Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Artificial intelligence techniques for integrative structural biology of intrinsically disordered proteins

Arvind Ramanathan[1,2], Heng Ma[1], Akash Parvatikar[3] and S Chakra Chennubhotla[3]

We outline recent developments in artificial intelligence (AI) and machine learning (ML) techniques for integrative structural biology of intrinsically disordered proteins (IDP) ensembles. IDPs challenge the traditional protein structure–function paradigm by adapting their conformations in response to specific binding partners leading them to mediate diverse, and often complex cellular functions such as biological signaling, self-organization and compartmentalization. Obtaining mechanistic insights into their function can therefore be challenging for traditional structural determination techniques. Often, scientists have to rely on piecemeal evidence drawn from diverse experimental techniques to characterize their functional mechanisms. Multiscale simulations can help bridge critical knowledge gaps about IDP structure-function relationships — however, these techniques also face challenges in resolving emergent phenomena within IDP conformational ensembles. We posit that scalable statistical inference techniques can effectively integrate information gleaned from multiple experimental techniques as well as from simulations, thus providing access to atomistic details of these emergent phenomena.

Addresses
[1] Data Science & Learning Division, Argonne National Laboratory, Lemont, IL 60439, United States

[2] Consortium for Advanced Science and Engineering (CASE), University of Chicago, Hyde Park, IL, United States
[3] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260, United States

Corresponding author: Ramanathan, Arvind (ramanathana@anl.gov)
URL: https://ramanathanlab.org (A. Ramanathan).

Current Opinion in Structural Biology 2021, **66**:216–224

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

Edited by **Margaret S Cheung** and **Vic Arcus**

For a complete overview see the Issue and the Editorial

Available online 6th January 2021

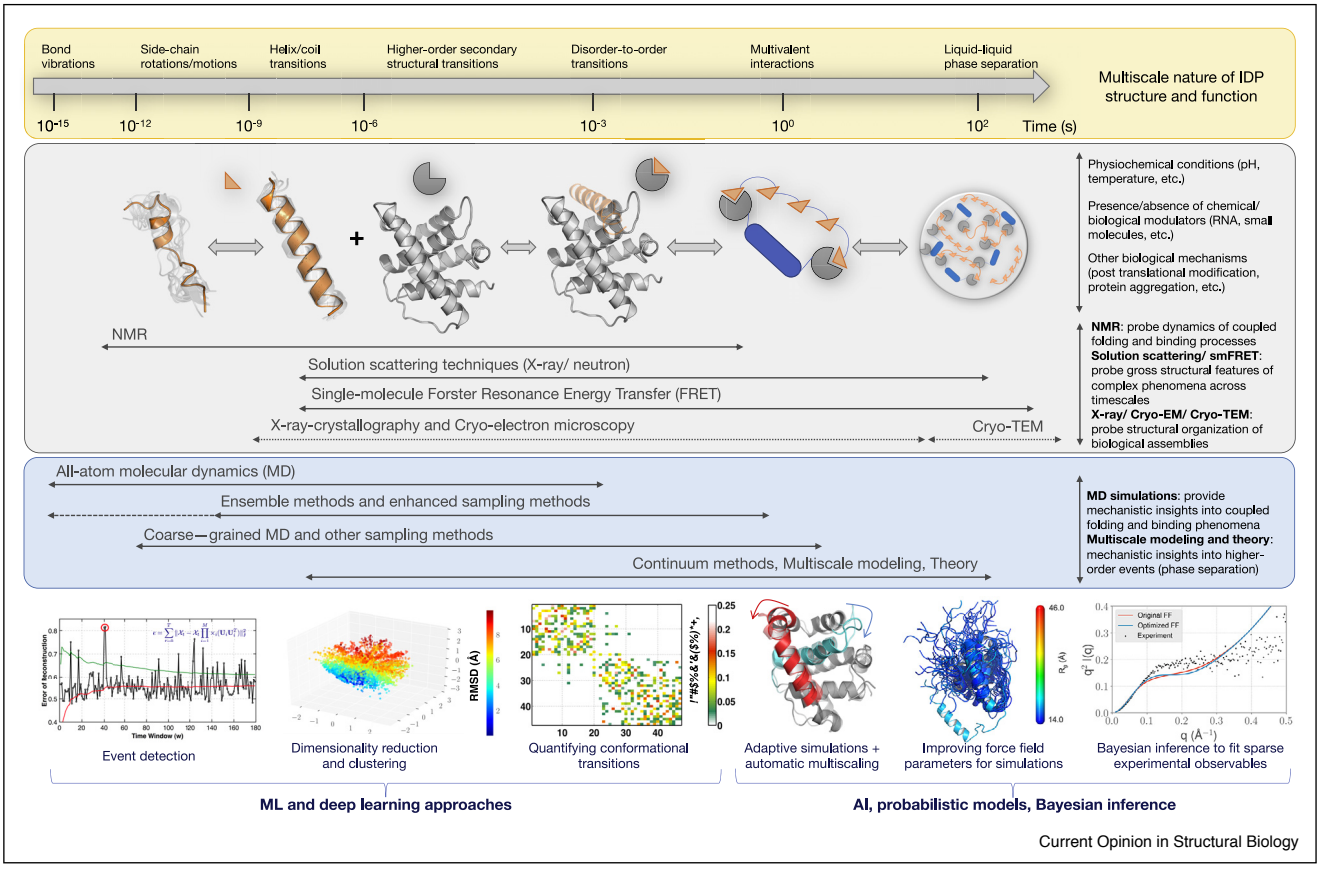
https://doi.org/10.1016/j.sbi.2020.12.001

Introduction

Our current understanding of protein structure-function relationships have been largely driven by the ability to visualize high-resolution three-dimensional (3D) structures of proteins with the aid of structure determination techniques including X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM) [1]. These traditional structure determination techniques have often been supported with evidence from biochemical/biophysical methods to map out the functional consequences of perturbing protein structures through mutations and/or other modifications and for drug-discovery, protein design and other applications. However, the discovery of intrinsically disordered proteins (IDPs), and proteins with intrinsically disordered regions (IDRs) have challenged this traditional structure–function relationship paradigm [2]. In particular, IDPs/IDRs adapt their 3D structures exquisitely in response to their substrates as well as post-translational modifications (such as phosphorylation) and/or based on other physiological conditions (such as pH, crowding, etc.) and can mediate context-specific functions within cells [3]. Indeed, IDPs/IDRs are known to be equally sensitive to perturbations to their primary sequence, where mutations can have devastating effects including misfolding, protein aggregation (e.g. Parkinsons, Alzheimers and other 'conformational diseases') and dysregulation of signaling pathways (e.g. cancer, diabetes, cardiovascular diseases) [4]. Given their central role in mediating complex biological functions within cells, understanding the structure-function paradigm of IDPs/IDRs remains an important challenge for modern biophysics.

The remarkable plasticity of IDPs/IDRs is enabled by their ability to undergo folding upon binding — one of the key mechanistic processes whereby an IDP/IDR adopts distinct secondary or even tertiary structure upon binding to a specific substrate (Figure 1). This coupled folding and binding processes occur at diverse length-scales and time-scales — beginning with finer conformational changes involving partial folding within IDR segments (e.g. helix–coil transitions) to disorder-to-order conformational transitions (e.g. formation of α-helix) upon binding to a particular substrate [5]. These local interactions can then drive the formation of higher-order interactions, whereby repeated 'segments' of hydrophobic/polar amino-acid residues can transiently interact (albeit specifically) to their target substrates. These multivalent interactions in turn lead to coacervation or liquid–liquid phase separation (LLPS), which has important biological implications, including compartmentalization (e.g. membraneless organelles) [6]. One of the key challenges then is to elucidate the mechanisms by which IDPs undergo coupled folding and binding processes leading to such diverse functions.

Figure 1



Role of AI/ML techniques in IDP/IDR biology. Conformational fluctuations within IDPs occur at a wide range of time-scales (top panel) and length-scales (middle panel). Further IDP systems are sensitive to physiological conditions, presence of biological modulators, and other mechanisms such as post-translational modifications. Solution scattering (X-ray/neutron), smFRET and NMR techniques provide access to probe IDP fluctuations over a wide range of length-scales and time-scales; while X-ray and cryo-electron microscopy/ tomography provide access to static snapshots across longer length scales. It is notable that even within cryo-EM and TEM datasets, inherent limitations in resolution can result in a lot of the flexible regions missing, leading to the use of multiscale molecular simulations to fill in the gaps. However, even with improvements in enhanced/adaptive sampling techniques, computational methods and computer hardware, it has been difficult to access details beyond *O*(μm) length-scales and *O*(ms) time-scales. We posit that AI/ML approaches will act as a 'glue' that can enable integrating insights from simulations with experiments while providing a platform to interpret mechanisms of IDP/IDR function.

Although there has been tremendous progress in using traditional structure determination techniques in extending the length-scales and time-scales for studying IDPs [7], these techniques alone cannot fully describe the range of conformational flexibility of IDPs/IDRs. Further, given the intrinsic limitations in the length-scales and time-scales that these techniques can access, often multiple experiments are needed to probe the mechanisms by which IDPs/IDRs function, leading to a piecemeal approach in interpreting IDPs/IDRs ensembles [8]. Molecular dynamics (MD) simulations, either via all-atom simulations or enhanced sampling techniques or multiscale coarse-grained methods provide a much needed 'boost' in terms of sampling IDP conformational landscapes, allowing one to obtain insights into complex phenomena such as LLPS [9]. Synergy between experiments and simulations have been quite successful in quantitatively probing how IDPs/IDRs function; however, such studies find it challenging when different experiments provide seemingly conflicting evidence that are not necessarily explained by simulations [10].

Motivating the need for AI/ML approaches in integrative IDP structural biology. Advances in machine learning (ML) and artificial intelligence (AI) techniques have recently made strides in a number of scientific disciplines including molecular biophysics [11]. We posit that AI/ML techniques can effectively act as a 'glue' to integrate disparate sources of experimental and simulation data and to infer functional mechanisms of IDP/IDRs. In this review, we include a broad definition of how AI/ML methods are applied, where traditional statistical inference methods

can be combined with methods that include neural networks. We examine how AI/ML techniques are being utilized in addressing the aforementioned challenges in IDP integrative structural biology, namely: (1) firstly, characterizing the conformational heterogeneity of IDP ensembles (Section 'AI/ML for characterizing IDP ensembles'); secondly, multiscaling (length-scales and time-scales) IDP ensembles to model emergent phenomena such as LLPS (Section 'AI/ML for multiscale simulations of IDP ensembles'); and finally, integration of sparse experimental observations with simulations to infer mechanisms of IDP function (Section 'Statistical inference for integrating experimental data with simulations'). Our review seeks to complement recent developments in AI/ML applications geared towards protein folding/dynamics [11]. Further, we seek to bridge these advances in the context of simulation techniques for studying emergent behavior [9]. We finally conclude with a perspective on how AI/ML techniques can be integral in elucidating structure-function relationships of IDP/IDRs (Section 'Challenges and outlook').

AI/ML for characterizing IDP ensembles

The range of conformations that IDPs can adapt is primarily attributed to the distribution of amino-acid residues along their primary sequences, where the ratio of charged residues to hydrophobic residues gives rise to specific patterning enabling them to vary their secondary (tertiary, and supra-molecular) structures in solution [12,13]. Since sequence based approaches by themselves are not sufficient to fully characterize IDP conformational landscapes, MD (and/or Monte Carlo) simulations are widely used to probe mechanisms of their functions, typically accessing timescales ranging O(10–100 μs) [14].

Dimensionality reduction methods to organize IDP conformational landscapes. AI/ML methods are necessary to quantify the statistical dependencies in atomistic fluctuations to obtain biophysically-relevant low-dimensional representations spanned by IDP landscapes. Dimensionality reduction methods summarize IDP ensembles in terms of a small number of collective variables or latent dimensions, where projections of the conformations from the simulations capture significant events along these dimensions [15]. These projections are referred to as *embeddings*, where each conformation is represented by the latent dimensions. An implicit requirement of these embedding techniques is that they group conformations in terms of biophysically-relevant observables (e.g. root-mean squared deviations/RMSD, radius of gyration/R_g). Most dimensionality reduction techniques are unsupervised — they exploit the intrinsic statistical structure within the data to discover dependencies without the need for explicit labels (e.g. within an IDP ensemble, there is no explicit notion of what constitutes a folded/partially folded/unfolded state made available to the ML algorithm). Dimensionality reduction techniques can leverage linear, non-linear, or hybrid methods to learn low-dimensional embeddings and here we provide a succinct summary of how they have been used to characterize IDP ensembles [16].

Principal component analysis (PCA) is one such linear embedding method widely popular in analyzing simulation trajectory datasets [15]. However, PCA and its derivative methods lack the ability to characterize conformational diversity purely based on covariance in positional fluctuations alone. One key observation from several MD simulations as well as experimentally determined IDP ensembles is that their positional fluctuations exhibit long-tail distributions — a natural consequence of their ability to undergo large conformational fluctuations and access *rare* states away from their mean positions. These anharmonic fluctuations within IDPs are posited to be functionally relevant, since such fluctuations enable them to access conformational states relevant for binding to their specific substrate. The anharmonicity also gives rise to non-orthogonal correlations between individual atoms/amino-acid residues (depending on the resolution at which the data is being analyzed) [17].

ML techniques such as anharmonic conformational analysis (ANCA) provide a convenient framework to analyze IDP ensembles especially in the context of disorder-to-order transitions [18]. ANCA uses fourth-order statistics to describe the atomic fluctuations and summarizes the internal motions using a small number of dominant anharmonic modes. In a recent study, time-resolved ANCA was used to characterize disorder-to-order transitions in the BCL*2* homology 3 domain, BECN1 (BCL*2*-interacting coiled–coiled protein) as it binds to the murine γ-herpesvirus 68 (M11) B-cell lymphoma *2* (BCL*2*) protein [19•]. This approach identified a small number of conformational states that acted as intermediates in enabling M11-BCL*2* to undergo partial unfolding in response to BECN1 binding. It identified a network of hydrophobic interactions, some farther than 10 Å from the BH3D binding cleft that underwent specific conformational changes upon binding. These interactions were validated using mutagenesis and isothermal calorimetry demonstrating that perturbing the intrinsic anharmonicity within M11 can adversely affect both protein stability and BECN1 binding.

Deep-learning methods in analyzing IDP ensembles. Long-tailed fluctuations in IDP ensembles is a characteristic indicator of multiscale behavior (Figure 1). Further, the linearity assumptions in PCA and ANCA can be limiting in extracting multiscale features from the conformational landscape, especially when such embeddings are non-trivial. Deep learning methods that leverage neural networks have proven to be successful in progressively extracting multiscale features from raw inputs [20].

Deep neural networks such as autoencoders employ an hourglass shaped architecture where data is compressed into a low-dimensional latent space in the early layers and then reconstructed back [21]. The latent space learns to capture most essential information required for accurate reconstruction in the original dimensional space. Variational autoencoders (VAE) is one such instantiation of autoencoders that enforce the latent space to be normally distributed. Several variations of the VAE neural network architecture have been used to characterize latent representations from protein folding trajectories, such as variational dynamics encoder (VDE) [22], variational approaches for Markov processes (VAMP) [23], reweighted autoencoded variational Bayes for enhanced sampling (RAVE) [24], and the convolutional variational autoencoder (CVAE) [25,26]. Although the conceptual use of the VAE is similar, their implementations can vary based on the essential features that they are used to learn. For example, within VDE, the loss function includes a term capturing the slowest processes in the simulation datasets, whereas the EncoderMap [27] utilizes a loss term that captures the proximity of conformations in the free-energy landscape. Complementary to these approaches, recurrent neural networks (RNNs) can serve as effective methods to learn time-dependent embeddings from MD simulations. RNNs, which are used extensively in natural language processing and image processing applications can be used to embed MD simulations to capture Boltzmann statistics from the system but also accurately reproduce the kinetics across multiple timescales [28••]. Another approach by Noe and colleagues used a deep learning approach that is trained on a potential energy function and builds a generative model for conformational ensembles that respects Boltzmann statistics [29••]. The uniqueness of this approach is that it is 'one-shot', meaning that it does not need any reaction coordinates and can produce unbiased samples, circumventing the expensive aspects of MD/Monte-Carlo simulations.

IDRs often function as linkers between several folded domains (in multidomain proteins). This gives rise to an exponential number of states that they can sample, making it further challenging to characterize such complex landscapes. Dynamic graphical models (DGM) propose to address this problem by considering multidomain proteins as assemblies of coupled subsystems where each system is governed by the states it can access as well as the states its neighbors can access [30•]. Although DGMs use fewer parameters than their deep learning counterparts, it is difficult to incorporate prior experimental knowledge and recover atomistic configurations from its encoded representations.

AI/ML for multiscale simulations of IDP ensembles

In the previous section, we described some of the recent developments applying ML approaches to characterize folding conformational landscapes. In this section, we examine how AI/ML methods can firstly, inform efficient sampling of their conformational landscapes and finally, enable multiscale simulations of emergent phenomena such as LLPS.

Determining reaction coordinates and enabling efficient sampling of IDP conformational landscapes. The latent representations learned from MD simulations provide information relevant to reaction coordinates (RCs; also referred to as collective variables, or order parameters) that correspond to conformational changes along biophysically relevant observables (e.g. R_g values, or helicity, etc.). In a recent paper, Romero and colleagues demonstrated that the CVAE-learned embeddings can be used to cluster conformations from long time-scale simulations of the lysosomal enzyzme glucocerebrosidase-1 (GCase) and its facilitator protein saposin C (SAPC) along several reaction coordinates [31]. The proposed conformational changes along the CVAE-determined RCs provided insights into key loop movements at the entrance of the substrate-binding site within GCase that are stabilized by direct interactions with SAPC. Note that this approach only used the raw simulation trajectories of GCase to infer the RCs and did not use any prior information (such as distance between residues or other features within GCase or SAPC). Similar insights can be drawn from other approaches as well [32,33,34•]; however, the consequences of selecting a particular method versus what RCs they extract, and how they represent interpretable (biophysically meaningful) RCs remains an open question.

RCs extracted from the analyses of MD simulations can be used to drive additional sampling of the conformational landscape. This is the basis for many adaptive and enhanced sampling approaches [35]. Techniques such as variational enhanced sampling (VES) [36•], VAMPnets [23], and RAVE [24] already include approaches for enhanced sampling. Both VES and VAMPnets utilize the variational approximation to enhance the sampling based on some set of reaction coordinates that can be determined by analyzing the MD simulations (see Section 'AI/ML for characterizing IDP ensembles'). However, RAVE utilizes the predictive information bottleneck principle as an RC, where it can predict the most likely future trajectory given a molecule's past trajectory. This principle, combined with the estimates for the most informative RCs (automatically determined from the information gain associated with sampling along subsets of RCs), the associated metastable states and equilibrium properties provides simultaneous access to uncover the unbiased kinetics for moving between different metastable states [37].

Generative adversarial networks (GAN) [38] have also been used for enhanced sampling, where on-the-fly training is used to modify the potential energy surface

in order to drive the system to a user-defined target distribution where the free-energy barrier is lowered. This approach, called targeted adversarial learning optimized sampling (TALOS) uses MD simulations (for 'generating' protein conformations) and a discriminator (differentiate samples generated by the biased sampler from those drawn from the desired target distribution) to automatically guide the sampling process [39••]. This approach is inspired from actor-critic reinforcement learning ideas and is complementary to approaches such as reinforcement based adaptive sampling (REAP) [40•].

While AI/ML-driven MD simulations have been demonstrated for smaller peptide/protein systems, there is a need for effective middleware that can orchestrate complex workflows and manage resources efficiently [35]. Conventional (non-deep learning) ML approaches take perhaps between a few seconds to may be a couple of hours to run and can easily be run concurrently with MD simulation jobs as long as the data is made available for analysis. Training deep learning models on the other hand, can potentially take several hours (and even days) similar to the same timeline as MD simulations, which means resource management and scheduling has to be managed to make use of available compute time effectively. To address these issues, DeepDriveMD [41] couples the CVAE [25] with adaptive MD simulations to accelerate folding of small proteins (up to 45 residues) on emerging supercomputers. DeepDriveMD's adaptive protocol could accelerate the sampling by at least 2.3× compared to traditional approaches. The adaptive sampling protocols used within simulation frameworks can be cast more generally as an optimization problem for balancing the cost of exploration (i.e. searching the IDP landscape) versus exploitation (i.e. utilizing existing knowledge to accelerate the search). The AdaptiveBandit [42•] technique uses a reinforcement learning based approach where an action-value function and an upper confidence bound selection algorithm allows for substantial improvement of the sampling strategy.

AI/ML approaches for learning force-field parameters and multiscale approaches. Sampling IDP landscapes implies the need to access a wide range of conformations, even those with relatively low probabilities. While enhanced and adaptive sampling techniques provide an opportunity to access such low-probability conformational states, the timescales that simulations can access is still limited [43]. Another potential challenge that limits the scale of sampling IDP/IDR landscape arises from the force field parameters used for these simulations. Several recent advances in force field parameter development do address these limitations specifically for IDP/IDR systems (see [44–46]); however, artifacts related to how they are parameterized and how they end up capturing interfacial dynamics between IDPs and water (or other solvent conditions) still affect the overall quality of sampling [47,48].

A complementary approach to this strategy is to use AI/ML to iteratively fit and refine force-field parameters in a data-driven fashion. One such approach, called ForceBalance-SAS [49•] (1) uses an initial 'best' set of parameters, (2) computes ensemble averaged small-angle scattering intensities from MD simulations, (3) measures the residual with respect to experimental data, along with the gradient and Hessian of the residual, and (4) optimizes this fitting process from (1–3) iteratively until convergence criteria are achieved. This process continues with the newly updated set of parameters and simulations, completing the cycle. ForceBalance-SAS can optimize parameters for IDPs with varying molecular weight and different charge-hydrophobicity characteristics, albeit in a system-specific manner. While ForceBalance-SAS fits to the global small-angle scattering profiles, the force field parameters also resulted in better agreements with NMR chemical shifts (local observables). Further, the learned parameters could be transferred and applied to other systems partially (for shorter timescale simulations).

For simulating emergent phenomena such as LLPS, coarse-graining is an essential step for making simulations tractable. While there are many approaches to coarse-grain simulations for LLPS (see review by Mittal and colleagues [9]), AI/ML approaches can aid in the development of data-driven representations from all-atom simulations for parameters needed at the coarse-grained resolution. One recent approach called lattice simulation engine for sticker and spacer interactions (LASSI) utilizes Boltzmann inversion, non-linear regression and a Gaussian process Bayesian optimization approach to parameterize the coarse-grained method for modeling sequence-specific phase-behaviors [50]. Additionally, force-field parameters simulating sequence-specific phase behavior could be enabled by an approach such as CAMELOT [51].

Deep learning approaches can also be used to automatically infer coarse-grained representations from all-atom simulations [52,53]. Advances in graph neural networks are aiding the development of accurate coarse-grained force field parameters [54••]. It however remains to be seen how these approaches can be in turn generalized for IDP systems [55]. Similarly, the Multiscale Machine-learned Modeling Infrastructure (MuMMI) [56] was developed to couple a continuum model with coarse-grained MD simulations using ML approaches to characterize how the oncogene RAS interacts with complex biological membranes. Complementary to this approach, adversarial autoencoders were coupled to multi-scale simulations of the severe acute respiratory coronavirus 2 (SARS-CoV-2) Spike protein in complex with the

angiotensin-converting enzyme 2 (ACE2) receptor protein to probe the mechanisms of its infectivity [57••]. Automatic coarse-graining approaches using AI approaches can be really attractive for tuning the scale of coarse-graining that needs to be performed such that IDP/IDR landscapes can be adaptively sampled to obtain precise atomistic scale information about LLPS. Further, the ability to simulate self-consistent ensembles at multiple resolutions (continuum → coarse-grained → all-atom) will be critical for integrative structural biology applications in the context of combining information from diverse experimental techniques (see next section).

Statistical inference for integrating experimental data with simulations

The previous sections outlined the use of AI/ML for characterizing IDP/IDR ensembles. But the true power of obtaining insights into the mechanisms of how IDPs function and how their functions can be exploited for therapeutic design [4], novel material discovery [58], and synthetic biology applications (e.g. membraneless organelles for transport) [1] comes from the integration of theory and simulations with experimental data. The challenge with experimental data, however is that it can be noisy, sparse, and often provide only partial information when investigating a particular phenomenon [10]. For example, solution scattering data for IDPs are usually summarized using the scattering intensities against a coarse structural measure such as R_g [59], and in the case of single molecular Forster resonance energy transfer (sm-FRET) experiments, a set of distances is measured across the IDP structure [60]. Simulations on the other hand, represent a full-scale system with all degrees of freedom (e.g. $3 \times N$, where N represents the individual atoms) implying a mismatch with the intrinsic dimensionality of experimental data. In such cases, how can one fit sparse experimental observables with simulation datasets? A second challenge arises when experimental datasets are unable to resolve flexible regions in a protein (e.g. cryo-electron microscopy) [61]. Given that often such flexible regions hold key insights in terms of understanding ensembles of multi-domain proteins, simulations can fill in the gaps by providing probable states that these regions occupy. But the intrinsic gap in terms of timescales that can be accessed by simulations often ends up making it difficult to extract such information. Thus, AI approaches, augmented with Bayesian approaches can be quite helpful in bridging the gaps between experiments and simulations [10,62].

There are two broad strategies for fitting simulation datasets with experiments. One strategy involves the use of unbiased simulations and then reweighting the generated ensembles using either maximum parsimony/entropy approaches or with Bayesian strategies that uses information known from simulations as a *prior* before the introduction of experimental observables. The complementary strategy involves the use of a biased simulations that are parameterized from experiments or using iterative approaches outlined in [49•] to refine the force field parameters to sample the IDP landscape of interest. Similarly, integrated experimental and computational simulations are also being used to understand energetics of interactions between an IDP and its binding partner [63]. Recent work by Lincoff and colleagues [64] also extends the experimental inferential structure determination using a Bayesian formulation that calculates the maximum log-likelihood of a conformational ensemble by accounting for the uncertainties across a variety of experimental data and back-calculation models. A similar integrated modeling approach by Gomes and colleagues [65••] demonstrated how conformational restraints imposed using NMR, SAXS, and sm-FRET approaches could reach agreement in the ensembles of Sic1 and phosphorylated Sic1.

Challenges and outlook

From quantitatively probing the complex conformational landscapes of IDPs to identifying disorder-to-order transitions or modeling emergent phenomena such as LLPS, AI/ML approaches are proving to be an indispensable tool for both experimental biophysicists as well as modelers. However, AI/ML approaches for MD simulations still face some challenges that need to be addressed.

Current AI/ML applications (barring a few like [18,29••]) tend to use fitting procedures in a blind manner, without much physical bearing, or paying attention to the underlying statistical physics of the system of interest. The resulting fitting procedure can end up overfitting and may not generalize to fully leverage the power of AI/ML in other domains [66,67]. In particular, transferring a AI/ML model learned across simulations can be challenging. AI/ML methods may also get stuck in regimes that are not entirely physical — leading to issues in how appropriate (weighted) sampling can be achieved. Although techniques such as cross-validation and regularization alleviate these problems, there is a need to develop rigorous statistical techniques as well as interactive tools that can assess the performance of AI/ML models. A second challenge arises when force field parameters are designed using AI/ML. Here, the challenge is in maintaining control over the versions of the force-field designed by AI/ML approaches — where initial conditions or datasets used for training, the inherent stochastic nature of how deep learning approaches work, and even program implementations (differences between how TensorFlow and PyTorch modules are implemented) — can result in highly divergent results, even if the physically represented parameters may in fact lie reasonably within the same range. While the AI/ML community already does similar activities through

rigorous benchmarking applications [68], a similar effort from the IDP/IDR community is needed to ensure robustness, reusability, and reproducibility of models across multiple studies. Efforts such as the IDP ensemble [69] database provide for such an opportunity; however, there is a need for the community-wide engagement to assess these intrinsic issues.

Further, many of the AI/ML results are considered *black box*, meaning that it is difficult to reason how the AI/ML model made its inference. Even though there have been some advances in enabling interactivity with the outputs from the AI/ML models [70], there is still the challenge of making it interactive when large datasets are streamed. Developments in interactive data analysis and virtual reality can aid this, although significant developments are needed to make these approaches practical for emerging datasets.

Finally, computational infrastructure to support AI/ML workflows in concert with simulations has been a long-standing challenge [71]. Traditional approaches run MD simulations continuously, store these large datasets and eventually analyze them with AI/ML methods. However, in the Exascale computing era, such approaches will become infeasible as the sheer volume of data generated by these machines can far exceed the capabilities of analyses that needs to be done (and occasionally, computing resources for AI/ML can exceed that of simulations). Approaches such as DeepDriveMD [41] are examining such emerging needs of complex workflows; however, we believe there is much research that needs to be done in order to understand how AI/ML workloads will interact with future simulation workloads. With newer developments in AI techniques, there is an opportunity to accelerate our understanding of how IDPs play a role in disease, developing novel means to design small-molecule inhibitors, designing new bio-materials, and engineering self-assembling systems for synthetic biology applications. We believe these represent exciting opportunities for the future.

Conflict of interest statement

Nothing declared.

Acknowledgements

AR thanks Anda Trifan for assistance in editing and proof-reading the manuscript. This research was supported by Argonne Laboratory Directed Research and Development Computing Expedition project (AR) and NIH/NIGMS GM105978 (SCC).

References and recommended reading

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

- • of special interest
- • of special interest

1. Shin Y, Brangwynne CP: **Liquid phase condensation in cell physiology and disease**. *Science* 2017:357 http://dx.doi.org/10.1126/science.aaf4382.

2. Uversky VN: **Intrinsically disordered proteins and their "mysterious" (meta)physics**. *Front Phys* 2019, **7**:10 http://dx.doi.org/10.3389/fphy.2019.00010.

3. Phillips AH, Kriwacki RW: **Intrinsic protein disorder and protein modifications in the processing of biological signals**. *Curr Opin Struct Biol* 2020, **60**:1-6 http://dx.doi.org/10.1016/j.sbi.2019.09.003.

4. Ruan H, Sun Q, Zhang W, Liu Y, Lai L: **Targeting intrinsically disordered proteins at the edge of chaos**. *Drug Discov Today* 2019, **24**:217-227 http://dx.doi.org/10.1016/j.drudis.2018.09.017.

5. Majumdar A, Dogra P, Maity S, Mukhopadhyay S: **Liquid–liquid phase separation is driven by large-scale conformational unwinding and fluctuations of intrinsically disordered protein molecules**. *J Phys Chem Lett* 2019, **10**:3929-3936 http://dx.doi.org/10.1021/acs.jpclett.9b01731.

6. Schuler B, Borgia A, Borgia MB, Heidarsson PO, Holmstrom ED, Nettels D, Sottini A: **Binding without folding — the biomolecular function of disordered polyelectrolyte complexes**. *Curr Opin Struct Biol* 2020, **60**:66-76 http://dx.doi.org/10.1016/j.sbi.2019.12.006.

7. Xie M, Yu L, Bruschweiler-Li L, Xiang X, Hansen AL, Brüschweiler R: **Functional protein dynamics on uncharted time scales detected by nanoparticle-assisted nmr spin relaxation**. *Sci Adv* 2019:5 http://dx.doi.org/10.1126/sciadv.aax5560.

8. Rout MP, Sali A: **Principles for integrative structural biology studies**. *Cell* 2019, **177**:1384-1403 http://dx.doi.org/10.1016/j.cell.2019.05.016.

9. Dignon GL, Zheng W, Mittal J: **Simulation methods for liquid–liquid phase separation of disordered proteins**. *Curr Opin Chem Eng* 2019, **23**:92-98 http://dx.doi.org/10.1016/j.coche.2019.03.004.

10. Orioli S, Larsen AH, Bottaro S, Lindorff-Larsen K: **How to learn from inconsistencies: integrating molecular simulations with experimental data**. In *Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly*. Edited by Strodel B, Barz B. Academic Press; 2020:123-176 http://dx.doi.org/10.1016/bs.pmbts.2019.12.006. vol 170 of Progress in Molecular Biology and Translational Science (Chapter 3).

11. Noé F, De Fabritiis G, Clementi C: **Machine learning for protein folding and dynamics**. *Curr Opin Struct Biol* 2020, **60**:77-84 http://dx.doi.org/10.1016/j.sbi.2019.12.005.

12. Miskei M, Horvath A, Vendruscolo M, Fuxreiter M: **Sequence-based prediction of fuzzy protein interactions**. *J Mol Biol* 2020, **432**:2289-2303 http://dx.doi.org/10.1016/j.jmb.2020.02.017.

13. Horvath A, Miskei M, Ambrus V, Vendruscolo M, Fuxreiter M: **Sequence-based prediction of protein binding mode landscapes**. *PLOS Comput Biol* 2020, **16**:1-19 http://dx.doi.org/10.1371/journal.pcbi.1007864.

14. Robustelli P, Piana S, Shaw DE: **Mechanism of coupled folding-upon-binding of an intrinsically disordered protein**. *J Am Chem Soc* 2020, **142**:11092-11101 http://dx.doi.org/10.1021/jacs.0c03217.

15. Tribello GA, Gasparotto P: **Using dimensionality reduction to analyze protein trajectories**. *Front Mol Biosci* 2019, **6**:46 http://dx.doi.org/10.3389/fmolb.2019.00046.

16. Ceriotti M: **Unsupervised machine learning in atomistic simulations, between predictions and understanding**. *J Chem Phys* 2019, **150**:150901 http://dx.doi.org/10.1063/1.5091842.

17. Burger VM, Ramanathan A, Savol AJ, Stanley CB, Agarwal PK, Chennubhotla CS: **Quasi-anharmonic analysis reveals intermediate states in the nuclear co-activator receptor binding domain ensemble**. *Biocomputing 2012. Pacific Symposium on Biocomputing* 2011:1 http://dx.doi.org/10.1142/9789814366496_0008.

18. Parvatikar A, Vacaliuc GS, Ramanathan A, Chennubhotla SC: **Anca: anharmonic conformational analysis of biomolecular simulations**. *Biophys J* 2018, **114**:2040-2043 http://dx.doi.org/10.1016/j.bpj.2018.03.021.

19. Ramanathan A, Parvatikar A, Chennubhotla SC, Mei Y, Sinha SC:
• **Transient unfolding and long-range interactions in viral bcl2 m11 enable binding to the becn1 bh3 domain**. *Biomolecules* 2020:10 http://dx.doi.org/10.3390/biom10091308
The authors show how higher-order statistical approaches can capture disorder-to-order transitions in the Beclin 1 BH3 domain as it binds to and interacts with the murine herpesvirus BCL2 protein.

20. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**:436-444 http://dx.doi.org/10.1038/nature14539.

21. Doersch C: *Tutorial on Variational Autoencoders*. 2016arXiv:1606.05908.

22. Hernández CX, Wayment-Steele HK, Sultan MM, Husic BE, Pande VS: **Variational encoding of complex dynamics**. *Phys Rev E* 2018, **97**:062412 http://dx.doi.org/10.1103/PhysRevE.97.062412.

23. Mardt A, Pasquali L, Wu H, Noé F: **Vampnets for deep learning of molecular kinetics**. *Nat Commun* 2018, **9**:5 http://dx.doi.org/10.1038/s41467-017-02388-1.

24. Ribeiro JML, Bravo P, Wang Y, Tiwary P: **Reweighted autoencoded variational Bayes for enhanced sampling (rave)**. *J Chem Phys* 2018, **149**:072301 http://dx.doi.org/10.1063/1.5025487.

25. Bhowmik D, Gao S, Young MT, Ramanathan A: **Deep clustering of protein folding simulations**. *BMC Bioinformatics* 2018, **19**:484 http://dx.doi.org/10.1186/s12859-018-2507-5.

26. Varolgüneş YB, Bereau T, Rudzinski JF: **Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders**. *Mach Learn Sci Technol* 2020, **1**:015012 http://dx.doi.org/10.1088/2632-2153/ab80b7.

27. Lemke T, Peter C: **Encodermap: dimensionality reduction and generation of molecule conformations**. *J Chem Theory Comput* 2019, **15**:1209-1215 http://dx.doi.org/10.1021/acs.jctc.8b00975.

28. Tsai ST, Kuo EJ, Tiwary P: *Learning Molecular Dynamics With*
•• *Simple Language Model Built Upon Long Short-Term Memory Neural Network*. 2020arXiv:2004.12360
This approach builds an embedding approach with long short-term memory networks for characterizing conformational dynamics for biological macromolecules. The simple models are able to learn non-trivial connectivity between metastable states.

29. Noé F, Olsson S, Köhler J, Wu H: **Boltzmann generators:**
•• **sampling equilibrium states of many-body systems with deep learning**. *Science* 2019:365 http://dx.doi.org/10.1126/science.aaw1147
This paper provides a 'one-shot' learning approach to learn protein conformational landscapes using a supervised training procedure using neural networks and invertible transformations between coordinates of the system of interest and Gaussian coordinates (of same dimensions).

30. Olsson S, Noé F: **Dynamic graphical models of molecular**
• **kinetics**. *Proc Natl Acad Sci U S A* 2019, **116**:15001-15006 http://dx.doi.org/10.1073/pnas.1901692116
The authors address a specific aspect of understanding relative movements between molecular 'subsystems' by using probabilistic graphical models.

31. Romero R, Ramanathan A, Yuen T, Bhowmik D, Mathew M, Munshi LB, Javaid S, Bloch M, Lizneva D, Rahimova A, Khan A, Taneja C, Kim SM, Sun L, New MI, Haider S, Zaidi M: **Mechanism of glucocerebrosidase activation and dysfunction in Gaucher disease unraveled by molecular dynamics and deep learning**. *Proc Natl Acad Sci U S A* 2019, **116**:5086-5095 http://dx.doi.org/10.1073/pnas.1818411116.

32. Rydzewski J, Valsson O: *Multiscale Reweighted Stochastic Embedding (mrse): Deep Learning of Collective Variables for Enhanced Sampling*. 2020arXiv:2007.06377.

33. Smith Z, Ravindra P, Wang Y, Cooley R, Tiwary P: **Discovering protein conformational flexibility through artificial-intelligence-aided molecular dynamics**. *J Phys Chem B* 2020, **124**:8221-8229 http://dx.doi.org/10.1021/acs.jpcb.0c03985.

34. Fakharzadeh A, Moradi M: **Effective Riemannian diffusion**
• **model for conformational dynamics of biomolecular systems**. *J Phys Chem Lett* 2016, **7**:4980-4987 http://dx.doi.org/10.1021/acs.jpclett.6b02208
This paper introduces robust framework for conformational free energy calculation methods. This approach can be used in general for computing the potential of mean force and minimum free energy path in an invariant manner for coordinate transformations unlike Euclidean methods.

35. Kasson PM, Jha S: **Adaptive ensemble simulations of biomolecules**. *Curr Opin Struct Biol* 2018, **52**:87-94 http://dx.doi.org/10.1016/j.sbi.2018.09.005.

36. Bonati L, Zhang YY, Parrinello M: **Neural networks-based**
• **variationally enhanced sampling**. *Proc Natl Acad Sci U S A* 2019, **116**:17641-17647 http://dx.doi.org/10.1073/pnas.1907975116
The paper provides a first demonstration of how neural networks could be used to drive atomistic simulations using the variational enhanced sampling principles. The sampling approach is demonstrated on prototypical systems such as alanine di-/tetra-peptide.

37. Lamim Ribeiro JM, Tiwary P: **Toward achieving efficient and accurate ligand-protein unbinding with deep learning and molecular dynamics through rave**. *J Chem Theory Comput* 2019, **15**:708-719 http://dx.doi.org/10.1021/acs.jctc.8b00869.

38. Goodfellow I: *Nips 2016 Tutorial: Generative Adversarial Networks*. 2016arXiv:1701.00160.

39. Zhang J, Yang YI, Noé F: **Targeted adversarial learning**
•• **optimized sampling**. *J Phys Chem Lett* 2019, **10**:5791-5797 http://dx.doi.org/10.1021/acs.jpclett.9b02173
The paper provides a framework for adversarial learning methods for enhanced sampling approaches of protein conformational landscapes. These approaches can be specifically adapted for studying IDP ensembles especially when conformational changes are not directly accessible via traditional approaches.

40. Shamsi Z, Cheng KJ, Shukla D: **Reinforcement learning based**
• **adaptive sampling: reaping rewards by exploring protein conformational landscapes**. *J Phys Chem B* 2018, **122**:8386-8395 http://dx.doi.org/10.1021/acs.jpcb.8b06521
This paper provides a framework for using reinforcement learning approaches with MD simulation methods.

41. Lee H, Turilli M, Jha S, Bhowmik D, Ma H, Ramanathan A: **Deepdrivemd: deep-learning driven adaptive molecular simulations for protein folding**. *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)* 2019:12-19.

42. Pérez A, Herrera-Nieto P, Doerr S, De Fabritiis G: **Adaptivebandit:**
• **a multi-armed bandit framework for adaptive sampling in molecular simulations**. *J Chem Theory Comput* 2020, **16**:4685-4693 http://dx.doi.org/10.1021/acs.jctc.0c00205
The authors demonstrate the use of reinforcement learning approaches for adaptive sampling of protein conformational landscapes.

43. Bhattacharya S, Lin X: **Recent advances in computational protocols addressing intrinsically disordered proteins**. *Biomolecules* 2019:9 http://dx.doi.org/10.3390/biom9040146.

44. Zerze GH, Zheng W, Best RB, Mittal J: **Evolution of all-atom protein force fields to improve local and global properties**. *J Phys Chem Lett* 2019, **10**:2227-2234 http://dx.doi.org/10.1021/acs.jpclett.9b00850.

45. Yang S, Liu H, Zhang Y, Lu H, Chen H: **Residue-specific force field improving the sample of intrinsically disordered proteins and folded proteins**. *J Chem Inform Model* 2019, **59**:4793-4805 http://dx.doi.org/10.1021/acs.jcim.9b00647.

46. Choi JM, Pappu RV: **Experimentally derived and computationally optimized backbone conformational statistics for blocked amino acids**. *J Chem Theory Comput* 2019, **15**:1355-1366 http://dx.doi.org/10.1021/acs.jctc.8b00572.

47. Zapletal V, Mládek A, Melková K, Louša P, Nomilner E, Jasenáková Z, Kubán V, Makovická M, Laníková A, Žídek L, Hritz J: **Choice of force field for proteins containing structured and intrinsically disordered regions**. *Biophys J* 2020, **118**:1621-1633 http://dx.doi.org/10.1016/j.bpj.2020.02.019.

48. Best RB: **Emerging consensus on the collapse of unfolded and intrinsically disordered proteins in water**. *Curr Opin Struct Biol* 2020, **60**:27-38 http://dx.doi.org/10.1016/j.sbi.2019.10.009.

49. Demerdash O, Shrestha UR, Petridis L, Smith JC, Mitchell JC,
• Ramanathan A: **Using small-angle scattering data and parametric machine learning to optimize force field**

parameters for intrinsically disordered proteins. *Front Mol Biosci* 2019, **6**:64 http://dx.doi.org/10.3389/fmolb.2019.00064
This paper provides a description of the ForceBalance-SAS algorithm that uses parametric machine learning method coupled to the ForceBalance approach for tuning force-field parameters. The approach has been demonstrated on three diverse IDPs with different sequence composition profiles.

50. Choi JM, Dar F, Pappu RV: **Lassi: a lattice model for simulating phase transitions of multivalent proteins**. *PLOS Comput Biol* 2019, **15**:1-39 http://dx.doi.org/10.1371/journal.pcbi.1007028.

51. Ruff KM, Harmon TS, Pappu RV: **Camelot: a machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences**. *J Chem Phys* 2015, **143**:243123 http://dx.doi.org/10.1063/1.4935066.

52. Zhang L, Han J, Wang H, Car REW: **Deepcg: constructing coarse-grained models via deep neural networks**. *J Chem Phys* 2018, **149**:034101 http://dx.doi.org/10.1063/1.5027645.

53. Wang Y, Ribeiro JML, Tiwary P: **Past-future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics**. *Nat Commun* 2019, **10**:3573 http://dx.doi.org/10.1038/s41467-019-11405-4.

54. •• Husic BE, Charron NE, Lemm D, Wang J, Pérez A, Krämer A, Chen Y, Olsson S, de Fabritiis G, Noé F, Clementi C: *Coarse Graining Molecular Dynamics With Graph Neural Networks*. 2020arXiv:2007.11412
The authors provide an approach for using graph neural networks to learn coarse-graining of all-atom simulations using prototypical systems and develop a novel embedding method that succeeds at reproducing the thermodynamics for small biomolecular systems.

55. Noé F: *Machine Learning for Molecular Dynamics on Long Timescales*. Cham: Springer International Publishing; 2020, 331-372 http://dx.doi.org/10.1007/978-3-030-40245-7_16.

56. Di Natale F, Bhatia H, Carpenter TS, Neale C, Kokkila-Schumacher S, Oppelstrup T, Stanton L, Zhang X, Sundram S, Scogland TRW, Dharuman G, Surh MP, Yang Y, Misale C, Schneidenbach L, Costa C, Kim C, D'Amora B, Gnanakaran S, Nissley DV, Streitz F, Lightstone FC, Bremer PT, Glosli JN, Ingólfsson HI: **A massively parallel infrastructure for adaptive multiscale simulations: Modeling ras initiation pathway for cancer**. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. New York, NY, USA: Association for Computing Machinery; 2019 http://dx.doi.org/10.1145/3295500.3356197.

57. •• Casalino L, Dommer A, Gaieb Z, Barros EP, Sztain T, Ahn SH, Trifan A, Brace A, Bogetti A, Ma H, Lee H, Turilli M, Khalid S, Chong L, Simmerling C, Hardy DJ, Maia JDC, Phillips JC, Kurth T, Stern A, Huang L, McCalpin J, Tatineni M, Gibbs T, Stone JE, Jha S, Ramanathan A, Amaro RE: **Ai-driven multiscale simulations illuminate mechanisms of sars-cov-2 spike dynamics**. *bioRxiv* 2020 http://dx.doi.org/10.1101/2020.11.19.390187
This paper presents an AI-driven workflow to study the mechanism of infectivity in the SARS-CoV-2 virus using multi-scale simulations coupled through an adversarial autoencoder network. The size and scale of how these simulations also demonstrate how AI-methods coupled to molecular dynamics simulations can provide novel scientific insights into such complex biological phenomena.

58. Dzuricky M, Roberts S, Chilkoti A: **Convergence of artificial protein polymers and intrinsically disordered proteins**. *Biochemistry* 2018, **57**:2405-2414 http://dx.doi.org/10.1021/acs.biochem.8b00056.

59. Lipfert J, Doniach S: **Small-angle x-ray scattering from rna, proteins, and protein complexes**. *Annu Rev Biophys Biomol Struct* 2007, **36**:307-327 http://dx.doi.org/10.1146/annurev.biophys.36.040306.132655.

60. Metskas LA, Rhoades E: **Single-molecule fret of intrinsically disordered proteins**. *Annu Rev Phys Chem* 2020, **71**:391-414 http://dx.doi.org/10.1146/annurev-physchem-012420-104917.

61. Lyumkis D: **Challenges and opportunities in cryo-em single-particle analysis**. *J Biol Chem* 2019, **294**:5181-5197 http://dx.doi.org/10.1074/jbc.REV118.005602.

62. Bottaro S, Lindorff-Larsen K: **Biophysical experiments and biomolecular simulations: a perfect match?** *Science* 2018, **361**:355-360 http://dx.doi.org/10.1126/science.aat4010.

63. Zou J, Simmerling C, Raleigh DP: **Dissecting the energetics of intrinsically disordered proteins via a hybrid experimental and computational approach**. *J Phys Chem B* 2019, **123**:10394-10402 http://dx.doi.org/10.1021/acs.jpcb.9b08323.

64. Lincoff J, Haghighatlari M, Krzeminski M, Teixeira JMC, Gomes GNW, Gradinaru CC, Forman-Kay JD, Head-Gordon T: **Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states**. *Commun Chem* 2020, **3**:74 http://dx.doi.org/10.1038/s42004-020-0323-0.

65. •• Gomes GNW, Krzeminski M, Namini A, Martin EW, Mittag T, Head-Gordon T, Forman-Kay JD, Gradinaru CC: **Conformational ensembles of an intrinsically disordered protein consistent with nmr, saxs, and single-molecule fret**. *J Am Chem Soc* 2020, **142**:15697-15710 http://dx.doi.org/10.1021/jacs.0c02088
Gomes *et al.* demonstrate an integrative approach for structural biology of IDP systems using diverse experimental observations. The method has been developed for Sic1, a prototypical IDP. Although there is no direct use of machine learning methods, the statistical approaches described to iteratively fit the different experimental datasets to make them consistent provides an idea of how AI/ML methods need to deal with diverse constraints into account.

66. Pant S, Smith Z, Wang Y, Tajkhorshid E, Tiwary P: **Confronting pitfalls of ai-augmented molecular dynamics using statistical physics**. *bioRxiv* 2020 http://dx.doi.org/10.1101/2020.06.11.146985.

67. Goolsby C, Moradi M: **Addressing the embeddability problem in transition rate estimation**. *bioRxiv* 2020 http://dx.doi.org/10.1101/707919.

68. Mattson P, Reddi VJ, Cheng C, Coleman C, Diamos G, Kanter D, Micikevicius P, Patterson D, Schmuelling G, Tang H, Wei G, Wu C: **Mlperf: An industry standard benchmark suite for machine learning performance**. *IEEE Micro* 2020, **40**:8-16.

69. Varadi M, Tompa P: *The Protein Ensemble Database*. Cham: Springer International Publishing; 2015, 335-349 http://dx.doi.org/10.1007/978-3-319-20164-1_11.

70. Chae J, Bhowmik D, Ma H, Ramanathan A, Steed C: **Visual analytics for deep embeddings of large scale molecular dynamics simulations**. *2019 IEEE International Conference on Big Data (Big Data)* 2019:1759-1764.

71. Fox G, Glazier JA, Kadupitiya J, Jadhao V, Kim M, Qiu J, Sluka JP, Somogyi E, Marathe M, Adiga A, Chen J, Beckstein O, Jha S: *Learning Everywhere: Pervasive machine Learning for Effective High-Performance Computation*. 2019arXiv:1902.10810.