Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Machine learning for protein folding and dynamics

Frank Noé[1], Gianni De Fabritiis[2,3] and Cecilia Clementi[4]

Many aspects of the study of protein folding and dynamics have been affected by the recent advances in machine learning. Methods for the prediction of protein structures from their sequences are now heavily based on machine learning tools. The way simulations are performed to explore the energy landscape of protein systems is also changing as force-fields are started to be designed by means of machine learning methods. These methods are also used to extract the essential information from large simulation datasets and to enhance the sampling of rare events such as folding/unfolding transitions. While significant challenges still need to be tackled, we expect these methods to play an important role on the study of protein folding and dynamics in the near future. We discuss here the recent advances on all these fronts and the questions that need to be addressed for machine learning approaches to become mainstream in protein simulation.

Addresses
[1] Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany
[2] Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Doctor Aiguader 88, 08003 Barcelona, Spain
[3] Institucio Catalana de Recerca i Estudis Avanats (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain
[4] Center for Theoretical Biological Physics, and Department of Chemistry, Rice University, 6100 Main Street, Houston, TX 77005, United States

Corresponding author: Clementi, Cecilia (cecilia@rice.edu)

Current Opinion in Structural Biology 2020, **60**:77–84

This review comes from a themed issue on **Folding and binding**

Edited by **Shachi Gosavi** and **Benjamin Schuler**

For a complete overview see the Issue and the Editorial

Available online 24th December 2019

https://doi.org/10.1016/j.sbi.2019.12.005

Introduction

During the last couple of decades advances in artificial intelligence and machine learning have revolutionized many application areas such as image recognition and language translation. The key of this success has been the design of algorithms that can extract complex patterns and highly non-trivial relationships from large amount of data and abstract this information in the evaluation of new data.

In the last few years these tools and ideas have also been applied to, and in some cases revolutionized problems in fundamental sciences, where the discovery of patterns and hidden relationships can lead to the formulation of new general principles. In the case of protein folding and dynamics, machine learning has been used for multiple purposes [1,2,3•, 4,5••,6].

As protein sequences contain all the necessary information to reach the folded structure, it is natural to ask if the ideas and algorithms that have proved very useful to associate labels to images can also help to associate a folded structure to a protein sequence. Indeed, protein structure prediction has greatly benefitted from the influx of idea from machine learning, as it has been demonstrated in the CASP competitions in the last few years, where several groups have used machine learning approaches of different kinds [1,2,7••,3•], and the AlphaFold team from DeepMind won the 2018 competition by a margin [8,9].

In addition to protein structure prediction, machine learning methods can help address other questions regarding protein dynamics. Physics-based approaches to protein folding usually involve the design of an energy function that guides the dynamics of the protein on its conformational landscape from the unfolded to the folded state. Different ideas have been used in the past several decades to design such energy functions, from first-principle atomistic force field [10,11] to simplified coarse-grained effective potential energies [12] encoding physical principles such as for instance the energy landscape theory of protein folding [13,14]. In this context, neural networks can help design these energy functions to take into account of multibody terms that are not easily modeled analytically [5••].

Another aspect where machine learning has made a significant impact is on the analysis of protein simulations. Even if we had an accurate protein force-field and we could simulate the dynamics of a protein long enough to sample its equilibrium distribution, there is still the problem of extracting the essential information from the simulation, and to relate it to experimental measurements. In this case, unsupervised learning methods can help to extract metastable states from high dimensional simulation data and to connect them to measurable observables [15].

In the following we review the recent contributions of machine learning in the advancement of these different aspects of the study of protein folding and dynamics. As the field is rapidly evolving, most probably these contributions will become even more significant in the near future.

Machine learning for protein structure prediction

Structure prediction consists in the inference of the folded structure of a protein from the sequence information. The most recent successes of machine learning for protein structure prediction arise with the application of deep learning to evolutionary information [16,17]. It has long been known that the mutation of one amino acid in a protein usually requires the mutation of a contacting amino acid in order to preserve the functional structure [18–21] and that the co-evolution of mutations contains information on amino acid distances in the three dimensional structure of the protein. Initial methods [16,22] to extract this information from co-evolution data were based on standard machine learning approaches but later methods based on deep residual networks have shown to perform better in inferring possible contact maps [1,2]. More recently, it has been shown that it is possible to predict distance matrices [4] from co-evolutionary information instead of just contact maps. This result was accomplished by using a probabilistic neural network to predict inter-residue distance distributions. From a complete distance matrix, it is relatively straightforward to obtain a protein structure, but of course the prediction of the distance matrix from co-evolution data is not perfect, nor complete. Yet, in [7••] it was shown that, if at least 32–64 sequences are available for a protein family, then this data are sufficient to obtain the fold class for 614 protein families with currently unknown structures, when the co-evolutionary information is integrated in the Rosetta structure prediction approach. Admittedly, the authors concede that this is not yet equivalent to obtain the crystal structure to the accuracy that would be useful, for instance, for drug discovery. However, it still represents a major achievement in structure prediction.

Every two years, the performance of the different methods for structure prediction is assessed in the CASP (Critical Assessment of Techniques for Protein Structure Prediction) competition, where a set of sequences with structures yet to be released are given to participants to predict the structure blindly. The extent of the impact of machine learning in structure prediction has been quite visible in the latest CASP competitions. The typical methodology in previous CASP editions for the top ranked predictions has been to use very complex workflows based on protein threading and some method for structure optimization like Rosetta [23]. Protein threading consists in selecting parts of the sequence for which there are good templates in the PDB and stitch them together [24]. A force-field can then be used to relax this object into a protein structure. The introduction of co-evolution information in the form of contact maps prediction provided a boost in the performance, at the expense of even more complex workflows.

Historically the difference between top predictors in CASP has been minimal — indicating that there was not a clearly better method, but rather an incremental improvement of the workflows. This situation created a barrier of entry to a certain extent for new ideas and models. However, in the latest edition of CASP (CASP13), the group of AlphaFold [9] ranked first with a very simplified workflow [8], heavily based on machine learning methods. The approach extended the contact and distance matrix predictions to predict histograms of distances between amino acids using a very deep residual network on co-evolutionary data. This approach allowed to take into account implicitly the possible errors and inaccuracy in the prediction itself. In addition, it used an autoencoder architecture derived from previous work on drawing [25] to replace threading all-together and generate the structure directly from the sequence and distance histograms. The use of an autoencoder guarantees an implicit, but much more elegant threading of the available structural information in the PDB to the predicted structure. In a second approach from the same group, a knowledge-based potential derived from the distance histograms was also used. The potential was simply minimized to converged structures. This last protein-specific potential minimization might look surprising at first, but it is actually very similar to well-known structure-based models for protein folding [26,13].

An alternative and interesting machine learning approach for structure predictions, which also offers wider applicability, is to use end-to-end differentiable models [27•,3•,28]. While the performance of these methods does not yet reach the performance of co-evolution based methods for cases where co-evolutionary information is high, they can be applied to protein design, and in cases where co-evolution data is missing. In [27•], a single end-to-end network is proposed that is composed by multiple transformations from the sequence to the protein backbone angles and finally to three-dimensional coordinates on which a loss function is computed in terms of root mean square deviations against known structures. In [3•] a sequence-conditioned energy function is parameterized by a deep neural network and Langevin dynamics is used to generate samples from the distribution. In [28] a generative adversarial model is used to produce realistic C_α distance matrices on blocks up to 128-residues, then standard methods are used to recreate the backbone and side chain structure from there. Incidentally, a variational autoencoder was also tested as a baseline with comparable results. This model is not conditioned on sequence, so it is useful for generating new structures and for in-painting missing parts in a crystal structure.

Folding proteins with machine learned force fields

State-of-the-art force fields can reproduce with reasonable accuracy the thermodynamical and structural

properties of globular proteins [10] or intrinsically disordered proteins (IDPs) [11]. Generally, force fields are designed by first assigning a functional form for all the different types of interactions (e.g. electrostatic, Van der Waals, etc.) between the atoms of different types, then optimizing the parameters in these interactions to reproduce as best as possible some reference data.

In the last few years, a new approach on the design of force-fields has emerged, that takes advantage of machine learning tools [29,30]. The idea is to use either a deep neural network or some other machine learning model to represent the classical energy function of a system as a function of the atomic coordinates, instead of specifying a functional form a priori [31•]. The model can then be trained on the available data to 'learn' to reproduce some desired properties, such as energies and forces as obtained from quantum mechanical calculations. As a neural network is a universal function approximator, this approach has the significant advantage that can approximate a large number of possible functional forms for the energy, instead of being constrained by a predefined one, and can in principle include multi-body correlations that are generally ignored in classical force-fields. The downside of this increased flexibility however resides in the fact that a very large amount of data is needed to train the machine learning model as the model may extrapolate poorly in regions of the conformational space where data are not available. So far, large amount of quantum chemical calculations have been used to train such force-fields, but in principle experimental data could also be included [32].

The machine learning approach to force field design has evolved rapidly in the last decade, but it has so far mostly been tested on small organic molecules. Some of the proposed methods are tailored to reproduce the thermodynamics of specific molecules (e.g. [33]), while others attempt to design transferable force-fields that are trained on a large number of small molecules and could in principle be used to simulate a much larger molecule such as a protein (e.g. [34•,35]). Indeed, quantum mechanical calculations on water, amino acids, and small peptides have been included in the latest generation of machine-learned classical force-fields (e.g. the development version of the ANI potential [36]). We are aware of one instance where a machine-learned force-field has been used to simulate a 50 ns molecular dynamics trajectory of a cellulose-binding domain protein (1EXG) in its folded state. Recently, a transferable machine-learned force-field has been tested on polypeptides. However, machine-learned force-fields have not (yet) been used for protein folding simulations, nor have they been used to predict thermodynamic or kinetic properties. While we believe that this will be possible and machine-learned force-fields will be widely used in protein simulations in the near future, at the moment there are still some significant challenges that need to be overcome towards this goal [6].

One fundamental challenge resides on the modeling of long-range interactions. If only quantum calculations on small molecules are used in the training of force-fields, interactions on scales larger than these molecules could easily be missed in the training. The locality of the machine-learned force-fields could be insufficient to capture electrostatic interactions, or long-range van der Waals interactions [37]. This problem could be addressed by separating the long-range effects in the force-field. For instance, atomic partial charges could be learned [38] simultaneously to local energy terms and used in electrostatic interactions that could be added to the machine-learned energy part to obtain a total energy that is used in the training.

Another main challenge resides in the software used for the simulations. Calculating energies and forces for a protein configuration by means of a trained neural network is several orders of magnitude faster than obtaining these quantities *ab initio* with quantum mechanical calculations, but it is still slower than with a standard classical force-field. In order to simulate protein folding, molecular dynamics trajectories of at least microseconds are needed and this timescale is not currently accessible with machine-learned force-fields. Research in this area has so far mostly focused on obtaining an accurate representation for the energy and forces for molecules and tests have been performed on small systems, mostly as a proof of concept. As this field mature, we believe that significant efforts will also be made to optimize the software for practical applications and molecular dynamics simulation with machine learned force-fields will become a viable alternative to current approaches. Additionally, the whole arsenal of methods that have been developed to enhance the sampling of protein configurational landscapes with classical force-fields (e.g. [39,40]) can also be used with machine-learned force-fields to reach longer timescales and larger system sizes.

Machine learning of coarse-grained protein folding models

In parallel to efforts for the design of atomistic force-fields, machine learning has also been used to obtain coarser models [42,43,5••], that could be applied to study larger systems and longer timescales with reduced computational resources. Coarse-grained models map groups of atoms in some effective interactive 'beads' and assign an effective energy function between the beads to try to reproduce some properties of a protein system. Different properties could be targeted, and different strategies have been used to design coarse-grained models, either starting from atomistic simulations (bottom-up) (e.g. [44,45]), experimental data (top-down) (e.g. [46]) or enforcing general 'rules' such as the minimal frustration principle for protein folding [13,14]. In principle, the same ideas used in the design of atomistic force-fields from quantum mechanical data can be used to make the next step in resolution and design coarse-grained

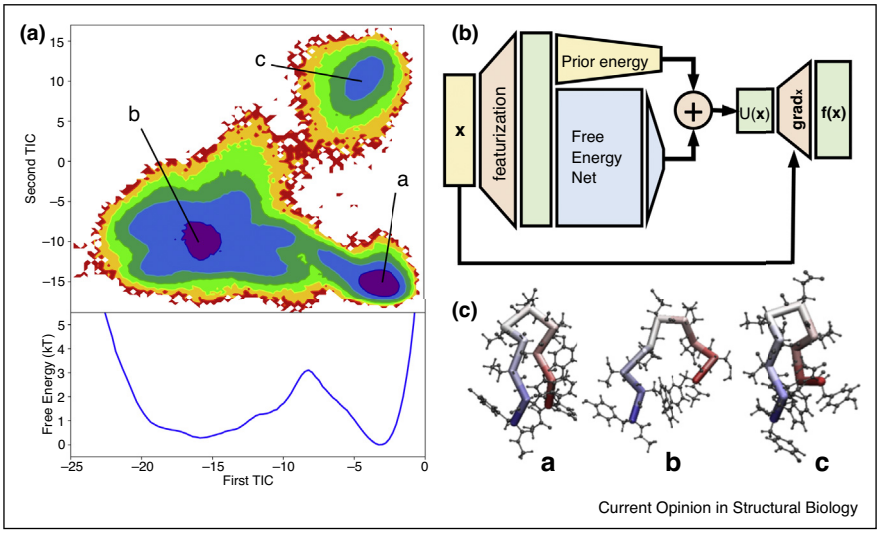
molecular models from all-atom molecular simulations [12]. One main problem in the design of models at a resolution coarser than atomistic is the fact that by renormalizing local degrees of freedom multi-body terms emerge in the effective energy function even if only pairwise interactions were used in a reference atomistic force-field. Such multi-body terms should then be taken into account in the energy function of the coarse-grained model to correctly reproduce the thermodynamics and dynamics of the model at finer resolution. Attempts have been made to include these terms in coarse-grained models, but it is challenging to define suitable and general functional forms to capture these effects in an effective energy function. For this reason, neural networks appear as a natural choice for the design of coarse-grained potentials, as they can automatically capture non-linearities and multi-body terms while agnostic on their specific functional form. Indeed, in the last few years, several groups have attempted to use machine learning methods to design coarse-grained potentials for different systems [42,43,5••]. Most recently, CGnet (see Figure 1), a neural network for coarse-grained molecular force-fields, has been proposed and has been used to model the folding/unfolding dynamics of a small protein [5••]. The CGnet applications presented so far have been system-specific. However similar ideas to what has been used in the design of transferable atomistic force-fields from quantum mechanical data could also been used to try to obtain more transferable coarse-grained models. In general, transferability remains an outstanding issue in the design of coarse models [47] and its requirement may decrease the ability of reproducing faithfully properties of specific systems. So far, the challenges in the definition of general and multi-body functional forms for coarse-grained models have not allowed to rigorously investigate the trade-off between transferability and accuracy for such models. The use of machine learning tools to design effective potential energy functions may soon allow to explore this question systematically.

Machine learning for analysis and enhanced simulation of protein dynamics

Machine learning has been quite impactful in the analysis of simulations of protein dynamics. In this context, two closely related aims are: Firstly, the extraction of collective variables (CVs) associated with the slowest dynamical processes and the metastable states (that can be defined from the knowledge of the slow CVs) from given protein molecular dynamics (MD) simulation data [15]; and finally, enhancing the simulations so as to increase the number of rare event transitions between them.

A cornerstone for the extraction of slow CVs, metastable states and their statistics are shallow machine learning methods such as Markov state models (MSMs) [48] and Master-equation models [49], which model the transitions between metastable states via a Markovian transition or rate matrix. A key advantage of MSMs is that they can be estimated from short MD simulations started from an arbitrary (non-equilibrium) distribution, and yet make predictions of the equilibrium distribution and long-timescale kinetics. While more complex models, for example, including memory, are conceivable, MSMs are simpler to estimate, easier to interpret and are motivated by the observation that if they are built in the slow

Figure 1



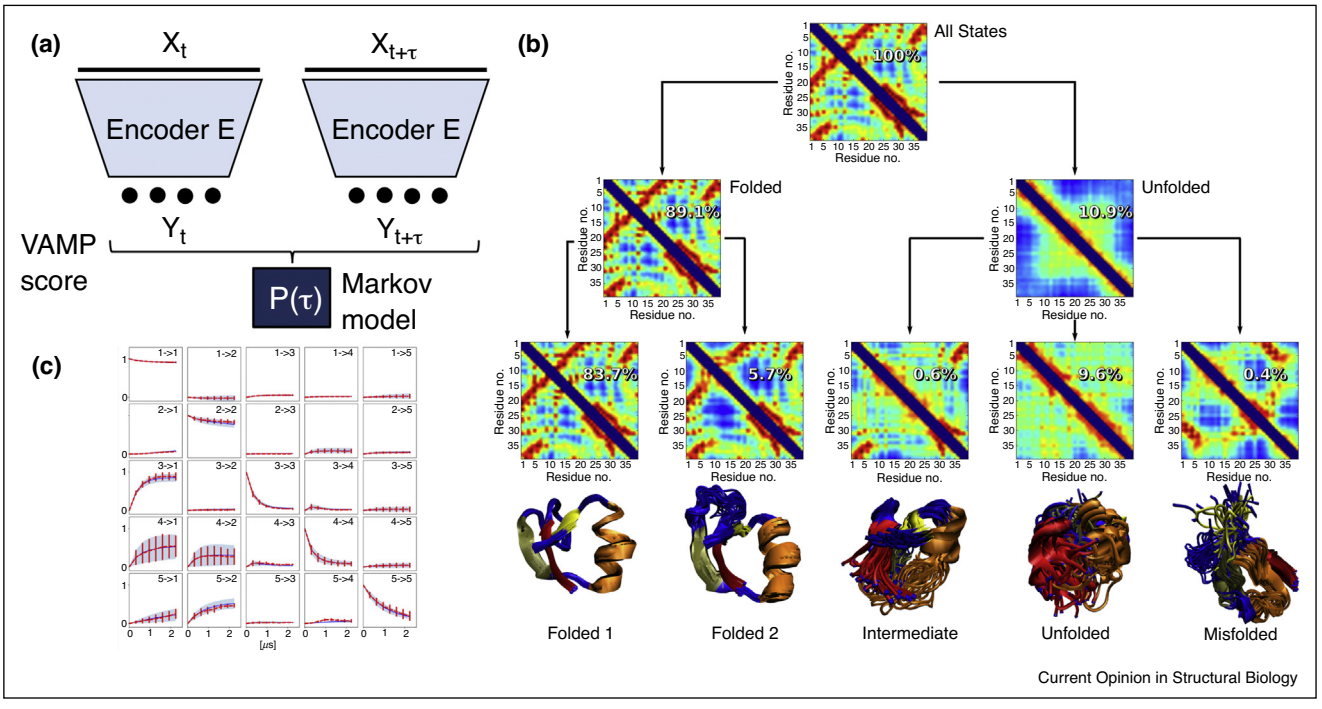
(a) Folding free energy landscape of the protein Chignolin as obtained with a coarse-grained model that uses a neural network to represent the effective energy (CGnet). Top panel: Free energy as obtained from CGnet, as a function of the first two collective coordinates obtained with the Time-Lagged Independent Component Analysis (TICA) method [41]. Bottom panel: Projection of the free energy on the first TICA coordinate. **(b)** The CGnet neural network architecture. **(c)** Representative Chignolin configurations in the three minima from (a). Figure adapted from [5••].

CVs of the molecule, the error made by the Markovian approximation is close to zero for practical purposes [48].

For this reason, much method development has been made in the past 10–15 years in order to optimize the pipeline for the construction of MSMs, that is: finding suitable molecular features to work with [50], reducing the dimensionality of feature space [51,52], clustering the resulting space [53,49], estimating the MSM transition matrix [54] and coarse-graining it [55,56]. While all steps of this pipeline have significantly improved over time, constructing MSMs this way is still very error prone and depends on significant expert knowledge. A critical step forward was the advent of the variational approach of conformation dynamics (VAC) [57] and later the more general variational approach of Markov processes (VAMP) [58]. These principles define loss functions that the best approximation to the slow CVs should minimize, and can thus be used to search over the space of features, discretization and transition matrices variationally [50]. Recently, VAMPnets have been proposed that use neural networks to find the optimal slow CVs and few-state MSM transition matrices by optimizing the VAMP score [59••] (Figure 2a), and hence replace the entire human-built MSM pipeline by a single end-to-end learning framework. VAMPnets have been demonstrated on several benchmark problems including protein folding (Figure 2b) and have been shown to learn high-quality MSMs without significant human intervention (Figure 2c). When used with an output layer that does perform a classification, VAMPnets can be trained to approximate directly the spectral components of the Markov propagator [59••,60].

The aim of enhancing MD sampling is closely connected to identifying the metastable states or slow CVs of a given molecular system. As the most severe sampling problems are due to the rare-event transitions between the most long-lived states, such as folding/unfolding transitions, identifying such states or the corresponding slow CVs on the fly can help to speed up the sampling. So-called adaptive sampling methods perform MD simulation in multiple rounds, and select the starting states for the new round based on a model of the slow CVs or metastable states found so far. Adaptive sampling for protein simulations has been performed using MSMs [61,62] and with neural network approximations of slow CVs [63,64]. Since adaptive sampling uses unbiased (but short) MD

Figure 2



VAMPnet and application to NTL9 protein folding. **(a)** A VAMPnet [59••] includes an encoder E which transforms each molecular configuration $\mathbf{x}_t$ to a latent space of 'slow reaction coordinates' $\mathbf{y}_t$, and is trained on pairs $(\mathbf{y}_t, \mathbf{y}_{t+\tau})$ sampled from the MD simulation using the VAMP score [58]. **(b)** Hierarchical decomposition of the NTL9 protein state space by a network with two and five output nodes. Mean contact maps are shown for all MD samples grouped by the network, along with the fraction of samples in that group. 3D structures are shown for the five-state decomposition, residues involved in α-helices or β-sheets in the folded state are colored identically across the different states. If the encoder performs a classification, the dynamical propagator $\mathbf{P}(\tau)$ is a Markov state model. **(c)** Chapman–Kolmogorov test comparing long-time predictions of the Koopman model estimated at $\tau = 320$ ns and estimates at longer lag times. Figure modified from [59••].

trajectories it is possible to reconstruct the equilibrium kinetics using MSMs, VAMPnets or similar methods. Recently, adaptive sampling has been used to sample protein-protein association and dissociation reversibly in all-atom resolution, involving equilibrium timescales of hours [65].

An alternative to adaptive sampling is to use enhanced sampling methods that speed up rare event sampling by introducing bias potentials, higher temperatures, etc., such as umbrella sampling, replica-exchange or metadynamics. Since these methods typically work in a space of few collective variables, they are also sensitive to making poor choices of collective variables, which can lead to sampling that is either not enhanced, or even slower than the original dynamics. Machine learning has an important role here as it can help these methods by learning optimal choices of collective variables iteratively during sampling. For example, shallow machine learning methods have been used to adapt the CV space during Metadynamics [66,67], adversarial and deep learning have used to adapt the CV space during variationally enhanced sampling (VES, [68]) [69,70]. A completely different approach to predict equilibrium properties of a protein system is the Boltzmann Generator [71••] that trains a deep generative neural network to directly sample the equilibrium distribution of a many-body system defined by an energy function, without using MD simulation.

Since enhanced sampling changes the thermodynamic state of the simulation, it is suitable for the reconstruction of the equilibrium distribution at a target thermodynamic state by means of reweighting Boltzmann probabilities, but generally loses information about the equilibrium kinetics. Ways to recover the kinetics include: Firstly, extrapolating to the equilibrium kinetics of rare event transitions by exploiting the Arrhenius relation [72]; secondly, learning a model of the full kinetics and thermodynamics by combining probability reweighting and MSM estimators in a multi-ensemble Markov model [73]; or finally, reweighting transition pathways [74]. Machine learning and particularly deep learning has not been used much in these methods, but certainly has potential to improve them.

Conclusions

Machine learning can provide a new set of tools to advance the field of molecular sciences, including protein folding and structure prediction. Nonetheless, physical and chemical knowledge and intuition will remain invaluable in the foreseeable future to design the methods and interpret the results obtained. In particular, machine learning can help us to extract new patterns from the data that are not immediately evident, but in virtually all areas reviewed above, machine learning methods that incorporate the relevant physical symmetries, invariances and conservation laws perform better than black-box methods. Furthermore, a trained scientist is still essential to provide meaning to the patterns and use them to formulate general principles.

Conflict of interest

Nothing declared.

Acknowledgements

We gratefully acknowledge funding from European Research Council (ERC CoG 772230 'ScaleCell' to FN), the Deutsche Forschungsgemeinschaft (CRC1114/A04 and GRK2433 DAEDALUS to FN), the MATH+ Berlin Mathematics research center (AA1-6 and EF1-2 to FN), the Einstein Foundation in Berlin (visiting fellowship to CC), the National Science Foundation (grants CHE-1265929, CHE-1740990, CHE-1900374, and PHY-1427654 to CC), the Welch Foundation (grant C-1570 to CC), MINECO (Unidad de Excelencia María de Maeztu MDM-2014-0370 and BIO2017-82628-P to GDF), FEDER (to GDF), and the European Union's Horizon 2020 research and innovation program under grant agreement no. 675451 (CompBioMed project to GDF).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ma J, Wang S, Wang Z, Xu J: **Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning**. *Bioinformatics* 2015, **31**:3506-3513.

2. Wang S, Sun S, Li Z, Zhang R, Xu J: **Accurate de novo prediction of protein contact map by ultra-deep learning model**. *PLoS Comput Biol* 2017, **13**:1-34.

3. • Ingraham J, Riesselman A, Sander C, Marks D: **Learning protein structure with a differentiable simulator**. *International Conference on Learning Representations* 2019

End-to-end-differentiable model for protein structure prediction solely from amino acid sequence information.

4. Xu J: **Distance-based protein folding powered by deep learning**. *Proc Natl Acad Sci U S A* 2019, **116**:16856-16865.

5. •• Wang J, Olsson S, Wehmeyer C, Pérez A, Charron NE, de Fabritiis G, Noé F, Clementi C: **Machine learning of coarse-grained molecular dynamics force fields**. *ACS Cent Sci* 2019, **5**:755-767

Coarse-grained models are extracted from atomistic simulations by using neural networks to capture multi-body terms in the effective energy function.

6. Noé F, Tkatchenko A, Müller K-R, Clementi C: **Machine learning for molecular simulation**. *Ann Rev Phys Chem* 2020, **71**: arXiv:1911.02792 http://dx.doi.org/10.1146/annurev-physchem-042018-052331 (in press).

7. •• Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D: **Protein structure determination using metagenome sequence data**. *Science* 2017, **355**:294-298

Rosetta structure prediction guided by residue-residue contacts inferred from evolutionary information and metagenome sequence data is used to generate structural models for 614 protein families with unknown structures.

8. Evans R, Jumper J, Kirkpatricks J, Sifre L, Green TFG, Qin C, Zidek A, Nelson A, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Jones DT, Silver D, Kavukcuoglu K, Hassabis D, Senior AW: **De novo structure prediction with deep-learning based scoring**. *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction* 2018.

9. Alphafold: Using AI for Scientific Discovery. https://deepmind.com/blog/alphafold/.

10. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE: **Systematic validation of protein force fields against experimental data**. *PLoS ONE* 2012, **7**:e32131.

11. Robustelli P, Piana S, Shaw DE: **Developing a molecular dynamics force field for both folded and disordered protein states**. *Proc Natl Acad Sci U S A* 2018, **115**:E4758-E4766.

12. Clementi C: **Coarse-grained models of protein folding: toy-models or predictive tools?** *Curr Opin Struct Biol* 2008, **18**:10-15.

13. Clementi C, Nymeyer H, Onuchic JN: **Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? Investigation for small globular proteins**. *J Mol Biol* 2000, **298**:937-953.

14. Davtyan A, Schafer NP, Zheng W, Clementi C, Wolynes PG, Papoian GA: **AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing**. *J Phys Chem B* 2012, **116**:8494-8503.

15. Noé F, Clementi C: **Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods**. *Curr Opin Struct Biol* 2017, **43**:141-147.

16. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C: **Protein 3d structure computed from evolutionary sequence variation**. *PLoS ONE* 2011, **6**:e28766.

17. Ovchinnikov S, Kamisetty H, Baker D: **Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information**. *Elife* 2014, **3**:e02030.

18. Altschuh D, Lesk A, Bloomer A, Klug A: **Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus**. *J Mol Biol* 1987, **193**:693-707.

19. Göbel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins**. *Proteins* 1994, **18**:309-317.

20. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T: **Identification of direct residue contacts in protein-protein interaction by message passing**. *Proc Natl Acad Sci U S A* 2008, **106**:67-72.

21. Szurmant H, Weigt M: **Inter-residue, inter-protein and inter-family coevolution: bridging the scales**. *Curr Opin Struct Biol* 2018, **50**:26-32.

22. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families**. *Proc Natl Acad Sci U S A* 2011, **108**:E1293-E1301.

23. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim B-H, Das R, Grishin NV, Baker D: **Structure prediction for CASP8 with all-atom refinement using Rosetta**. *Proteins* 2009, **77**:89-99.

24. Jones DT, Taylort WR, Thornton JM: **A new approach to protein fold recognition**. *Nature* 1992, **358**:86-89.

25. Gregor K, Danihelka I, Graves A, Rezende DJ, Wierstra D: *Draw: A Recurrent Neural Network for Image Generation*. 2015arXiv:1502.04623.

26. Taketomi H, Ueda Y, Gō N: **Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions**. *Int J Pept Protein Res* 1975, **7**:445-459.

27. • AlQuraishi M: **End-to-end differentiable learning of protein structure**. *Cell Syst* 2019, **8** 292–301.e3
Neural network model for structure prediction without the use of co-evolutionary information.

28. Anand N, Huang P-S: **Generative modeling for protein structures.**. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems, NIPS'18*. USA: Curran Associates Inc.; 2018, 7505-7516.

29. Behler J, Parrinello M: **Generalized neural-network representation of high-dimensional potential-energy surfaces**. *Phys Rev Lett* 2007, **98**:146401.

30. Rupp M, Tkatchenko A, Müller K-R, Lilienfeld OAV: **Fast and accurate modeling of molecular atomization energies with machine learning**. *Phys Rev Lett* 2012, **108**:058301.

31. • Schütt KT, Sauceda HE, Kindermans P-J, Tkatchenko A, Müller K-R: **SchNet — a deep learning architecture for molecules and materials**. *J Chem Phys* 2018, **148**:241722
New neural network architecture based on continuous convolutions to learn transferable force-fields from quantum chemical calculations.

32. Chen J, Chen J, Pinamonti G, Clementi C: **Learning effective molecular models from experimental observables**. *J Chem Theory Comput* 2018, **14**:3849-3858.

33. Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller K-R: **Machine learning of accurate energy-conserving molecular force fields**. *Sci Adv* 2017, **3**:e1603015.

34. • Smith JS, Isayev O, Roitberg AE: **ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost**. *Chem Sci* 2017, **8**:3192-3203
Neural networks are used to design transferable force-fields from large amount of quantum chemical calculations.

35. Smith JS, Nebgen B, Lubbers N, Isayev O, Roitberg AE: **Less is more: sampling chemical space with active learning**. *J Chem Phys* 2018, **148**:241733.

36. Isayev O: https://github.com/isayev/ASE_ANI.

37. Hermann J, DiStasio RA, Tkatchenko A: **First-principles models for van der Waals interactions in molecules and materials: concepts, theory, and applications**. *Chem Rev* 2017, **117**:4714-4758.

38. Nebgen B, Lubbers N, Smith JS, Sifain AE, Lokhov A, Isayev O, Roitberg AE, Barros K, Tretiak S: **Transferable dynamic molecular charge assignment using deep neural networks**. *J Chem Theory Comput* 2018, **14**:4687-4698.

39. Laio A, Parrinello M: **Escaping free energy minima**. *Proc Natl Acad Sci U S A* 2002, **99**:12562-12566.

40. Preto J, Clementi C: **Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics**. *Phys Chem Chem Phys* 2014, **16**:19181-19191.

41. Pérez-Hernández G, Paul F, Giorgino T, De Fabritiis G, Noé F: **Identification of slow molecular order parameters for Markov model construction**. *J Chem Phys* 2013, **139**.

42. John ST, Csányi G: **Many-body coarse-grained interactions using Gaussian approximation potentials**. *J Phys Chem B* 2017, **121**:10934-10949.

43. Zhang L, Han J, Wang H, Car R, Dee WE: *PCG: Constructing Coarse-Grained Models Via Deep Neural Networks*. 2018arXiv:1802.08549.

44. Noid WG, Chu J-W, Ayton GS, Krishna V, Izvekov S, Voth GA, Das A, Andersen HC: **The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models**. *J Chem Phys* 2008, **128**:244114.

45. Shell MS: **The relative entropy is fundamental to multiscale and inverse thermodynamic problems**. *J Phys Chem* 2008, **129**:144108.

46. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink S-J: **The MARTINI coarse-grained force field: extension to proteins**. *J Chem Theory Comput* 2008, **4**:819-834.

47. Noid WG: **Perspective: coarse-grained models for biomolecular systems**. *J Chem Phys* 2013, **139**:090901.

48. Prinz J-H, Wu H, Sarich M, Keller BG, Senne M, Held M, Chodera JD, Schütte C, Noé F: **Markov models of molecular kinetics: generation and validation**. *J Chem Phys* 2011, **134**:174105.

49. Buchete NV, Hummer G: **coarse master equations for peptide folding dynamics**. *J Phys Chem B* 2008, **112**:6057-6069.

50. Scherer MK, Husic BE, Hoffmann M, Paul F, Wu H, Noé F: **Variational selection of features for molecular kinetics**. *J Chem Phys* 2019, **150**:194108.

51. Perez-Hernandez G, Paul F, Giorgino T, De Fabritiis G, Noé F: **Identification of slow molecular order parameters for Markov model construction**. *J Chem Phys* 2013, **139**:015102.

52. Schwantes CR, Pande VS: **Improvements in Markov state model construction reveal many non-native interactions in the folding of ntl9**. *J Chem Theory Comput* 2013, **9**:2000-2009.

53. Husic BE, Pande VS: **Ward clustering improves cross-validated Markov state models of protein folding**. *J Chem Theory Comput* 2017, **13**:963-967.

54. Trendelkamp-Schroer B, Wu H, Paul F, Noé F: **Estimation and uncertainty of reversible Markov models**. *J Chem Phys* 2015, **143**:174101.

55. Deuflhard P, Weber M: **Robust perron cluster analysis in conformation dynamics**. In *Linear Algebra Appl, vol 398C*. Edited by Dellnitz M, Kirkland S, Neumann M, Schütte C. New York: Elsevier; 2005:161-184.

56. Noé F, Wu H, Prinz J-H, Plattner N: **Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules**. *J Chem Phys* 2013, **139**:184114.

57. Nüske F, Keller BG, Pérez-Hernández G, Mey ASJS, Noé F: **Variational approach to molecular kinetics**. *J Chem Theory Comput* 2014, **10**:1739-1752.

58. Wu H, Noé F: *Variational Approach for Learning Markov Processes from Time Series Data*. 2017arXiv:1707.04659.

59. •• Mardt A, Pasquali L, Wu H, Noé F: **Vampnets: deep learning of molecular kinetics**. *Nat Commun* 2018, **9**:5
A deep learning framework for molecular kinetics using neural networks, that extract Markov State Models directly from time series of molecular coordinates.

60. Chen W, Sidky H, Ferguson AL: *Nonlinear Discovery of Slow Molecular Modes Using State-Free Reversible Vampnets*. 2019arXiv:1902.03336.

61. Doerr S, Fabritiis GD: **On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations**. *J Chem Theory Comput* 2014, **10**:2064-2069.

62. Hruska E, Abella JR, Nüske F, Kavraki LE, Clementi C: **Quantitative comparison of adaptive sampling methods for protein dynamics**. *J Chem Phys* 2018, **149**:244119.

63. Chen W, Ferguson AL: **Molecular enhanced sampling with autoencoders: on-the-fly collective variable discovery and accelerated free energy landscape exploration**. *J Comput Chem* 2018, **39**:2079-2102.

64. Ribeiro JML, Bravo P, Wang Y, Tiwary P: **Reweighted autoencoded variational Bayes for enhanced sampling (rave)**. *J Chem Phys* 2018, **149**:072301.

65. Plattner N, Doerr S, Fabritiis GD, Noé F: **Protein–protein association and binding mechanism resolved in atomic detail**. *Nat Chem* 2017, **9**:1005-1011.

66. McCarty J, Parrinello M: **A variational conformational dynamics approach to the selection of collective variables in metadynamics**. *J Chem Phys* 2017, **147**:204109.

67. Sultan MM, Pande VS: **tICA-metadynamics: accelerating metadynamics by using kinetically selected collective variables**. *J Chem Theory Comput* 2017, **13**:2440-2447.

68. Valsson O, Parrinello M: **Variational approach to enhanced sampling and free energy calculations**. *Phys Rev Lett* 2014, **113**:090601.

69. Zhang J, Yang YI, Noé F: **Targeted adversarial learning optimized sampling**. *ChemRxiv* 2019 http://dx.doi.org/10.26434/chemrxiv.7932371.

70. Bonati L, Zhang Y-Y, Parrinello M: *Neural Networks Based Variationally Enhanced Sampling*. 2019arXiv:1904.01305.

71. •• Noé F, Olsson S, Köhler J, Wu H: **Boltzmann generators — sampling equilibrium states of many-body systems with deep learning**. *Science* 2019, **365**:eaaw1147
Deep learning methods for the generation of equilibrium structures from the Boltzmann distribution without the need of molecular dynamics.

72. Tiwary P, Parrinello M: **From metadynamics to dynamics**. *Phys Rev Lett* 2013, **111**:230602.

73. Wu H, Paul F, Wehmeyer C, Noé F: **Multiensemble Markov models of molecular thermodynamics and kinetics**. *Proc Natl Acad Sci U S A* 2016, **113**:E3221-E3230.

74. Donati L, Keller BG: **Girsanov reweighting for metadynamics simulations**. *J Chem Phys* 2018, **149**:072335.