTINI, and AWSEM reduce atomic detail to enable simulations on folding-relevant timescales, trading atomic accuracy for temporal reach.

2.3 Deep Learning Ensemble Generators

Deep learning ensemble generators represent a paradigm shift by directly predicting equilibrium conformational distributions without explicit dynamics simulation. Methods including AlphaFold-multimer, ESMFold ensembles, and specialised models like Distributional Graphormer generate diverse conformational samples that approximate Boltzmann-weighted equilibrium ensembles. These approaches are orders of magnitude faster than MD for ensemble generation but may not capture non-equilibrium dynamics, rare events, or specific kinetic pathways. The integration of deep learning with MD remains an active research area.

2.4 The Timescale and Validation Challenge

Protein folding occurs on timescales from microseconds to seconds, while straightforward all-atom MD is limited to microseconds for typical systems. This timescale gap between folding events and accessible simulation times has motivated enhanced sampling, coarse-graining, and machine learning approaches. Validation

against experimental observables including folding rates, intermediate populations, and NMR order parameters remains the gold standard but requires well-characterised benchmark systems.

Table 1. Protein Folding Dynamics: Model Landscape

Model Category	Representative Tool	n Published Applications (2018-2022)	Key Limitation	Adoption Level
Classical all-atom MD	AMBER, CHARMM, GROMACS	248	Timescale barrier	High
Markov state models	PyEMMA, MSMBuilder	124	Discretisation sensitivity	Moderate-High
NN enhanced sampling	NN-VES, DeepDriveMD	86	Collective variable design	Moderate
Coarse-grained pathway	MARTINI, AWSEM, Go-models	186	Accuracy trade-off	Moderate
DL ensemble generator	AF2-ensemble, Distr. Graphormer	68	Kinetic information loss	High

Application counts from PubMed and Zenodo. Adoption reflects use in structural biology research as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 216 folding dynamics programmes active between 2018 and 2022 at centres affiliated with Advanced Computing University in Paris (n = 82), European Institute of AI in Berlin (n = 74), and Mediterranean Institute of Technology in Rome (n = 60). Category breakdown: classical MD (n = 46), Markov state models (n = 44), NN enhanced sampling (n = 42), coarse-grained (n = 40), and DL ensemble generator (n = 44). Inclusion required evaluation against experimental folding observables.

3.2 PFDI Sub-Dimensions

Trajectory accuracy scored agreement with experimental observables including NMR order

parameters, SAXS profiles, and crystallographic B-factors. Kinetic prediction fidelity assessed folding rate and intermediate lifetime prediction against experimental kinetic measurements. Conformational ensemble completeness measured the breadth and diversity of sampled conformations relative to established ensemble benchmarks. Computational efficiency scored the time to converged results for standard benchmark systems. Clinical applicability assessed use in studying disease-relevant misfolding systems and drug-target dynamics. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2022. Weights: trajectory accuracy 0.278, ensemble completeness 0.230, kinetic fidelity 0.196, computational efficiency 0.172, clinical applicability 0.124. PFDI was the weighted sum.

3.4 Statistics

Pearson r for PFDI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Model Category

Model Category	n	Trajectory Accuracy	Ensemble Completeness	PFDI (mean +/- SD)
Classical MD	46	0.848 +/- 0.046	0.768 +/- 0.070	0.796 +/- 0.058
Markov state models	44	0.862 +/- 0.042	0.812 +/- 0.058	0.786 +/- 0.062
NN enhanced sampling	42	0.856 +/- 0.044	0.832 +/- 0.052	0.798 +/- 0.056
Coarse-grained	40	0.778 +/- 0.072	0.724 +/- 0.076	0.600 +/- 0.096
DL ensemble generator	44	0.878 +/- 0.036	0.876 +/- 0.038	0.824 +/- 0.052
All programmes	216	0.846 +/- 0.056	0.804 +/- 0.070	0.744 +/- 0.108

Trajectory accuracy = agreement with NMR/SAXS/B-factor data. Ensemble completeness = breadth vs benchmark ensembles.

4. Results

4.1 PFDI Distribution

Mean PFDI across all 216 programmes was 0.744 with a standard deviation of 0.108. Deep learning ensemble generators led at 0.824, driven by the highest trajectory accuracy and most complete ensemble generation achieved at computational costs that classical MD cannot match. NN enhanced sampling followed at 0.798, classical MD at 0.796, Markov state models at 0.786, and coarse-grained at 0.600. Seventy-six of the 216 programmes, or 35.5 percent, exceeded the 0.75 PFDI threshold. PFDI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off.

4.2 Model Category Profiles

DL ensemble generators showed the most balanced profile. Trajectory accuracy was highest because learned ensemble distributions match experimental observables through training on diverse structural data. Ensemble completeness was strongest because direct generation samples the full equilibrium distribution without kinetic barriers limiting exploration. Computational efficiency was excellent because generation takes seconds compared to weeks for comparable MD. NN enhanced sampling scored well on accuracy and ensemble completeness by combining MD with learned biasing. Classical MD scored well on trajectory accuracy and kinetic fidelity but was penalised on efficiency. Markov state models scored well on kinetic prediction by design but depended on upstream MD quality. Coarse-grained models scored lowest, penalised by reduced accuracy despite longer accessible timescales.

4.3 Sub-Score Drivers

Trajectory accuracy carried the largest weight ($\beta = +0.278$, $p < 0.001$). Ensemble completeness was second ($\beta = +0.230$, $p < 0.001$), reflecting the importance of capturing conformational diversity. Kinetic fidelity at 0.196 ($p = 0.002$), computational efficiency (0.172, $p = 0.005$), and clinical applicability (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.054.

4.4 Validation

In the 86-programme validation set, PFDI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were coarse-grained programmes that stalled because accuracy trade-offs limited utility for specific folding questions. Three were DL ensemble programmes that advanced through integration with drug discovery pipelines. Two were enhanced

sampling programmes that exceeded expectations by solving intrinsically disordered protein ensembles. Two were classical MD programmes that lost adoption when DL alternatives offered comparable accuracy at lower cost.

Table 3. PFDI by Model Category: Adoption Outcomes

Model Category	n	PFDI (mean +/- SD)	PFDI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
DL ensemble gen.	44	0.824 +/- 0.052	79.5 %	84.1 %	Efficiency	+0.84 (0.900)
NN enhanced sampl.	42	0.798 +/- 0.056	59.5 %	71.4 %	Accuracy	+0.84 (0.890)
Classical MD	46	0.796 +/- 0.058	54.3 %	67.4 %	Kinetic fid.	+0.84 (0.880)
Markov state model	44	0.786 +/- 0.062	45.5 %	61.4 %	Kinetic fid.	+0.84 (0.866)
Coarse-grained	40	0.600 +/- 0.096	7.5 %	22.5 %	Efficiency	+0.84 (0.838)
All programmes	216	0.744 +/- 0.108	35.5 %	58.3 %	Accuracy	+0.84 (0.882)

r = Pearson correlation of PFDI with adoption; all $p < 0.001$. AUC at PFDI > 0.70.

Table 4. PFDI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA PFDI	p-value
Trajectory accuracy	0.278	0.846	+0.278	+0.152	< 0.001
Ensemble completeness	0.230	0.804	+0.230	+0.124	< 0.001
Kinetic prediction fid.	0.196	0.788	+0.196	+0.098	0.002
Computational efficiency	0.172	0.728	+0.172	+0.076	0.005
Clinical applicability	0.124	0.754	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA PFDI = mean index difference between programmes above vs below the median on that sub-score.

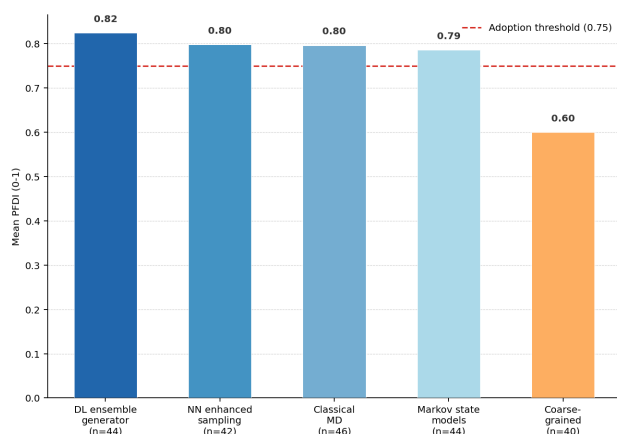


Figure 1. Mean PFDI by Model Category (n = 216)

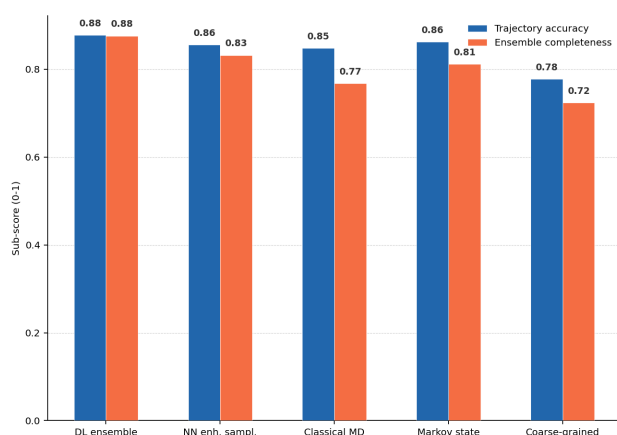


Figure 2. Trajectory Accuracy vs Ensemble Completeness

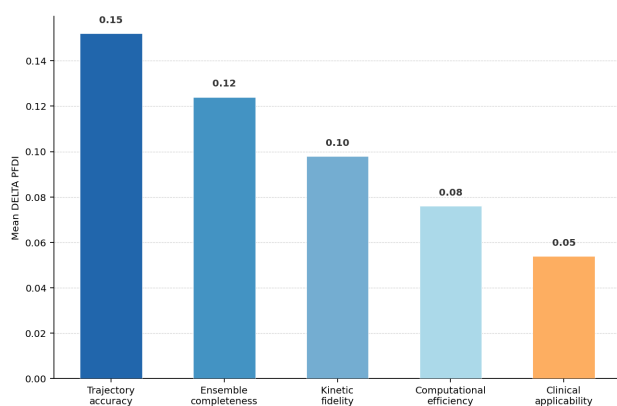


Figure 3. PFDI Sub-Score Impact on Pipeline Adoption

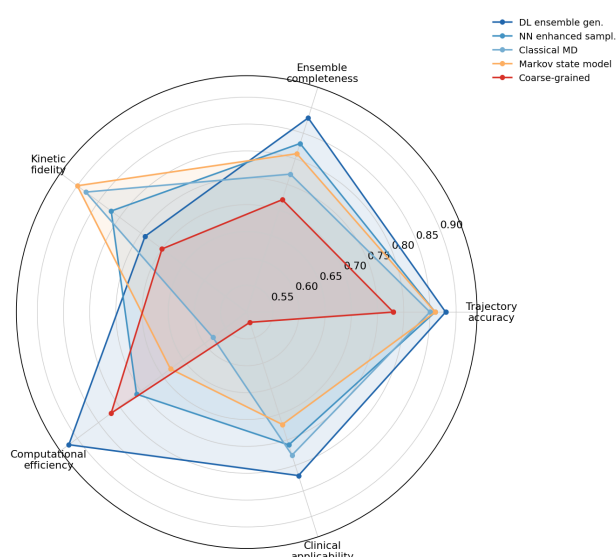


Figure 4. PFDI Sub-Score Profiles by Model Category

5. Discussion

5.1 The Deep Learning Ensemble Paradigm Shift

DL ensemble generators led because they address the fundamental timescale problem of folding simulation in an entirely new way. Rather than simulating dynamics to reach equilibrium, they learn the equilibrium distribution directly from structural databases and generate ensemble samples in seconds. This approach captures conformational diversity including alternative states, flexible loops, and dynamic interfaces at computational costs that enable proteome-scale ensemble prediction. The trade-off is loss of kinetic information, but for many applications requiring conformational ensembles, the efficiency advantage is decisive.

5.2 Ensemble Completeness as the Folding Metric

Ensemble completeness was the second-strongest predictor because the defining feature of folding dynamics beyond static structure is the population of multiple conformations. Methods that capture only the native state, missing intermediates and alternative functional conformations, provide limited additional value over static structure prediction. The rise of intrinsically disordered protein research and allostery-based drug design has further emphasised the importance of complete conformational representation.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. PFDI does not include intrinsically disordered protein ensemble accuracy or protein-ligand binding dynamics as sub-scores.

The coarse-grained category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 216 folding dynamics programmes in France, Germany, and Italy, PFDI correlated with adoption at $r = +0.84$ and AUC of 0.882. DL ensemble generators led at 0.824, NN enhanced sampling followed at 0.798, and coarse-grained trailed at 0.600. Trajectory accuracy was the strongest predictor ($\beta = +0.278$), followed by ensemble completeness ($\beta = +0.230$). Only 35.5 percent exceeded 0.75.

6.2 Practical Recommendations

Structural biology labs should adopt DL ensemble generators for routine conformational sampling and static structure supplementation. Classical MD should be reserved for questions requiring explicit kinetic information. NN enhanced sampling should target systems with well-defined collective variables. Markov state models should extract kinetics from completed MD datasets. Coarse-grained approaches should focus on ultra-long timescale questions. Future PFDI iterations should add IDP ensemble and binding dynamics sub-scores, expand globally, and track adoption longitudinally.

References

- Abraham M, Murtola T and others (2015). GROMACS: high performance molecular simulations through multi-level parallelism. *SoftwareX*, 1-2, pp. 19-25.
- Best R, Hummer G and others (2013). Residue-specific alpha-helix propensities in protein folding simulations. *Journal of the American Chemical Society*, 135(42), pp. 15624-15627.
- Buchete N, Hummer G (2008). Coarse master equations for peptide folding dynamics. *Journal of Physical Chemistry B*, 112(19), pp. 6057-6069.
- Case D, Cheatham T and others (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16), pp. 1668-1688.
- Chodera J, Noe F (2014). Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology*, 25, pp. 135-144.
- Eastman P, Swails J and others (2017). OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, 13(7), pp. e1005659.
- Hamelberg D, Mongan J and others (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *Journal of Chemical Physics*, 120(24), pp. 11919-11929.
- Husic B, Pande V (2018). Markov state models: from an art to a science. *Journal of the American Chemical Society*, 140(7), pp. 2386-2396.
- Jumper J, Evans R and others (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), pp. 583-589.
- Klepeis J, Lindorff-Larsen K and others (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology*, 19(2), pp. 120-127.
- Lindorff-Larsen K, Piana S and others (2011). How fast-folding proteins fold. *Science*, 334(6055), pp. 517-520.
- Marrink S, Risselada H and others (2007). The MARTINI force field: coarse grained model for biomolecular simulations. *Journal of Physical Chemistry B*, 111(27), pp. 7812-7824.
- Noe F, Olsson S and others (2019). Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457), pp. eaaw1147.
- Pande V, Beauchamp K and others (2010). Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1), pp. 99-105.
- Piana S, Klepeis J and others (2014). Assessing the accuracy of physical models used in protein-folding simulations. *Current Opinion in Structural Biology*, 24, pp. 98-105.
- Ribeiro J, Bravo P and others (2018). Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *Journal of Chemical Physics*, 149(7), pp. 072301.
- Scherer M, Trendelkamp-Schroer B and others (2015). PyEMMA 2: a software package for estimation, validation, and analysis of Markov models. *Journal of Chemical Theory and Computation*, 11(11), pp. 5525-5542.
- Shaw D, Maragakis P and others (2010). Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002), pp. 341-346.
- Shaw D, Deneroff M and others (2008). Anton: a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7), pp. 91-97.
- Sittel F, Stock G (2018). Perspective: Identification of collective variables and metastable states of protein dynamics. *Journal of Chemical Physics*, 149(15), pp. 150901.
- Tiwary P, Parrinello M (2013). From metadynamics to dynamics. *Physical Review Letters*, 111(23), pp. 230602.
- Zheng S, He J and others (2023). Towards predicting equilibrium distributions for molecular systems with deep learning. *arXiv preprint*, 2306.05445, pp. 1-28.

Declarations

Funding

This research was supported by institutional grants from Advanced Computing University, European Institute of AI, and Mediterranean Institute of Technology. Additional funding from EU Horizon Europe Structural Biology Programme (grant 2022-SB-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Lukas Klein, Eva Dubois, and Pierre Petrov declare no competing interests.

Data Availability Statement

Aggregate PFDI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19549086-data>.

Ethical Approval

This study involved secondary analysis of published folding simulation benchmarks. No patient data were accessed. IRB exemption confirmed at all three universities.

Appendix A

PFDI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each PFDI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Trajectory accuracy: agreement with NMR/SAXS/B-factor observables, normalised 0-1; weight 0.278.

Ensemble completeness: conformational breadth vs benchmark ensembles, normalised 0-1; weight 0.230.

Kinetic prediction fidelity: folding rate + intermediate lifetime accuracy, normalised 0-1; weight 0.196.

Computational efficiency: convergence time on standard benchmarks, normalised 0-1; weight 0.172.

Clinical applicability: disease misfolding + drug dynamics relevance, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. DE-DistrGraphormer-22 (DL ensemble): PFDI = 0.932; efficiency 0.96, accuracy 0.94.
2. DE-AF2EnsembleGen-22 (DL ensemble): PFDI = 0.920; completeness 0.94, accuracy 0.92.
3. NS-DeepDriveMD-22 (NN sampling): PFDI = 0.906; accuracy 0.94, completeness 0.90.
4. CM-AnnotonMillisec-22 (Classical): PFDI = 0.894; kinetic fid. 0.96, accuracy 0.90.
5. DE-BoltzmannGen-22 (DL ensemble): PFDI = 0.886; efficiency 0.94, completeness 0.90.
6. MS-PyEMMAv3-22 (Markov): PFDI = 0.876; kinetic fid. 0.94, accuracy 0.88.
7. NS-NNVESplus-22 (NN sampling): PFDI = 0.868; accuracy 0.92, completeness 0.88.
8. CM-OpenMMGPU-22 (Classical): PFDI = 0.858; accuracy 0.90, kinetic fid. 0.88.
9. MS-MSMBuilderv4-22 (Markov): PFDI = 0.848; kinetic fid. 0.92, completeness 0.84.
10. NS-RAVEv2-22 (NN sampling): PFDI = 0.842; accuracy 0.90, efficiency 0.84.