

# UC Berkeley

## UC Berkeley Previously Published Works

### Title

Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data.

### Permalink

<https://escholarship.org/uc/item/7df6638c>

### Journal

The Journal of Chemical Physics, 158(17)

### Authors

Zhang, Oufan  
Haghighatlari, Mojtaba  
LI, JIE  
[et al.](#)

### Publication Date

2023-05-07

### DOI

10.1063/5.0141474

Peer reviewed

# Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data

Cite as: *J. Chem. Phys.* **158**, 174113 (2023); doi: [10.1063/5.0141474](https://doi.org/10.1063/5.0141474)

Submitted: 5 January 2023 • Accepted: 11 April 2023 •

Published Online: 5 May 2023










View Online



Export Citation



CrossMark

Oufan Zhang,<sup>1</sup>  Mojtaba Haghghatlari,<sup>1,a)</sup>  Jie Li,<sup>1</sup>  Zi Hao Liu,<sup>2,3</sup>  Ashley Namini,<sup>2</sup>   
João M. C. Teixeira,<sup>2,3,b)</sup>  Julie D. Forman-Kay,<sup>2,3</sup>  and Teresa Head-Gordon<sup>1,4,c)</sup> 

## AFFILIATIONS

<sup>1</sup>Kenneth S. Pitzer Theory Center and Department of Chemistry, University of California, Berkeley, California 94720, USA

<sup>2</sup>Molecular Medicine Program, Hospital for Sick Children, Toronto, Ontario M5S 1A8, Canada

<sup>3</sup>Department of Biochemistry, University of Toronto, Toronto, Ontario M5G 1X8, Canada

<sup>4</sup>Department of Bioengineering and Chemical and Biomolecular Engineering, University of California, Berkeley, California 94720, USA

**Note:** This paper is part of the JCP Special Topic on Machine Learning Hits Molecular Simulations.

**a) Current address:** Pfizer, 610 Main St, Cambridge, MA 02139, USA.

**b) Current address:** Department of Biomedical Sciences, University of Padova, Padova, Italy.

**c) Author to whom correspondence should be addressed:** [thg@berkeley.edu](mailto:thg@berkeley.edu)

## ABSTRACT

The structural characterization of proteins with a disorder requires a computational approach backed by experiments to model their diverse and dynamic structural ensembles. The selection of conformational ensembles consistent with solution experiments of disordered proteins highly depends on the initial pool of conformers, with currently available tools limited by conformational sampling. We have developed a Generative Recurrent Neural Network (GRNN) that uses supervised learning to bias the probability distributions of torsions to take advantage of experimental data types such as nuclear magnetic resonance J-couplings, nuclear Overhauser effects, and paramagnetic resonance enhancements. We show that updating the generative model parameters according to the reward feedback on the basis of the agreement between experimental data and probabilistic selection of torsions from learned distributions provides an alternative to existing approaches that simply reweight conformers of a static structural pool for disordered proteins. Instead, the biased GRNN, DynamICE, learns to physically change the conformations of the underlying pool of the disordered protein to those that better agree with experiments.

Published under an exclusive license by AIP Publishing. <https://doi.org/10.1063/5.0141474>

## I. INTRODUCTION

Many proteins adopt a well-defined three-dimensional structure to carry out their function. Despite the widely accepted protein structure–function paradigm, it is increasingly appreciated that all proteomes also encode intrinsically disordered proteins and regions (IDPs/IDRs), which do not adopt a well-defined 3D structure but instead form fluctuating and heterogeneous structural ensembles.<sup>1–4</sup> The structural disorder of IDPs/IDRs is central to their functional roles but is also implicated in diseases, including autism spectrum disorder, cancer, and many others.<sup>5–7</sup> More recently, IDPs have been found to be over-represented in biomolecular condensates<sup>8,9</sup> and have been suggested to promote phase separation due to their structural plasticity, low-complexity sequence

domains, and multivalency.<sup>10–12</sup> Hence, the structural characterization of IDPs/IDRs constitutes a new frontier in structural biology in order to understand their biological function, requiring a computational approach backed by experiments to model their diverse and dynamic structural ensembles.

The conformational heterogeneity of IDPs can be ascertained using various types of biophysical experiments, including Nuclear Magnetic Resonance (NMR), small angle x-ray scattering (SAXS), single molecule fluorescence resonance energy transfer (smFRET), and any other available solution experimental measurements.<sup>4,13</sup> However, because solution experiments only measure ensemble and/or time averages given the dynamic nature of disordered protein states, computational methods must be used to complete the

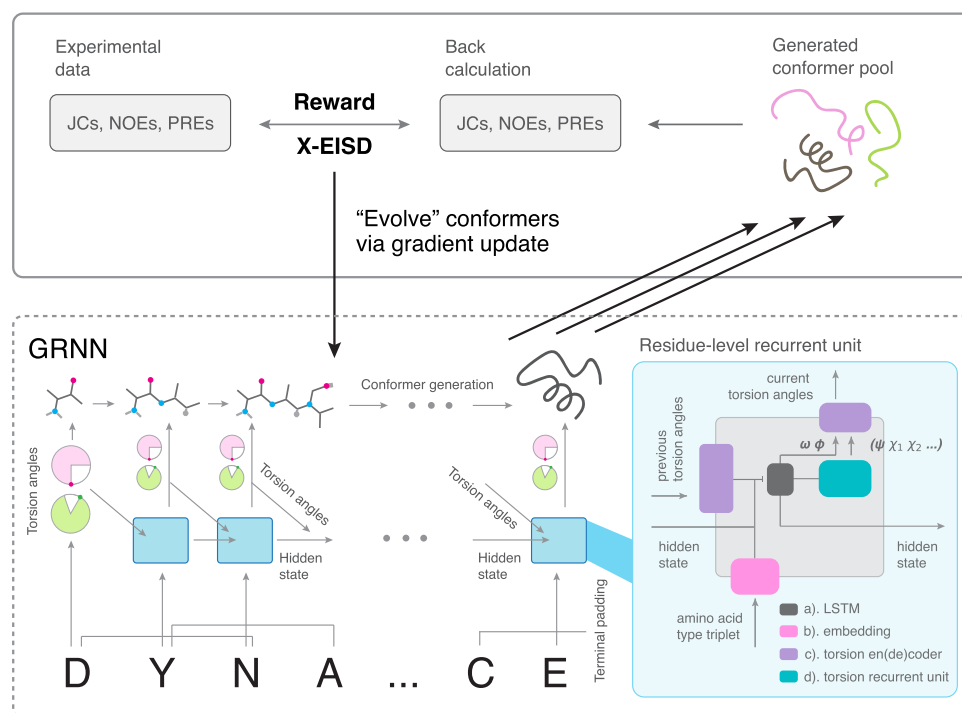
atomic scale structural ensemble. Hence, a number of computational approaches have been developed for generating and evaluating disordered structural ensembles that are consistent with the collective experimental restraints.

The creation of large structural pools of unfolded or IDP conformations can be derived from a variety of sources, such as molecular dynamics (MD) simulations using a force field,<sup>14–16</sup> or structural builders, such as TraDES,<sup>17</sup> Flexible-Meccano,<sup>18</sup> FastFloppyTail,<sup>19</sup> and IDPConformerGenerator.<sup>20</sup> To optimize agreement with experiments, most methods have typically focused on either biasing molecular simulations using experimental data, as in the case of the ensemble-biased metadynamics method,<sup>21</sup> or selecting a collection of structures from a pre-generated pool of candidate conformers that best fit the available experimental data, such as ENSEMBLE,<sup>22–25</sup> Mollack,<sup>26–28</sup> the energy-minima mapping and weighting method,<sup>29,30</sup> and ASTEROIDS.<sup>18,31–34</sup>

In recent years, Bayesian models have emerged as an ideal framework to account for the multiple and different sources of uncertainties in the IDP problem, most typically experimental and back-calculation model errors, as originally proposed by Stultz and co-workers.<sup>26–28</sup> These robust statistical approaches provide a confidence level in the calculated structural ensemble models given their

undetermined nature and variable quality of the restraining experimental solution data.<sup>26–28,35–43</sup> Among some of the most visible developments are maximum parsimony inspired methods exemplified by the Bayesian weighing (BW) method<sup>26</sup> and maximum entropy inspired techniques represented by the Bayesian ensemble refinement method,<sup>35</sup> metainference,<sup>37,42</sup> Bayesian/maximum entropy (BME),<sup>43,44</sup> and the Bayesian inference of ensembles (BioEn) method.<sup>39,41</sup> The Head-Gordon lab developed the Extended Experimental Inferential Structure Determination (X-EISD) method that treats experimental and model errors as Gaussian random variables and can use their joint probabilities in a Monte Carlo sampling or maximization procedure for refining the computational ensembles given experimental data.<sup>36,40</sup>

However, in order for these Bayesian approaches to be successful, the underlying structural pool must cover a representative conformational space such that the most important conformers can be weighted more heavily than more irrelevant conformations for the optimization to be effective. However, the “putative” disordered ensemble may not contain a relevant pool of structures. For example, some structural builder approaches<sup>17–19</sup> can generate structures that are unphysical, with large steric clashes and a lack of Boltzmann weighting. While MD-generated ensembles do



**FIG. 1.** Schematic of the design of the DynamICE GRNN and its interplay with a supervised learning strategy using a Bayesian reward function to evolve new conformer generation to create new IDP ensembles given the data. Top: The generated conformer pools are evaluated by their agreement with the experimental data to formulate a feedback to the GRNN to generate new conformers with better agreement with the data through gradient updates. Bottom: The GRNN generates new torsion angles, which are sequentially translated to Cartesian coordinates to generate new conformers. Each of the residue-level recurrent units (sky blue) takes as input a residue triplet, torsion angles of the previous residue, and the previous hidden state to compute an internal state of a target central residue. The hidden states are vectors that pass information between sequential recurrent units. The internal states inside the recurrent unit are defined by (a) a multi-layer long short-term memory (LSTM) neural network; (b) a dictionary-like embedding layer that encodes each unique amino acid type triplet with a vector; (c) the two-way en(de)coder between a torsion angle and a Gaussian smeared probability vector; and (d) a recurrent unit that handles torsion angle generation within a next residue build step. These components are described in detail in the section titled Method.

contain energetically weighted states, they can have structural biases toward overly compact states using many popular force fields and, thus, may be poor descriptions for disordered protein states.<sup>45</sup> While new IDP-specific force fields have been introduced,<sup>46,47</sup> in some cases, they no longer describe folded states<sup>48</sup> and/or tend to become too unstructured and featureless to be consistent with the solution data<sup>49</sup> and, therefore, can result in underlying biases in the resulting ensemble. Although there are force fields that can better describe both IDPs and folded proteins,<sup>48</sup> MD approaches can also be expensive, requiring sampling on timescales of tens to hundreds of microseconds.

The recent advent of machine learning models, most notably AlphaFold2<sup>50</sup> and RoseTTAFold,<sup>51</sup> has made stunning breakthroughs in producing target structures of monomeric folded proteins of quality similar to experimental structures.<sup>52–58</sup> Other examples are deep convolutional neural networks that predict structures as distance maps<sup>54–57</sup> and natural language processing that encodes protein sequences using recurrent neural networks (RNNs).<sup>58,59</sup> To create a diverse and representative protein structural space, the machine learning field has also seen an emergence of generative neural networks,<sup>60</sup> predominantly employing variational autoencoders (VAEs) and generative adversarial networks (GANs) to learn from native protein databases to propose structural variants of folded states<sup>61,62</sup> or to learn from MD to provide a less computationally expensive alternative for conformational sampling.<sup>63,64</sup>

Advances in structure prediction and generation for folded proteins foreshadow an exciting frontier in applying machine learning methods to the integrative modeling of IDP ensembles.<sup>65,66</sup> Recently, Gupta *et al.* used a VAE to compress MD generated conformers for the disordered  $\alpha\beta40$  and ChiZ proteins to a low-dimensional latent space, which were sampled to reconstruct conformers and subsequently validated in a subsequent and independent step against NMR chemical shifts (CS) and SAXS data.<sup>67</sup> Janson *et al.* also proposed a GAN model called idpGAN that learns from coarse-grained simulations to generate disordered proteins<sup>68</sup> but is only assessed with experimental data at a later stage. Even the highly successful AlphaFold2 for folded proteins predicts regions of disorder with low confidence.<sup>69</sup>

Here, we introduce a machine learning approach that learns the probability of the next residue torsions  $X_{i+1} = [\phi_{i+1}, \psi_{i+1}, \omega_{i+1}, \chi_{i+1}]$  from the previous residue in the sequence  $X_i$  using a generative recurrent neural network (GRNN) model to build new conformational states of a disordered protein ensemble. This work is distinguished further by a supervised learning step that biases the probability distributions of torsions of the GRNN to take advantage of experimental data types such as three-bond J-couplings (JCs), nuclear Overhauser effects (NOEs), and paramagnetic resonance enhancements (PREs) from NMR spectroscopy. The resulting biased-GRNN machine learning model (Fig. 1), which we call DynamICE (dynamic IDP creator with experimental restraints), learns to structurally change the conformations of the underlying pool to those that better agree with solution experiments, using the X-EISD Bayesian model and enforcing realistic energetic states through a Lennard-Jones potential. We show that updating the DynamICE model parameters according to the reward feedback on the basis of the agreement between structures and data provides a conceptual advance over existing approaches that simply reweight

static structural pools for disordered proteins.<sup>35,40,43</sup> The significance of the DynamICE approach is that we evolve the underlying structural ensemble to agree with the experimental data as opposed to iterative guesswork about relevant sub-populations and subsequent reweighting of arbitrary conformations.

We apply our DynamICE machine learning (ML) approach to four protein cases: the unfolded state of the human salivary histatin 5 (Hst5), amyloid- $\beta$  1-40 ( $A\beta40$ ), the *Drosophila* DrkN-terminal SH3 domain (uDrkN-SH3), and the  $\alpha$ -synuclein ( $\alpha$ -Syn) IDP to demonstrate its ability to evolve new conformers driven by better agreement with solution experimental data.

- Hst5 possesses antimicrobial activity in oral fluid; while the Hst5 molecules remain unstructured in aqueous solutions, they prefer to adopt  $\alpha$ -helical conformations in DMSO solvent that mimics the polar aprotic membrane environment.<sup>70</sup>
- $A\beta40$  is one of the important hallmarks of Alzheimer's disease characterized by insoluble fibrils and plaques in the extracellular space within the brain.<sup>71</sup> However, the monomeric form of  $A\beta40$  has been classified as an IDP,<sup>14–16</sup> and understanding its disordered conformational ensembles is relevant to preventing the selected unfolded sub-states from associating with and folding into toxic oligomers or ordered fibril states that can have a great therapeutic value.<sup>72</sup>
- DrkN-SH3, which exists in ~1:1 equilibrium between folded and unfolded states under non-denaturing conditions, is a popular test case with abundant experimental data made available for ensemble reweighting used for disordered proteins.<sup>23</sup> The structural features of the unfolded state, uDrkN-SH3, can help explain the lack of stability of the folded beta-structured domain and are highly valuable for understanding folding thermodynamic equilibrium in general.
- Finally,  $\alpha$ -Syn is an IDP that shows transient sampling of extended and helical conformations in the aqueous phase as opposed to the highly ordered helical state in the membrane. Because the disordered state is the precursor to the toxic fibers found in Parkinson's disease, knowing more about this state will be beneficial for potential therapeutic targeting.<sup>73</sup>

These examples are both biologically and structurally diverse, with important basic research and translational therapeutic implications.

## II. METHOD

In this section, we describe the design of the GRNN and supervised machine learning algorithms. Some more technical details of protein representation, conformer generation, and DynamICE training are further described in [Appendices A and B](#).

### A. Generative recurrent neural network architecture

RNNs are designed to handle sequential information by determining the current outputs from past information along with the current inputs. Previous work by AlQuraishi<sup>58</sup> has developed an end-to-end differentiable model that encodes protein sequences in



the torsional space using RNN to predict novel folds given the primary sequences and mutation information. Similarly, in this work, we use an advanced multi-layer long short-term memory (LSTM) network<sup>74</sup> to iteratively predict the distribution of an accessible angle range of the torsion angles in the current residue given those of the last residue and its associated hidden state to generate protein conformers. LSTMs can preserve long-term memory while ignoring certain short-term inputs through a dedicated mechanism.<sup>74</sup> The basic LSTM cell contains two internal states, the hidden state  $h_t$  and the cell state  $c_t$ , and can be described through the following set of equations:

$$\dot{i}_t = \sigma(W^i x_t + U^i h_{t-1}), \quad (1)$$

$$f_t = \sigma(W^f x_t + U^f h_{t-1}), \quad (2)$$

$$o_t = \sigma(W^o x_t + U^o h_{t-1}), \quad (3)$$

$$\tilde{c}_t = \tanh(W^c x_t + U^c h_{t-1}), \quad (4)$$

$$c_t = \dot{i}_t \odot \tilde{c}_t + f_t \odot c_{t-1}, \quad (5)$$

$$h_t = o_t \odot \tanh c_t, \quad (6)$$

where  $[W^i, W^f, W^o, W^c, U^i, U^f, U^o, \text{ and } U^c]$  are the trainable parameters of the model;  $x_t$  is the input to the cell at the current time-step,  $\tilde{c}_t$  contains the information to be added to the cell state; and  $\dot{i}_t$ ,  $f_t$ , and  $o_t$  represent the update gate, forget gate, and output gate, respectively, which are numbers between (0 and 1) that control how much information should be updated, discarded, or retrieved from the cell state.  $\sigma$  denotes the sigmoid function, and  $\odot$  represents element-wise multiplication.

The recurrent units inherently formulate a conditional probability between individual torsion angles at the local level that is chained to create a global representation of the entire chain. The eight torsion angle vectors representing the backbone and sidechain torsion angles in a residue are concatenated with a 64 length embedding layer that encodes the amino acid type of a triplet of the previous, current, and subsequent residues. We use alanine as terminal padding for the last amino acid type triplet. Together, they are transformed through a two-layer, fully connected multi-layer perceptron (MLP) with a Rectified Linear unit (ReLU) activation for each layer. Torsions of residues with less than five sidechain angles are padded with zero.

The GRNN architecture has two recurrent units, one for recursion between residues and another inside each residue-level recursion for iteratively processing torsion angles within a residue. This design allows the model to capture correlations of torsion angles between residues as well as correlations between torsion angles within a residue. In the residue-level recurrent unit, the multi-layer perceptron (MLP) outputs are passed to a RNN cell connected to two linear layers corresponding to the  $\omega$  and  $\phi$  torsion angles. The torsion-level recurrent unit enclosed iterates through the rest of the torsion angles ( $\psi, \chi_1, \chi_2, \dots$ ) using the generated  $\phi$  angle. Along with the torsion angle vectors and the MLP outputs, a one-hot encoding of torsion angle types is passed to a RNN cell connected to a linear layer. Each linear layer uses a softmax activation to transform the output into a vector that represents the probability

distribution of a torsion angle. The residue-level RNN cell contains two stacked LSTMs with a hidden size of 200 and dropouts of 0.1, while the torsion-level RNN cell uses one LSTM with the same hidden size and dropout configurations. The GRNN is implemented using PyTorch. We illustrate the design of the recurrent units in the supplementary material (Fig. 1) and describe the details of the pre-training procedure of the generative model and conformer generation in Appendices A and B.

## B. GRNN combined with supervised learning

After the pre-training step, we bias the GRNN toward generating new underlying conformers such that the resulting ensembles better agree with the measured experimental data. The generation of a conformer is formulated as a Markovian decision process (MDP), where the torsional probability distributions in the GRNN naturally define the transitional probabilities between intermediate states. At each torsional generation step, the GRNN implements a mapping from the current state, namely the previously built conformer sequence, to probabilities of selecting possible actions that determine the next set of internal coordinates to build. As a consequence of sampling from the internal coordinate distribution to then generate 3D conformers, the GRNN model receives a reward measuring the agreement between the back-calculation and the experimental data, which provides feedback to the model on how the internal coordinate distributions themselves should change.

This can be considered a reinforcement learning (RL) type of problem and solution. However, unlike a common RL model where the gradient of the rewards is assumed to be unknown, in our GRNN, the gradient of the rewards is readily defined as there are analytical functions that interconvert between the internal coordinates and 3D conformers and between 3D conformers and their experimental observables. However, what is unknown is how the underlying  $\phi, \psi, \omega$  distribution changes given the reward. This is such as the RL problem in which the policy gradient is not defined but must be learned as a set of actions that change the underlying distribution, i.e., the gradient is not being formulated conformer by conformer to improve the X-EISD score but instead by how the  $\phi, \psi, \omega$  distribution changes to improve the entire ensemble. In particular, an optimal strategy of actions that maximizes the expected return, which can be approximated as the sum of rewards  $r_\Theta$  with network parameter  $\Theta$  through sampling the state-action space  $s_T$ ,

$$J(\Theta) = \mathbb{E}_{s_T \sim p(s_T)} [r_\Theta(s_T)] \quad (7)$$

$$\approx - \sum_{s_T} \gamma(s_T) (V(s_T, \Theta) - \hat{V}(s_T))^2, \quad (8)$$

is analogous to minimizing the loss between the back-calculation  $V(s_T, \Theta)$  of an ensemble of sampled structures (trajectories of torsion angles) and the target experimental observables  $\hat{V}$  with a parameter  $\gamma$  that weighs the loss of different states.<sup>75</sup>

To embrace the uncertainties  $\sigma_{\text{exp}}$  associated with these experimental data, we also devise a “flat-bottom” loss by only performing gradient updates on the terms of which the back-calculations are outside of the experimental uncertainty ranges,

$$J(\Theta) = - \sum_{V(s_T, \Theta) \notin [\hat{V} - \sigma_{\text{exp}}, \hat{V} + \sigma_{\text{exp}}]} \gamma(s_T) (V(s_T, \Theta) - \hat{V}(s_T))^2. \quad (9)$$

Even so, it is probably more accepted to define our model as a GRNN that uses the reward function to bias the probability distributions of the torsions through agreement with experimental data using supervised learning.

We train models that are biased with J-couplings (JCs), nuclear Overhauser effects (NOEs), and paramagnetic relaxation enhancements (PREs). J-couplings (JCs) are defined by the backbone  $\phi$  torsion angle  $H_N - N - C_\alpha - H_\alpha$ , and the ensemble average is back-calculated using the Karplus equation<sup>76</sup> as

$$V(\phi) = \langle A \cos(\phi - \phi_0)^2 + B \cos(\phi - \phi_0) + C \rangle, \quad (10)$$

where  $\phi_0$  is a reference state offset of  $60^\circ$  and  $A$ ,  $B$ , and  $C$  are back-calculation parameters sampled as random Gaussian variables<sup>40</sup> with mean and standard deviation values provided in the work of Vuister *et al.*<sup>77</sup> NOEs and PREs back-calculations are modeled as the ensemble averaged distance  $D$  of  $N$  structures using the ENSEMBLE approach,<sup>23–25</sup>

$$D = \left( \frac{\sum_{i=1}^N d_i^{-6}}{N} \right)^{-1/6}. \quad (11)$$

For joint optimization with multiple data types, the total reward function sums up the reward for each data type according to Eq. (9) with a weight hyperparameter. We describe the details of the GRNN training procedure in [Appendices A and B](#).

To keep the gradient information of the back-calculations generated from the sampled torsion angles, we utilize Gumbel-Softmax<sup>78</sup> as a differentiable reparameterization trick that allows sampling from a categorical distribution of  $i$  classes during the forward pass of a neural network. The sample vector  $y_i$  from the generated torsion distribution with probabilities  $p_i$  is expressed as

$$y_i = \frac{\exp((\log(p_i) + g_i)/\lambda)}{\sum_i \exp((\log(p_i) + g_i)/\lambda)}, \quad (12)$$

where  $g_i$  denotes noise generated from a Gumbel distribution and the softmax function is taken over the reparameterized distribution with a temperature hyperparameter  $\lambda$ . We use an annealing schedule that starts from 1 and gradually decreases the temperature by an order of 0.98 for each training iteration. This annealing process balances between model accuracy and variance associated with temperature: the models are trained robustly with low variance at high temperatures initially, and as the model parameters began to converge, the temperature lowering ensures accuracy without causing significant instability.<sup>78</sup> This recast of a stochastic generation process allows the model to trace the rewards based on distance type restraints to specific torsion angles through an internal to Cartesian coordinates conversion (see [Appendices A and B](#)), thereby overcoming the difficulties of a generative model defined in a torsional space being less sensitive to tertiary contact restraints, such as NOEs and PREs, as compared to local and angular restraints, such as backbone J-couplings.

The best model is selected based on the X-EISD score of the generated structures during the validation steps. The use of a Bayesian model for validation furnishes the GRNN model with a better probabilistic interpretation of disordered protein ensembles by modulating different sources of uncertainties in the experimental

data types. We briefly summarize the details of the X-EISD calculation and reweighting approach in [Appendices A and B](#) and refer readers to Refs. 36 and 40.

### C. Timings for the unbiased and biased GRNN algorithms

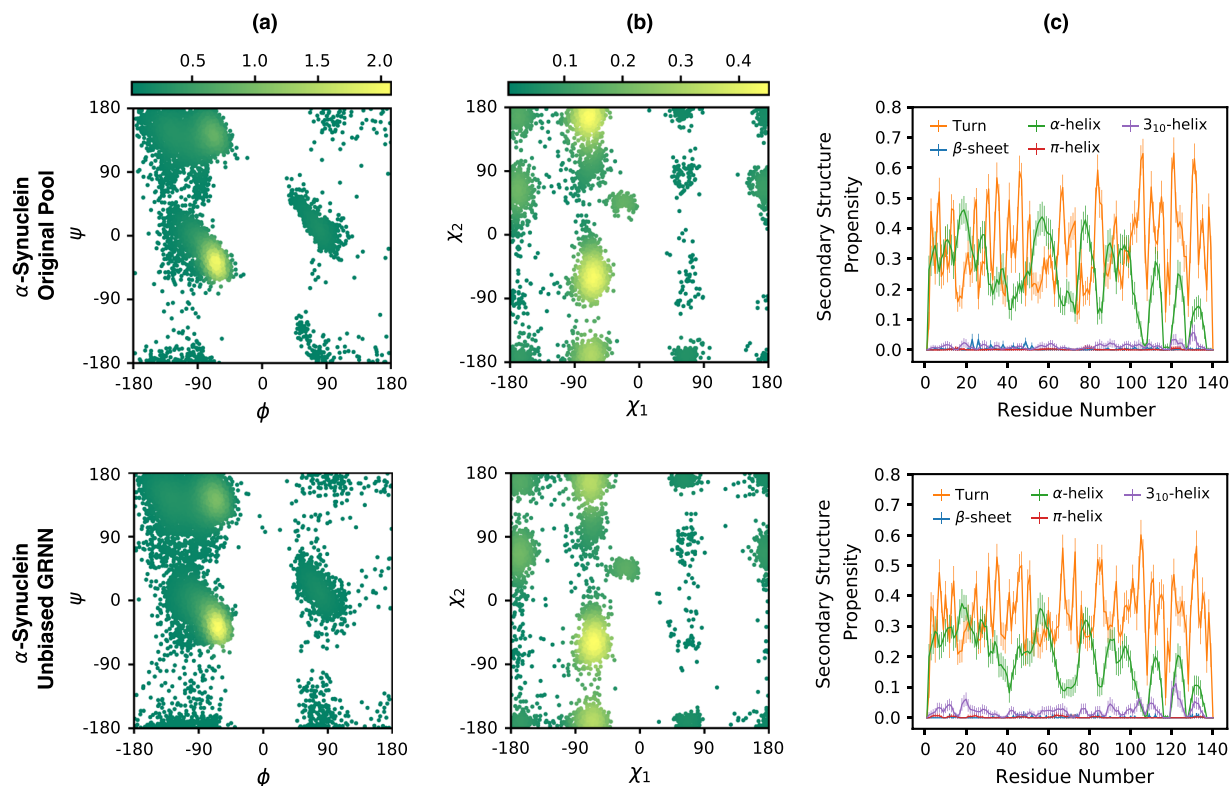
The computational cost of both algorithms scales approximately linearly with system size. On a single GTX 1070 Ti graphics processing unit (GPU), the small protein Hst5 required 0.2 min/iteration for the unbiased GRNN and 0.74 min/iteration for the biased GRNN for 6000 samples using a 100 sample batch size. For the largest protein evaluated here,  $\alpha$ -Syn, the unbiased GRNN required 0.83 and 3.9 min/iteration for the biased GRNN step for 4000 samples using 100 sample batch sizes. It usually takes between 100 and 200 iterations to converge the unbiased and biased GRNNs.

## III. RESULTS

All initial conformer pools for Hst5, A $\beta$ 40, uDrkN-SH3, and  $\alpha$ -Syn are generated using IDPConformerGenerator.<sup>20</sup> IDPConformerGenerator is a flexible software platform that can be used to create conformers based on torsion angles from any secondary structure combination. In this work, we use it to randomly sample only loop and extended state torsion angles to make a conformer pool lacking helix for uDrkN-SH3, since the uDrkN-SH3 protein is known to have local regions of helical structure, in order to generate a clear test case with a starting pool that is missing relevant conformers. For Hst5 and A $\beta$ 40, we also start with conformer pools containing only loop and extended state structures. Correspondingly, for  $\alpha$ -Syn, we randomly sample only loop and helical state torsion angles to make a conformer pool without extended conformations, although the  $\alpha$ -Syn protein is largely known to be featureless in regard to secondary structure signatures. All of these are intended to be challenging cases, i.e., the underlying conformational pools are poor “start states” for a reweighting algorithm. However, as a control, we also start with randomly sampled torsion angles comprising loop and extended states for  $\alpha$ -Syn, which should favor the reweighting algorithm, i.e., a starting pool in which relevant conformations are (fortuitously) present, to see how the DynamICE compares for this special case.

We begin with the GRNN that learns the torsional statistics of the backbone and side chains of the given protein sequence from the respective starting conformational pools from the IDPConformerGenerator. We perform this “pre-training” step for three reasons: (1) to demonstrate that our unbiased GRNN can make reasonable IDP structural ensembles, as almost all ML models for IDP ensemble generation in the literature are generative only. (2) IDPConformerGenerator is a discrete ensemble, but the underlying generative model is defined as a continuous distribution. (3) Finally, and most importantly, it defines the initial condition that the underlying structural distribution shifts away from to conform to the data. This last condition is important as we use pre-training to get the physical initial conditions of the backbone but especially the sidechains, because there is no energy function during biased GRNN optimization and the experimental data are sparse.

[Figure 2](#) compares the Ramachandran plot for the backbone and sidechain torsions from the unbiased GRNN model and



**FIG. 2.** Properties of ensembles for the  $\alpha$ -Syn IDP from the original pool comprised of loops and helices and from the unbiased GRNN. (a) Ramachandran plots displaying the backbone torsion angle distributions and (b) histograms displaying the  $\chi_1 - \chi_2$  distributions from 100 structures of the training data (top) and unbiased GRNN (bottom). The density values are scaled by  $1 \times 10^{-4}$ . (c) Secondary structure propensities per residue among 50 independently drawn ensembles of 100 structures. The error bars are shown as  $\pm 1$  standard deviation. The corresponding plots for the other proteins are shown in the supplementary material, Figs. 2 and 3.

the original training conformer pools and their agreement with the percentage of secondary (local) structure per residue of all the major secondary structure categories for the  $\alpha$ -Syn IDP; the supplementary material, Figs. 2 and 3, provides the same results for the loop/extended states used for the uDrkN-SH3,  $\alpha$ -Syn, Hst5, and A $\beta$ 40 proteins. Table I provides quantitative metrics to evaluate the underlying structural differences of the original pool and unbiased GRNN ensembles in terms of global shape characteristics, such as the radius of gyration  $R_g$ , end-to-end distance  $R_{ee}$ , and asphericity  $\delta^*$ , which measures the anisotropy of the ensemble. Table I in the supplementary material also demonstrates that the unbiased GRNN models and their respective original conformer pools also yield similar root mean squared error (RMSD) of the back calculated NOE, with respect to various experimental data types, providing additional evidence that the unbiased GRNNs are robust for all four proteins.

After the pre-training step, we bias the GRNN toward generating new underlying conformers such that the resulting ensembles better agree with the measured experimental data through an additional biasing step. For uDrkN-SH3, A $\beta$ 40, and both aqueous and DMSO solvent Hst5, we perform a GRNN optimization with JCs and NOEs to create biased ensembles. We chose these two data types as our previous study with X-EISD has shown that dual reweighting

optimization of local data, such as JCs, and long-ranged restraints, such as NOEs, can yield ensembles that simultaneously improve other data types.<sup>40</sup> For  $\alpha$ -Syn, we train DynamICE by jointly optimizing JCs and PREs. Given the longer-ranged contacts for PREs compared to NOEs, the biased GRNN must concertedly drive a greater number of torsional changes to meet each distance restraint. Furthermore, the reported experimental error estimates tend to be uncertain, implying that the distances measured are less precise due to the dynamics of the probe,<sup>23,79</sup> contamination of the diamagnetic protein,<sup>80</sup> and perturbations to the IDP structural ensemble due to the probe label.<sup>81,82</sup> PREs are a more difficult class of experimental data as they tend to be biased toward longer-ranged contacts between different parts of the sequence compared to NOEs, which require the model to cooperatively change the torsion angles of a large number of residues to meet a contact restraint. To best demonstrate the ability of DynamICE, despite the torsional representation limitation, we optimize both a subset of PRE data that only contains contacts equal to or less than ten residues apart as well as the full set of PRE data.

Figure 3 and the supplementary material, Fig. 4, demonstrate that DynamICE applied to Hst5 shows strong agreement with JCs measured in both aqueous and organic phases. While the RMSD

**TABLE I.** Evaluation of the unbiased GRNN model, the standard reweighted ensemble optimization, and the DynamICE model for disordered states and variable solvent conditions for Hst5, A $\beta$ 40, uDrkN-SH3, and  $\alpha$ -Syn.<sup>a</sup> The experimental data RMSDs include J-couplings (JCs), nuclear Overhauser effects (NOEs), paramagnetic relaxation enhancement (PREs), single molecule FRET (smFRET), chemical shifts (CS), and small angle x-ray scattering (SAXS).<sup>b</sup> Global metrics of the ensembles include adius of gyration  $R_g$ , end-to-end distance  $R_{ee}$ , and ensemble asphericity  $\delta^*$ .<sup>b</sup> All values are reported in terms of mean and standard deviation (in parentheses) over 50 ensembles of 100 structures each. The chemical shift RMSDs by atom types are shown in the supplementary material, Table IV. The uDrkN-SH3 MD ensemble uses trajectories randomly sampled from the 30  $\mu$ s simulation using the a99SB-disp force field.<sup>48</sup> Here, we report models trained using the “flat-bottom” loss function Eq. (9), which better interprets the experimental errors and uncertainties, and the models trained using the simple squared loss Eq. (8) are given in the supplementary material, Table II.

	Experimental data type RMSD								
	JC (Hz)	NOE (Å)	PRE (Å)	smFRET $\langle E \rangle$	CS (ppm)	SAXS (Intensity)	$R_g$ (Å)	$R_{ee}$ (Å)	$\delta^*$
Hst5 ensembles UNOPTIMIZED (loop/extended) and OPTIMIZED with JCs and NOEs taken in the aqueous phase									
Unbiased GRNN	0.662 (0.043)	0.359 (0.005)			0.138 (0.004)	0.001 (0.001)	13.59 (1.79)	32.22 (10.49)	0.410 (0.169)
Reweight	0.582 (0.019)	0.371 (0.006)			0.141 (0.004)	0.001 (0.001)	13.26 (2.00)	31.40 (10.58)	0.402 (0.178)
DynamICE	0.317 (0.029)	0.363 (0.025)			0.150 (0.004)	0.001 (0.001)	14.14 (2.05)	34.62 (10.61)	0.422 (0.180)
Hst5 ensembles UNOPTIMIZED (loop/extended) and OPTIMIZED with JCs and NOEs taken in DMSO organic solvent									
Unbiased GRNN	1.274 (0.046)	2.933 (0.234)			0.171 (0.004)		13.59 (1.79)	32.22 (10.49)	0.410 (0.169)
Reweight	1.308 (0.033)	2.158 (0.176)			0.164 (0.004)		13.17 (1.95)	31.42 (10.30)	0.397 (0.177)
DynamICE	0.532 (0.037)	1.173 (0.146)			0.161 (0.004)		11.67 (1.36)	27.33 (8.50)	0.349 (0.165)
A $\beta$ 40 ensembles UNOPTIMIZED (loop/extended) and OPTIMIZED with JCs and NOEs									
Unbiased GRNN	0.797 (0.030)	1.361 (0.115)		0.146 (0.034)	0.734 (0.025)		18.64 (3.40)	44.85 (16.38)	0.448 (0.192)
Reweight	0.716 (0.017)	1.142 (0.094)		0.156 (0.037)	0.795 (0.026)		18.58 (3.59)	44.68 (16.41)	0.455 (0.193)
DynamICE	0.471 (0.034)	1.308 (0.127)		0.105 (0.034)	0.822 (0.028)		17.65 (3.07)	41.62 (15.42)	0.417 (0.193)
uDrkN-SH3 ensembles UNOPTIMIZED (loop/extended) and OPTIMIZED with JCs and NOEs									
Unbiased GRNN	1.440 (0.028)	6.343 (0.429)	7.711 (1.193)	0.228 (0.032)	0.495 (0.007)	0.007 (0.001)	23.16 (4.81)	55.51 (21.21)	0.431 (0.202)
MD	0.730 (0.033)	4.764 (0.254)	5.104 (0.650)	0.074 (0.034)	0.400 (0.011)	0.003 (0.001)	19.69 (3.89)	45.56 (18.07)	0.369 (0.174)
Reweight	1.398 (0.017)	5.208 (0.365)	7.213 (1.381)	0.208 (0.027)	0.493 (0.009)	0.007 (0.001)	22.35 (4.33)	52.95 (19.26)	0.421 (0.192)
DynamICE	0.693 (0.033)	5.242 (0.410)	6.346 (1.073)	0.119 (0.035)	0.478 (0.010)	0.004 (0.001)	20.28 (3.68)	48.58 (17.52)	0.401 (0.193)
$\alpha$ -Syn ensembles UNOPTIMIZED (helix/loop) and OPTIMIZED with JCs and PREs (all data)									
Unbiased GRNN	0.622 (0.032)		9.923 (0.351)	0.103 (0.004)	0.612 (0.019)	0.017 (0.002)	33.99 (7.61)	78.66 (33.80)	0.426 (0.196)
Reweight	0.528 (0.048)		6.372 (0.158)	0.112 (0.004)	0.638 (0.026)	0.014 (0.001)	35.09 (7.30)	83.48 (32.45)	0.444 (0.201)
DynamICE	0.524 (0.017)		8.992 (0.355)	0.145 (0.005)	0.566 (0.002)	0.025 (0.002)	43.81 (9.44)	104.25 (39.38)	0.454 (0.198)



TABLE I. (Continued.)

	Experimental data type RMSD								
	JC (Hz)	NOE (Å)	PRE (Å)	smFRET $\langle E \rangle$	CS (ppm)	SAXS (Intensity)	$R_g$ (Å)	$R_{ee}$ (Å)	$\delta^*$
$\alpha$ -Syn ensembles UNOPTIMIZED (loop/extended) and OPTIMIZED with JCs and PREs (all data)									
Unbiased GRNN	0.704 (0.022)		10.088 (0.395)	0.108 (0.005)	0.558 (0.003)	0.013 (0.001)	37.18 (7.90)	84.73 (34.32)	0.443 (0.187)
Reweight	0.622 (0.015)		6.200 (0.175)	0.119 (0.005)	0.550 (0.004)	0.014 (0.001)	38.49 (8.63)	87.89 (35.17)	0.432 (0.202)
DynamICE	0.655 (0.009)		9.365 (0.309)	0.133 (0.010)	0.588 (0.003)	0.026 (0.002)	41.77 (9.91)	94.80 (40.57)	0.416 (0.192)

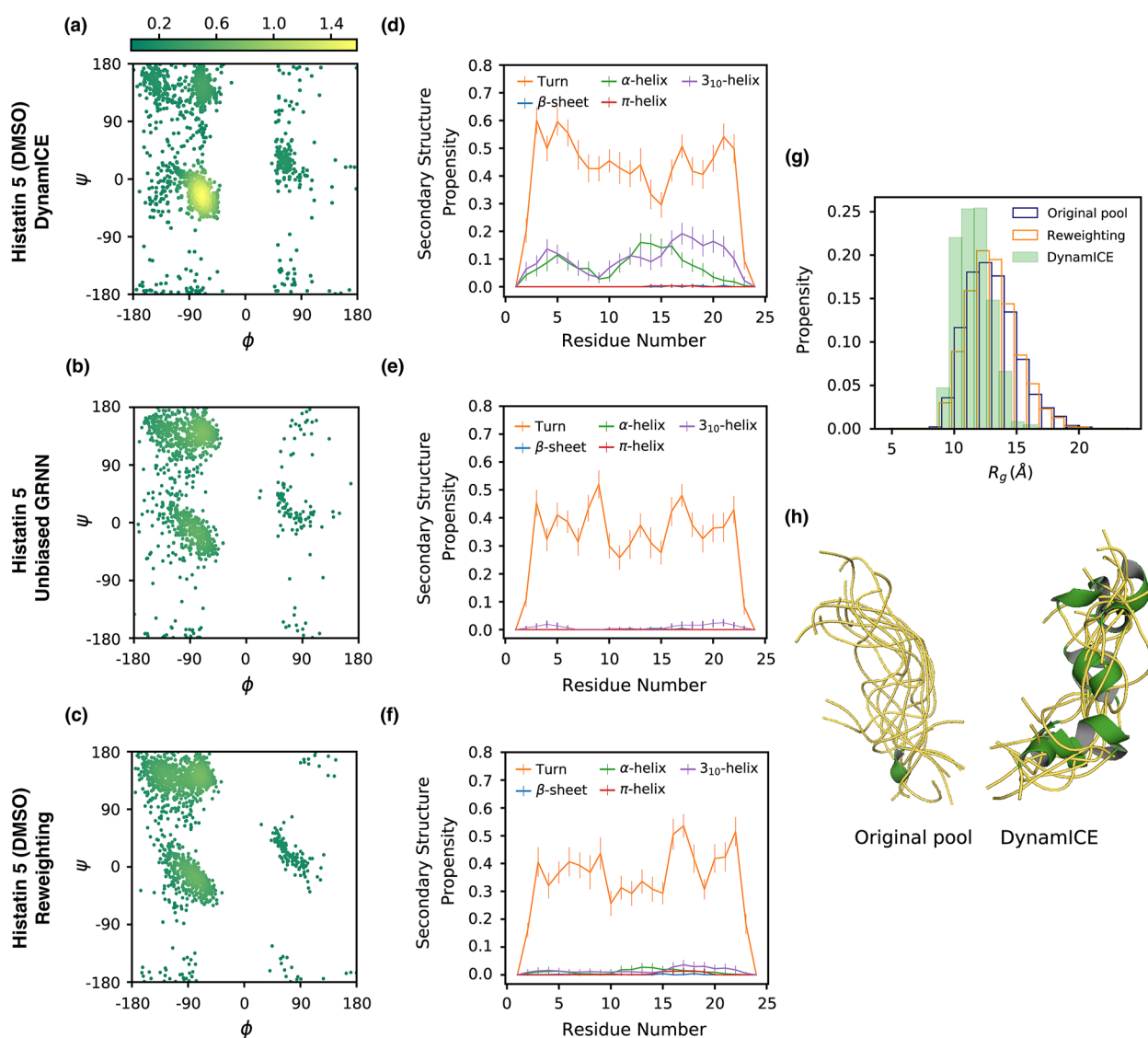
<sup>a</sup>The experimental (expt.) and back-calculation (bc) errors for CS [ $\sigma_{\text{exp}} = 0.03\text{--}0.3$  ppm;  $\sigma_{\text{bc}} = 0.3\text{--}0.5$  ppm (hydrogen), 1.2–1.4 ppm (carbon)]; JCs ( $\sigma_{\text{exp}} = 0.5$  Hz,  $\sigma_{\text{bc}}^A = 0.14$  Hz,  $\sigma_{\text{bc}}^B = 0.03$  Hz,  $\sigma_{\text{bc}}^C = 0.08$  Hz); NOEs ( $\sigma_{\text{exp}} = 5.0$  Å;  $\sigma_{\text{bc}} = 0.0001$  Å); PREs ( $\sigma_{\text{exp}} = 5.0$  Å;  $\sigma_{\text{bc}} = 0.0001$  Å); smFRET  $\langle E \rangle$  ( $\sigma_{\text{exp}} = 0.02$ ;  $\sigma_{\text{bc}} = 0.0074$ ); and SAXS ( $\sigma_{\text{exp}} = 0.0008\text{--}0.002$ ;  $\sigma_{\text{bc}} = 0.006$ ).

<sup>b</sup> $\delta^*$  measures the anisotropy of the structures, ranging from 0 (sphere) to 1 (rod).

with respect to the NOE data taken in the aqueous phase is marginally better compared to the reweighting optimized ensembles, this is not unexpected as the NOEs measured in the aqueous phase are predominantly sequential contacts ( $i, i + 1$ ) between the alpha carbon and amide hydrogen (Table I and the supplementary material, Fig. 4). This provides limited additional information besides an indication of the unfolded nature of histatins in the aqueous phase, which is already provided by the JC data. Remarkably, for the Hst5 ensembles in DMSO, the DynamICE model shows significantly improved NOE agreement due to its ability to create new helical conformations, which are lacking in the original and unoptimized GRNN pool and consequently in the reweighting optimization (Table I and Fig. 3). When trained with the mean squared error (MSE) loss function [Eq. (8)] that ignores experimental and back-calculation NOE uncertainties, the model can better drive toward longer helical conformations (Table II and Fig. 6 in the supplementary material). It has been argued that since the anti-fungal activities of histatins may be a result of their interaction with cell membranes, the structure–function relationship of histatins is highly relevant to the conformations they adopt in a membrane environment.<sup>70</sup> The ability of DynamICE to contrast the structural ensembles by solvent conditions is an example of how DynamICE can provide functional insight into the experimental data.

DynamICE improves the RMSD of JCs for A $\beta$ 40 by changing the underlying torsions to shift toward a more compact conformer ensemble than the original unbiased A $\beta$ 40 ensemble (Fig. 4). While the A $\beta$ 40 reweighting optimization shows a more visible improvement in NOEs than the DynamICE model, the RMSD improvement is minor and does not necessarily differentiate the two ensembles given the generous back-calculation uncertainties of NOE distances. The DynamICE algorithm promotes a helix sub-population near residue 23 and a turn around residues 13–16 [Figs. 4(d) and 4(h)] as noted in previous studies,<sup>15,83,84</sup> leading to a more spherical and compact conformer ensemble as measured by  $\langle R_g \rangle$ ,  $\langle R_{ee} \rangle$ , and  $\langle \delta^* \rangle$  in Table I. This qualitative change in compaction predicted by DynamICE is independently validated by the smFRET data, although it was not directly optimized.

Table I and Fig. 5 show that the optimization using DynamICE reduces the RMSD of NOE data with respect to the unoptimized ensembles for uDrkN-SH3 and yields similar RMSDs for the NOEs compared to reweighting. However, DynamICE shows a significant RMSD improvement for JCs by changing the conformers of the underlying ensemble. As a result of these new members, an independent validation shows that global configurational metrics, such as PREs, smFRET, and SAXS, are also improved, although these additional experimental data types were not optimized directly. Figure 5 illustrates that these improvements in the optimized and validated metrics for uDrkN-SH3 arise because the backbone torsion angles shift toward the helical region after DynamICE optimization [Fig. 5(a)], leading to a substantial increase in the percentage of helical content from nearly zero to around 10%–30% at residues 10–20 and 30–45, as shown in Fig. 5(d), and are supported by JCs and the NOE data, which include a number of  $i$  to  $i + 3$  or  $i + 4$  contacts around residues 15–20 and 30–40. By contrast, the reweighting optimization barely changes the torsion angle profiles from the unbiased pools [Figs. 5(b) and 5(c)], nor is there a shift in the secondary structure assignments [Figs. 5(e) and 5(f)] due to a lack of relevant conformers in the initial pool to further refine ensembles using the JCs and NOEs data. While reweighting optimization yields ensembles that slightly shift toward more compact and globular-like conformers as measured by  $\langle R_g \rangle$ ,  $\langle R_{ee} \rangle$ , and  $\langle \delta^* \rangle$ , the DynamICE model exhibits a more pronounced shift in  $R_g$  [Fig. 5(g)] and  $R_{ee}$  distributions (Table I) to even more compact disordered states. These new sub-populations of helical structures [Fig. 5(h)] and more compact conformers generated by DynamICE, which are not available in the original pool used in the reweighting scheme, are responsible for better agreement with the overall SAXS intensity profile [Fig. 5(i)]. Thus, by generating physically different conformers, DynamICE directly overcomes the deficiencies of the static initial ensemble. We also performed an additional analysis to use the X-EISD score for the reported MD ensemble result for uDrk-SH3 by Robustelli *et al.*<sup>48</sup> and compared it to DynamICE (Table I). We can see that both ensembles are within the generated blue RMSD uncertainties for J-couplings, NOEs, and PREs, but more importantly, the qualitative



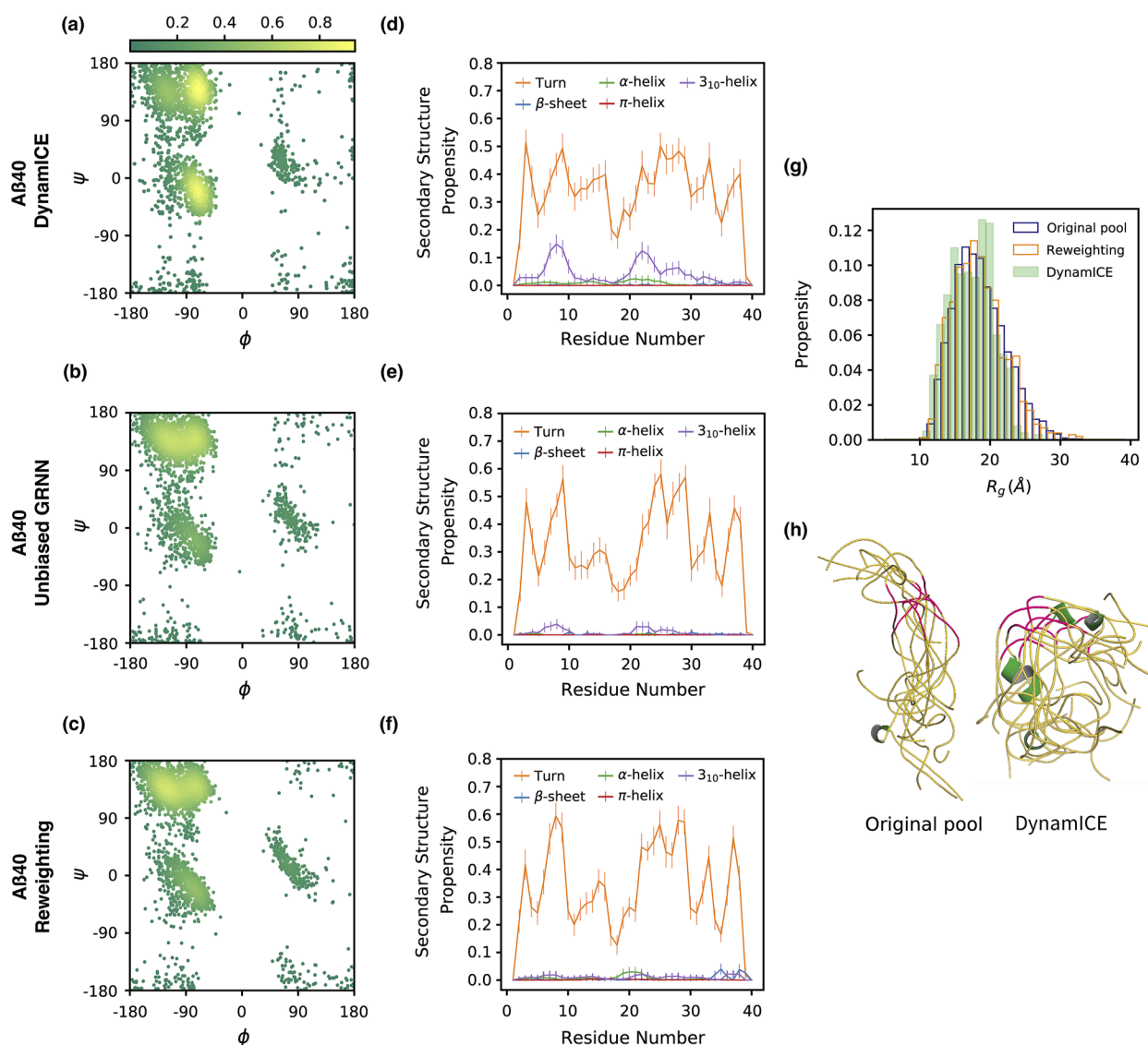
**FIG. 3.** Properties of Hst5 in DMSO from the unbiased ensemble and generated by the DynamICE model compared with reweighting optimization using JCs and NOEs. Ramachandran plots displaying the backbone torsion angle distributions from the (a) biased GRNN, (b) unbiased GRNN, and (c) reweighting optimization. The density values are scaled by  $1 \times 10^{-4}$ . Secondary structure propensities per residue of the (d) biased GRNN, (e) unbiased GRNN, and (f) reweighting optimization. (g) Comparison of the radius of gyration distributions before and after optimization with reweighting and DynamICE. (h) Example ensembles of conformers from the Hst5 original pool and the DynamICE model (helices in green and loops in yellow). Statistical errors from 50 independently drawn ensembles of 100 structures. The error bars are shown as  $\pm 1$  standard deviation. While Table I reports the results for the biased GRNN model trained with the “flat-bottom” loss Eq. (9), we note that the results are even better for Hst5 in DMSO when trained with the simple squared loss function Eq. (8), as reported in the supplementary material, Table II.

trends are the same in the prediction of much more compact ensembles.

We note that the favored helical regions created by DynamICE are comprised of more 3–10 helices as opposed to  $\alpha$ -helices determined in the optimized ensembles of uDrkN-SH3 using ENSEMBLE in previous studies.<sup>24,40</sup> While the total amount of helix is comparable between the original ENSEMBLE implementation and

DynamICE [the supplementary material, Fig. 7(a)], there are several aspects worth noting. First, the ENSEMBLE method used all the available data (chemical shifts, smFRET, SAXS, hydrodynamic radius, etc.) and assumed that no errors existed in the experiment or back-calculation, and the underlying original ensemble was dominated by  $\alpha$ -helices as opposed to 3–10 helices. We note that if we use Eq. (9), which assumes no experimental or back-calculation



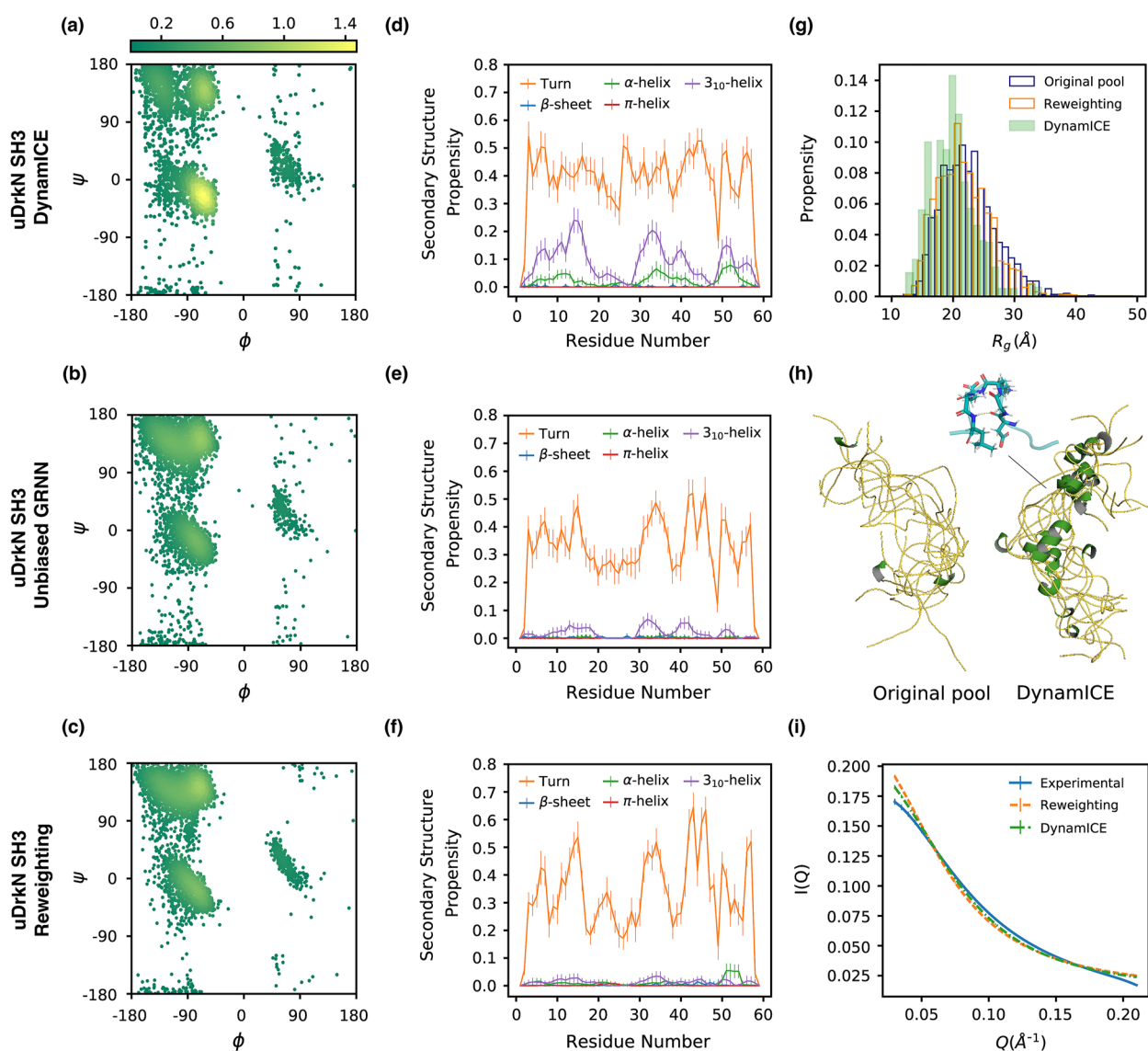


**FIG. 4.** Properties of the A $\beta$ 40 unbiased ensemble and generated by the DynamICE model compared with reweighting optimization using JCs and NOEs. Ramachandran plots displaying the backbone torsion angle distributions from the (a) DynamICE, (b) unbiased GRNN, and (c) reweighting optimization. The density values are scaled by  $1 \times 10^{-4}$ . Secondary structure propensities per residue of the (d) DynamICE, (e) unbiased GRNN, and (f) reweighting optimization. (g) Comparison of the radius of gyration distributions before and after optimization with reweighting and DynamICE. (h) Examples of conformers from the Hst5 original pool and the DynamICE model (helices in green; loops in yellow; and residues 13–16 highlighted in magenta). Statistical errors from 50 independently drawn ensembles of 100 structures. The error bars are shown as  $\pm 1$  standard deviation.

uncertainties, we find better agreement with ENSEMBLE in the position of the helical regions [the supplementary material, Fig. 7(b)]. However, because the experimental J-coupling data cannot differentiate between the two sub-classes of helices and the  $i, i + 3$ , and  $i, i + 4$  NOEs can support both the 3–10 and  $\alpha$  helices, DynamICE determines more 3–10 helices (Fig. 5). This offers an interesting question as to whether the DynamICE ensemble is a physical outcome, i.e., IDPs may form incipient alpha-helices by first forming 3–10 helices,

or whether another data type, such as chemical shifts, could drive toward one of the helical subclasses.

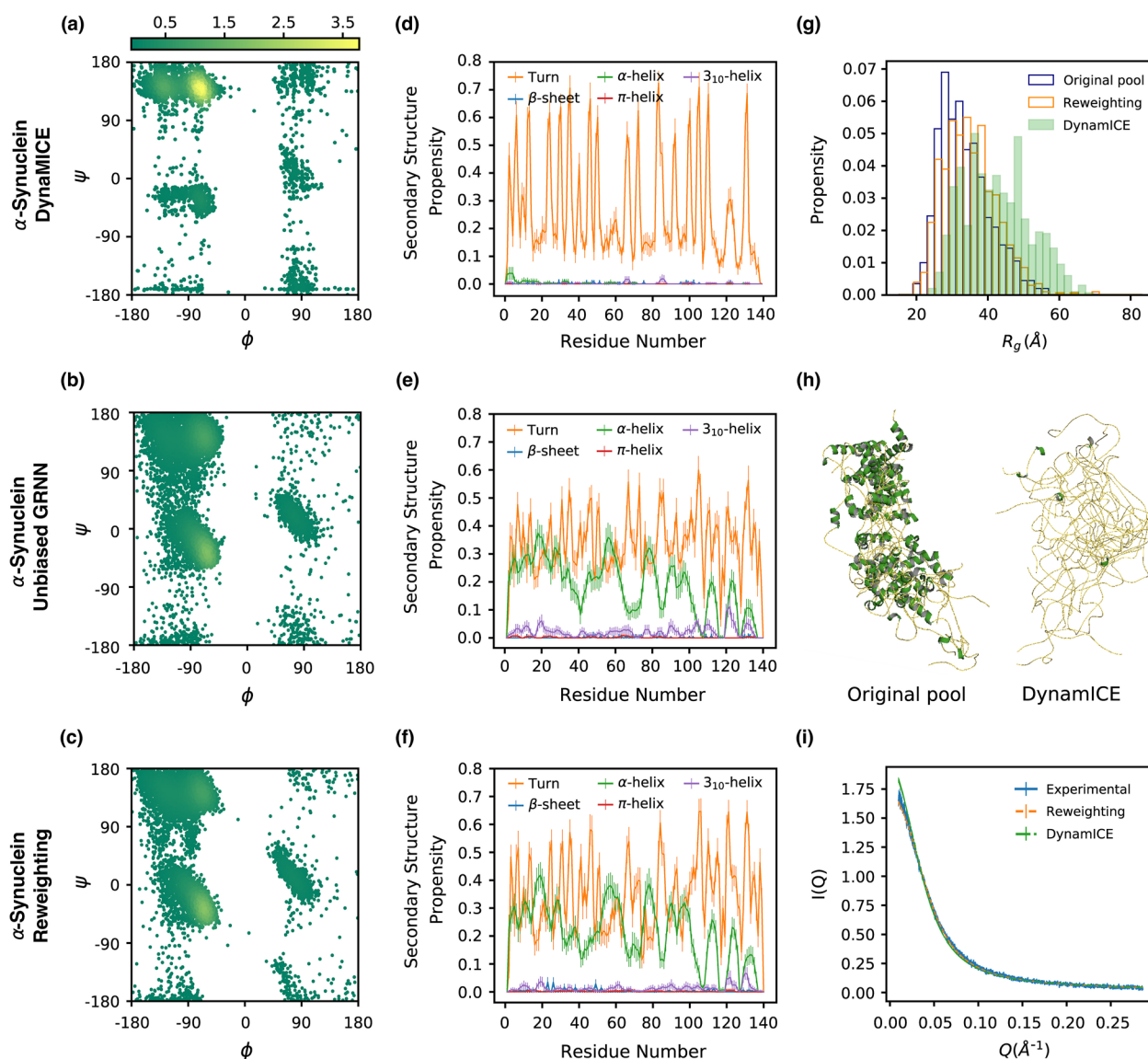
The DynamICE results for the  $\alpha$ -Syn IDP starting from an unbiased pool containing loops and helices are shown in Fig. 6 and Table I. As in the case for uDrkN-SH3, both the reweighting and DynamICE models achieve improvements in JCs and PRE data types compared with the unbiased GRNN for  $\alpha$ -Syn (Table I). Given the uncertainties in PRE distances of  $5 \text{\AA}$  the reweighting and DynamICE



**FIG. 5.** Properties of the uDrkN-SH3 domain unbiased ensemble and generated by the DynamICE model compared with reweighting optimization using JCs and NOEs. Ramachandran plots displaying the backbone torsion angle distributions from the (a) DynamICE, (b) unbiased GRNN, and (c) reweighting optimization. The density values are scaled by  $1 \times 10^{-4}$ . Secondary structure propensities per residue of the (d) DynamICE, (e) unbiased GRNN model, and (f) reweighting optimization. (g) Comparison of the radius of gyration distributions before and after optimization with reweighting and DynamICE. (h) Examples of conformers from the uDrkN-SH3 original pool and the DynamICE model (helices in green and loops in yellow). Conformers generated with the DynamICE model also exhibit short cooperative secondary structures, such as  $\beta$ -turns. (i) SAXS intensity curves for DynamICE and reweighting optimized ensembles compared with the experimental data. Statistical errors from 50 independently drawn ensembles of 100 structures. The error bars are shown as  $\pm 1$  standard deviation.

RMSDs are not distinguishable. However, unlike the case of uDrkN-SH3, in which DynamICE drove to more compact ensembles, the combination of J-coupling and PRE data drives the backbone torsion angles toward the polyproline-II region ( $\phi = -90^\circ$  to  $-25^\circ$  and  $\psi = 120^\circ$ – $150^\circ$ ) after DynamICE optimization [Fig. 6(a)] compared to the unbiased and reweighted ensembles [Figs. 6(b) and 6(c)]. In particular, we see that the DynamICE optimization has largely

eliminated helical torsions when compared to the unbiased and reweighted ensembles [Figs. 6(d)–6(f)], as these conformational states are not unambiguously supported by the JC and PRE experimental data. Since the DynamICE model introduces more extended conformations [Fig. 6(h)], it noticeably shifts the  $R_g$  [Fig. 6(g)] and  $R_{ee}$  (Table I) distributions in comparison with the unoptimized and reweighted ensembles and exhibits good agreement with the SAXS



**FIG. 6.** Properties of the  $\alpha$ -Syn unbiased ensemble of helices/loops and ensembles optimized by the DynamICE model compared with reweighting optimization using JCs and PREs (all data). Ramachandran plots displaying the backbone torsion angle distributions from the (a) DynamICE, (b) unbiased GRNN, and (c) reweighting optimization. The density values are scaled by  $1 \times 10^{-4}$ . Secondary structure propensities per residue of the (d) DynamICE, (e) unbiased GRNN, and (f) reweighting optimization. (g) Comparison of the radius of gyration distributions before and after optimization with reweighting optimization and DynamICE. (h) Examples of conformers from the  $\alpha$ -Syn original pool and the DynamICE model (helices in green and loops in yellow). (i) SAXS intensity curves for DynamICE and reweighting optimized ensembles compared with the experimental data. SAXS intensity is scaled by 0.001. Statistical errors from 50 independently drawn ensembles of 100 structures. The error bars are shown as  $\pm 1$  standard deviation.

data [Fig. 6(i)]. Since no qualitative differences are found when starting from the loop/extended vs helix/loop pools, we believe that the limited experimental data of JCs and NOEs as formulated can't distinguish between the two qualitatively different ensembles optimized by the DynamICE and reweighting approaches, and both are reasonably validated by the other data types. In fact, Table III and Fig. 5 in the supplementary material, show that if we restrict the PREs to be short-ranged in sequence separation (labels less than

ten amino acids apart), the support for more extended states is still strong, although long-ranged PRE data are missing.

#### IV. DISCUSSION

Presently, most methods for creating disordered ensembles that are consistent with available experimental solution data are

separated into two steps. The first is to create a static pool of conformations, and the second is to improve upon that pool by reweighting different sub-populations of conformations to improve a score that reflects better experimental agreement.<sup>35,36,39,43,44</sup> If the underlying static pool is insufficient, i.e., if relevant conformations are absent, there is little that can be solved with reweighting approaches. Instead, the first step needs to be revisited to create new structural pools in the hope that the new underlying basis set of conformations can be made more consistent with experimental observables. Alternatively, methods have been developed that generate new conformations by using unbiased<sup>48</sup> or biased molecular dynamics simulations with experimental data,<sup>37,38</sup> with the benefit that such ensembles model physical interactions among side chains and can create transient local structure details, such as the formation and packing of hydrophobic cores and local folding of transiently structured domains with recognizable secondary structure motifs, but can be computationally costly.

This work offers a conceptual alternative to such existing approaches through a machine learning method that simultaneously physically changes the conformations of the underlying pool to evolve new structural ensembles that agree with experimental solution data at minimal computational cost and with no inherent biases. In particular, DynamICE biases the probability of the residue torsions of a chain molecule, generating new sub-populations of disordered states using a reward mechanism that simultaneously improves agreement with experimental data based on X-EISD scores. Currently, DynamICE biases the probability distributions of torsions to take advantage of experimental data types, such as J-couplings, NOEs, and PREs, but extensions to other data types, such as smFRET, chemical shifts, and SAXS, are certainly possible and are under way in our development of this nascent method. Given the computational timings reported in the section titled Method, the DynamICE training costs are negligible compared to molecular dynamics methods and do not simply yield evolving conformer snapshots but instead are an optimized ensemble generator. Even so, if a given MD force field is suitable for this class of protein, we could use it as a structural prior or even use it to replace IDPConformerGenerator, which we will explore in future work.

As a proof-of-concept of the DynamICE method, we applied this approach by biasing toward experimental <sup>3</sup>JCs and NOEs for the unfolded state of the DrkN SH3 domain, human salivary histatin 5 in aqueous and DMSO solutions, and A $\beta$ 40 peptide, and <sup>3</sup>JCs and PREs for the  $\alpha$ -synuclein IDP. We showed that the DynamICE approach generates ensembles of vastly different underlying structural characteristics from their starting pools, to better conform to their individual experimental data restraints. However, driving a model that uses an internal coordinate representation of protein conformers to meet distance restraints, such as NOEs and PREs, is not yet fully optimal. To utilize more effectively the distance/contact-based data in the reward function, we could consider more hidden states in the LSTM that replace the prediction of individual residue torsions with the prediction of larger structural fragments to maintain secondary structure motifs. It is also a limitation of the torsion-based (local) protein representation, and in the future, we could explore the use of a message passing neural network (MPNN), which can represent the 3D coordinates of the protein conformers directly to better handle distance restraints.

Posing IDP conformer generation as a problem evaluated by rewards introduces several benefits over the recent generative models that simply learn to represent a conformational landscape or reweighting methods that require that all relevant conformations be present. As such, the DynamICE method provides a natural framework to combine the scoring and conformer generation steps simultaneously with the experimental data, as opposed to requiring a separation of the scoring and conformer search of a starting/training conformer pool, and still considers the various errors and uncertainties of a Bayesian model. We believe that the DynamICE approach is a paradigm shift in how to address the overall conformational search problem for disordered states of proteins by allowing the underlying structural pools to *evolve* toward experimental data under a Bayesian model that reflects statistical uncertainties. In summary, by showing the ability of DynamICE to differentiate among disordered and ordered states across variable solvent conditions for the four proteins used here, our approach will allow for greater functional insight to better support the structure–function relationship.

## SUPPLEMENTARY MATERIAL

See the supplementary material for additional training results.

## ACKNOWLEDGMENTS

All authors acknowledge the funding and support from the National Institute of Health under Grant No. 5R01GM127627-04. J.D.F.-K. also acknowledges the support from the Natural Sciences and Engineering Research Council of Canada (Grant No. 2016-06718) and from the Canada Research Chairs Program.

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

## Author Contributions

**Oufan Zhang:** Conceptualization (equal); Formal analysis (equal); Methodology (lead); Software (lead); Validation (equal); Writing – original draft (lead); Writing – review & editing (equal). **Mojtaba Haghghatdari:** Conceptualization (lead); Methodology (equal); Software (equal). **Jie Li:** Formal analysis (supporting); Validation (equal); Writing – review & editing (supporting). **Zi Hao Liu:** Software (equal); Writing – review & editing (equal). **Ashley Namini:** Data curation (equal); Validation (equal). **João M. C. Teixeira:** Software (equal); Writing – review & editing (equal). **Julie Forman–Kay:** Funding acquisition (equal); Writing – review & editing (equal). **Teresa Head-Gordon:** Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Resources (lead); Supervision (lead); Writing – original draft (equal); Writing – review & editing (lead).

## DATA AVAILABILITY

The DynamICE package has been implemented as a publicly accessible Python package at <https://github.com/THGLab/DynamICE> for the reproducibility of future studies. The data



that support the findings of this study are available from the corresponding author upon reasonable request.

## APPENDIX A: DYNAMICE MODEL AND ALGORITHM

### 1. Protein conformer representation

Assuming ideal bond lengths and bond angles, a protein conformer can be represented by a sequence of the backbone and sidechain torsion angles for each residue  $j$  ( $\omega_j, \phi_j, \psi_j, \chi_{j1}, \chi_{j2}, \dots, \chi_{j5}$ ). By parameterizing protein structures in the torsional space, the generative model covers conformations in a reduced dimension while preserving local chemical connectivity. The torsional space is discretized into  $2^\circ$  bins over the range of  $[-180, 180]$  such that each torsion angle is represented by a vector of size 180 with elements corresponding to the relative probability of finding the angle at each angle bin. The relative probability of each bin is calculated by a Gaussian distribution with a  $1^\circ$  standard deviation  $\sigma$  to allow for flexibility,

$$P_\phi(i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(2i-180-\phi)^2}{2\sigma^2}\right), \quad (\text{A1})$$

where  $i$  is the bin index in the range of  $(1, 2, \dots, 180)$  and periodic boundaries are enforced.

### 2. Conformation generation

To initiate the generation of a new protein conformer, a set of torsion angles of the first residue along with its protein sequence is provided to the GRNN model. The model repeatedly takes the torsion angles of the current residue to generate the probability distributions from which the torsion angles of the next residue are sampled until it reaches the last residue. The torsion angles are translated to Cartesian coordinates to generate a conformer. A Lennard-Jones potential is computed using Amber14SB parameters<sup>85</sup> with a user-definable threshold to reject severe clashes at each residue iteration during the conformer building process. The building and validation of conformers are supported by a conformer generator module adapted from IDPCConformerGenerator.<sup>20</sup>

### 3. Training procedure for the generative model

Separate models are trained for the disordered states of Hst5, A $\beta$ 40, uDrkN-SH3 domain, and  $\alpha$ -Syn. The Hst5 and A $\beta$ 40 pools both contain 8000 conformers and are split into 6000 for training, 1000 for validation, and 1000 for testing. The uDrkN-SH3 pool contains 7373 conformers and is split into 6000 for training, 600 for validation, and 737 for testing. The  $\alpha$ -Syn pool contains 4903 conformers, in total, and is split into 4000 for training, 400 for validation, and 503 for testing. We use categorical cross-entropy loss,

$$L_\Theta = -\frac{1}{N} \sum_{i=1}^N \sum_{t_i} \hat{p}(t_i|t_1, t_2, \dots, t_{i-1}) \log p_\Theta(t_i|t_1, t_2, \dots, t_{i-1}), \quad (\text{A2})$$

where  $N$  represents the number of angle bins,  $\hat{p}(t_i|t_1, t_2, \dots, t_{i-1})$  represents the actual probability of a specific torsion at the  $t_i$ th step, and  $p_\Theta(t_i|t_1, t_2, \dots, t_{i-1})$  is the probability predicted by the

neural network with parameters  $\Theta$ . The model is trained using the Adam optimizer<sup>86</sup> in batches of size 100. To achieve convergence, we employed an initial learning rate of 0.0005 and reduced the learning rate by a factor of 0.8 when the loss function plateaued. The generative models are trained for 300 epochs.

### 4. DynamICE supervised learning procedure

Torsion angles unrelated to the experimental observables being optimized, if unrestrained, can lead to a noisy action space during training. Thus, only the strongly relevant model parameters are updated. This includes parameters of the torsion en(de)coder and LSTM in the residue-level recurrent unit. We restrain parameters in the torsion recurrent unit to preserve the side chain torsion correlation learned from the pre-training stage, allowing the side chain torsion distribution to shift as the backbone torsion angle changes during the training stage.

We tested JC:NOE (PRE) reward weight hyperparameters of 1:1, 1:2, and 1:4. For the A $\beta$ 40 and uDrk-SH3 DynamICE models, a JC:NOE reward weight of 1:4 yields the best result; for the Hst5 and  $\alpha$ -Syn DynamICE models, we report the result using a JC:NOE (PRE) weight of 1:2. Experimental data points are weighed equally in the loss function, except for the supplementary material, Fig. 7 model, where the short-ranged NOEs (within five residues) are dynamically weighed higher during training ( $\gamma_{i,i+<=5}^{epoch<=100} : \gamma_{rest}^{epoch<=100} = 2 : 1; \gamma_{i,i+<=5}^{epoch=100-200} : \gamma_{rest}^{epoch=100-200} = 4 : 1$ ). In each epoch, 50 molecules are sampled, and model weights are updated by taking gradient steps on the loss function using the Adam optimizer with a learning rate cap of 0.0005.

### 5. Internal-Cartesian conversion

For evaluations on the distance-based experimental data types, the conformers, which are represented by torsion angle trajectories in the generative model, need to be reconstructed in terms of Cartesian coordinates. We use the SidechainNet package<sup>59</sup> for internal to Cartesian conversion following the natural extension reference frame (NeRF) algorithm.<sup>87</sup>

## APPENDIX B: X-EISD MODEL

### 1. Bayesian framework

The X-EISD method applies a maximum likelihood estimator to formulate a log-likelihood as the degree to which a simulated ensemble is in agreement with a set of experimental data, given both the experimental and back-calculation uncertainties modeled as optimized Gaussian random variables under a Bayesian framework. X-EISD can be applied to generate an aggregated score of multiple data types, as shown in Eq. (B1),

$$\log p(X, \xi|D, I) = \log p(X|I) + \sum_{j=1}^M \log [p(d_j|X, \xi_j, I)p(\xi_j|I)] + C, \quad (\text{B1})$$

where  $X$  is a set of conformers,  $\xi$  denotes the various uncertainties,  $D$  is the experimental data, and  $I$  is any other prior information. We refer readers to Refs. 36 and 40 for more detailed descriptions of the approach.

## 2. Ensemble reweighting and characterization

For reweighting optimizations of all the protein cases, we use X-EISD as a probabilistic score in a simple direct maximization, performing 10 000 attempts to exchange one conformer with another for an ensemble with 100 starting structures and accepting the exchange if the new ensemble receives a higher X-EISD score than the previous one,

$$\text{acc}(i \rightarrow j) = X - \text{EISD}(i) > X - \text{EISD}(j). \quad (\text{B2})$$

Reweighting optimizations with each set of data type conditions (JCs and NOEs/PREs) are repeated 100 times. In addition to the data types included during the RL-like training, we validate the generated ensembles with chemical shifts (CS), smFRET  $\langle E \rangle$ , and SAXS. We use UCBSHift<sup>88</sup> for chemical shift calculations and CRY SOL software program<sup>89</sup> for SAXS intensities, and the efficiencies of the energy transfer are treated using in-house scripts as reported previously.<sup>40</sup> The preparations of experimental data and back-calculation uncertainties for the reported data types are also described previously.<sup>40</sup> The experimental J-couplings, NOEs, and chemical shifts for histatin 5 are reported in Raj *et al.*,<sup>70</sup> and we use the experimental SAXS data in an aqueous solution from Sagar *et al.*<sup>90</sup> The experimental data for A $\beta$ 40 used are reported by Ball *et al.*<sup>15,91</sup> The various experimental data types for the DrkN SH3 unfolded domain are from the previous work of the Forman-Kay and Gradi-naru group,<sup>92,93</sup> and those for  $\alpha$ -Synuclein are reported in Ferrie and Petersson<sup>19</sup> Ensemble global metrics, including radius of gyration  $R_g$ , end-to-end distance  $R_{ee}$ , and asphericity  $\delta^*$  [which measures the anisotropy of the structural ranging from 0 (sphere) to 1 (rod)], are calculated using the MDTraj package.<sup>94</sup>

## REFERENCES

- P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm," *J. Mol. Biol.* **293**, 321–331 (1999).
- H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
- J. D. Forman-Kay and T. Mittag, "From sequence and forces to structure, function, and evolution of intrinsically disordered proteins," *Structure* **21**, 1492–1499 (2013).
- A. Bhowmick, D. H. Brookes, S. R. Yost, H. J. Dyson, J. D. Forman-Kay, D. Gunter, M. Head-Gordon, G. L. Hura, V. S. Pande, D. E. Wemmer, P. E. Wright, and T. Head-Gordon, "Finding our way in the dark proteome," *J. Am. Chem. Soc.* **138**, 9730–9742 (2016).
- V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Intrinsically disordered proteins in human diseases: Introducing the D<sub>2</sub> concept," *Annu. Rev. Biophys.* **37**, 215–246 (2008).
- V. N. Uversky, V. Davé, L. M. Iakoucheva, P. Malaney, S. J. Metallo, R. R. Pathak, and A. C. Joerger, "Pathological unfoldomics of uncontrolled chaos: Intrinsically disordered proteins and human diseases," *Chem. Rev.* **114**, 6844 (2014).
- B. Tsang, I. Pritišanac, S. W. Scherer, A. M. Moses, and J. D. Forman-Kay, "Phase separation as a missing mechanism for interpretation of disease mutations," *Cell* **183**, 1742–1756 (2020).
- J. A. Toretzky and P. E. Wright, "Assemblages: Functional units formed by cellular phase separation," *J. Cell Biol.* **206**, 579–588 (2014).
- W. Borchers, A. Bremer, M. B. Borgia, and T. Mittag, "How do intrinsically disordered protein regions encode a driving force for liquid–liquid phase separation?," *Curr. Opin. Struct. Biol.* **67**, 41–50 (2021).
- P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation," *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
- Y.-H. Lin, J. D. Forman-Kay, and H. S. Chan, "Theories for sequence-dependent phase behaviors of biomolecular condensates," *Biochemistry* **57**, 2499–2508 (2018).
- K. M. Ruff, R. V. Pappu, and A. S. Holehouse, "Conformational preferences and phase behavior of intrinsically disordered low complexity sequences: Insights from multiscale simulations," *Curr. Opin. Struct. Biol.* **56**, 1–10 (2019).
- G.-N. W. Gomes, M. Krzeminski, A. Namini, E. W. Martin, T. Mittag, T. Head-Gordon, J. D. Forman-Kay, and C. C. Gradi-naru, "Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET," *J. Am. Chem. Soc.* **142**, 15697–15710 (2020).
- K. A. Ball, A. H. Phillips, P. S. Nerenberg, N. L. Fawzi, D. E. Wemmer, and T. Head-Gordon, "Homogeneous and heterogeneous tertiary structure ensembles of amyloid- $\beta$  peptides," *Biochemistry* **50**, 7612–7628 (2011).
- K. A. Ball, A. H. Phillips, D. E. Wemmer, and T. Head-Gordon, "Differences in  $\beta$ -strand populations of monomeric A $\beta$ 40 and A $\beta$ 42," *Biophys. J.* **104**, 2714–2724 (2013).
- K. A. Ball, D. E. Wemmer, and T. Head-Gordon, "Comparison of structure determination methods for intrinsically disordered amyloid- $\beta$  peptides," *J. Phys. Chem. B* **118**, 6405–6416 (2014).
- H. J. Feldman and C. W. V. Hogue, "A fast method to sample real protein conformational space," *Proteins: Struct., Funct., Bioinf.* **39**, 112–131 (2000).
- V. Ozenne, R. Schneider, M. Yao, J.-r. Huang, L. Salmon, M. Zweckstetter, M. R. Jensen, and M. Blackledge, "Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution," *J. Am. Chem. Soc.* **134**, 15138–15148 (2012).
- J. R. Ferrie and E. J. Petersson, "A unified de novo approach for predicting the structures of ordered and disordered proteins," *J. Phys. Chem. B* **124**, 5538–5548 (2020).
- J. M. C. Teixeira, Z. H. Liu, A. Namini, J. Li, R. M. Vernon, M. Krzeminski, A. A. Shamandy, O. Zhang, M. Haghighatlari, L. Yu, T. Head-Gordon, and J. D. Forman-Kay, "IDPConformerGenerator: A flexible software suite for sampling the conformational space of disordered protein states," *J. Phys. Chem. A* **126**, 5985–6003 (2022).
- J. R. Allison, P. Várnai, C. M. Dobson, and M. Vendruscolo, "Determination of the free energy landscape of  $\alpha$ -synuclein using spin label nuclear magnetic resonance measurements," *J. Am. Chem. Soc.* **131**, 18314–18326 (2009).
- W.-Y. Choy and J. D. Forman-Kay, "Calculation of ensembles of structures representing the unfolded state of an SH<sub>3</sub> domain," *J. Mol. Biol.* **308**, 1011–1032 (2001).
- J. A. Marsh, C. Neale, F. E. Jack, W.-Y. Choy, A. Y. Lee, K. A. Crowhurst, and J. D. Forman-Kay, "Improved structural characterizations of the drkN SH<sub>3</sub> domain unfolded state suggest a compact ensemble with native-like and non-native structure," *J. Mol. Biol.* **367**, 1494–1510 (2007).
- J. A. Marsh and J. D. Forman-Kay, "Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints," *J. Mol. Biol.* **391**, 359–374 (2009).
- M. Krzeminski, J. A. Marsh, C. Neale, W.-Y. Choy, and J. D. Forman-Kay, "Characterization of disordered proteins with ENSEMBLE," *Bioinformatics* **29**, 398–399 (2013).
- C. K. Fisher, A. Huang, and C. M. Stultz, "Modeling intrinsically disordered proteins with Bayesian statistics," *J. Am. Chem. Soc.* **132**, 14919–14927 (2010).
- C. K. Fisher and C. M. Stultz, "Constructing ensembles for intrinsically disordered proteins," *Curr. Opin. Struct. Biol.* **21**, 426–431 (2011).
- C. K. Fisher, O. Ullman, and C. M. Stultz, "Efficient construction of disordered protein ensembles in a Bayesian framework with optimal selection of conformations," *Pac. Symp. Biocomput.* **2012**, 82–93.
- A. Huang and C. M. Stultz, "The effect of a  $\Delta$ K280 mutation on the unfolded state of a microtubule-binding repeat in tau," *PLoS Comput. Biol.* **4**, e1000155 (2008).
- M.-K. Yoon, V. Venkatachalam, A. Huang, B.-S. Choi, C. M. Stultz, and J. J. Chou, "Residual structure within the disordered C-terminal segment of p21<sup>Waf1/Cip1/Sdi1</sup> and its implications for molecular recognition," *Protein Sci.* **18**, 337–347 (2009).
- M. R. Jensen, L. Salmon, G. Nodet, and M. Blackledge, "Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts," *J. Am. Chem. Soc.* **132**, 1270–1272 (2010).



- <sup>32</sup>R. Schneider, J.-R. Huang, M. Yao, G. Communie, V. Ozenne, L. Mollica, L. Salmon, M. R. Jensen, and M. Blackledge, "Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy," *Mol. BioSyst.* **8**, 58–68 (2012).
- <sup>33</sup>M. R. Jensen, R. W. Ruigrok, and M. Blackledge, "Describing intrinsically disordered proteins at atomic resolution by NMR," *Curr. Opin. Struct. Biol.* **23**, 426–435 (2013).
- <sup>34</sup>M. Schwalbe, V. Ozenne, S. Bibow, M. Jaremko, L. Jaremko, M. Gajda, M. R. Jensen, J. Biernat, S. Becker, E. Mandelkow, M. Zweckstetter, and M. Blackledge, "Predictive atomic resolution descriptions of intrinsically disordered hTau40 and  $\alpha$ -synuclein in solution from NMR and small angle scattering," *Structure* **22**, 238–249 (2014).
- <sup>35</sup>G. Hummer and J. Köfinger, "Bayesian ensemble refinement by replica simulations and reweighting," *J. Chem. Phys.* **143**, 243150 (2015).
- <sup>36</sup>D. H. Brookes and T. Head-Gordon, "Experimental inferential structure determination of ensembles for intrinsically disordered proteins," *J. Am. Chem. Soc.* **138**, 4530–4538 (2016).
- <sup>37</sup>M. Bonomi, C. Camilloni, A. Cavalli, and M. Vendruscolo, "Metainference: A Bayesian inference method for heterogeneous systems," *Sci. Adv.* **2**, e1501177 (2016).
- <sup>38</sup>M. Bonomi, G. T. Heller, C. Camilloni, and M. Vendruscolo, "Principles of protein structural ensemble determination," *Curr. Opin. Struct. Biol.* **42**, 106–116 (2017).
- <sup>39</sup>J. Köfinger, L. S. Stelzl, K. Reuter, C. Allande, K. Reichel, and G. Hummer, "Efficient ensemble refinement by reweighting," *J. Chem. Theory Comput.* **15**, 3390–3401 (2019).
- <sup>40</sup>J. Lincoff, M. Haghghatlati, M. Krzeminski, J. M. C. Teixeira, G.-N. W. Gomes, C. C. Gradinaru, J. D. Forman-Kay, and T. Head-Gordon, "Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states," *Commun. Chem.* **3**, 74 (2020).
- <sup>41</sup>J. Köfinger, B. Różycki, and G. Hummer, "Inferring structural ensembles of flexible and dynamic macromolecules using Bayesian, maximum entropy, and minimal-ensemble refinement methods," in *Biomolecular Simulations* (Springer, 2019), pp. 341–352.
- <sup>42</sup>M. C. Ahmed, L. K. Skaanning, A. Jussupow, E. A. Newcombe, B. B. Kragelund, C. Camilloni, A. E. Langkilde, and K. Lindorff-Larsen, "Refinement of  $\alpha$ -synuclein ensembles against SAXS data: Comparison of force fields and methods," *Front. Mol. Biosci.* **8**, 654333 (2021).
- <sup>43</sup>S. Bottaro, T. Bengtson, and K. Lindorff-Larsen, "Integrating molecular simulation and experimental data: A Bayesian/maximum entropy reweighting approach," in *Structural Bioinformatics: Methods and Protocols*, edited by Z. Gáspári (Springer, New York, 2020), pp. 219–240.
- <sup>44</sup>S. Bottaro and K. Lindorff-Larsen, "Biophysical experiments and biomolecular simulations: A perfect match?," *Science* **361**, 355 (2018).
- <sup>45</sup>M. Liu, A. K. Das, J. Lincoff, S. Sasmal, S. Y. Cheng, R. M. Vernon, J. D. Forman-Kay, and T. Head-Gordon, "Configurational entropy of folded proteins and its importance for intrinsically disordered proteins," *Int. J. Mol. Sci.* **22**, 3420 (2021).
- <sup>46</sup>S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, "Water dispersion interactions strongly influence simulated structural properties of disordered protein states," *J. Phys. Chem. B* **119**, 5113–5123 (2015).
- <sup>47</sup>J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell, Jr., "CHARMM36m: An improved force field for folded and intrinsically disordered proteins," *Nat. Methods* **14**, 71–73 (2017).
- <sup>48</sup>P. Robustelli, S. Piana, and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states," *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4758–E4766 (2018).
- <sup>49</sup>A. M. Fluit and J. J. de Pablo, "An analysis of biomolecular force fields for simulations of polyglutamine in solution," *Biophys. J.* **109**, 1009–1018 (2015).
- <sup>50</sup>J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature* **596**, 583–589 (2021).
- <sup>51</sup>M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science* **373**, 871–876 (2021).
- <sup>52</sup>M. Torrissi, G. Pollastri, and Q. Le, "Deep learning methods in protein structure prediction," *Comput. Struct. Biotechnol. J.* **18**, 1301–1310 (2020).
- <sup>53</sup>G. Masrati, M. Landau, N. Ben-Tal, A. Lupas, M. Kosloff, and J. Kosinski, "Integrative structural biology in the era of accurate structure prediction," *J. Mol. Biol.* **433**, 167127 (2021).
- <sup>54</sup>S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLoS Comput. Biol.* **13**, e1005324 (2017).
- <sup>55</sup>J. Schaarschmidt, B. Monastyrskyy, A. Kryshchafyovych, and A. M. J. Bonvin, "Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age," *Proteins: Struct., Funct., Bioinf.* **86**, 51–66 (2018).
- <sup>56</sup>B. Adhikari, J. Hou, and J. Cheng, "DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks," *Bioinformatics* **34**, 1466–1472 (2018).
- <sup>57</sup>M. AlQuraishi, "Machine learning in protein structure prediction," *Curr. Opin. Chem. Biol.* **65**, 1–8 (2021).
- <sup>58</sup>M. AlQuraishi, "End-to-end differentiable learning of protein structure," *Cell Syst.* **8**, 292–301 (2019).
- <sup>59</sup>J. E. King and D. R. Koes, "SidechainNet: An all-atom protein structure dataset for machine learning," *Proteins: Struct., Funct., Bioinf.* **89**, 1489–1496 (2021).
- <sup>60</sup>P. Hoseini, L. Zhao, and A. Shehu, "Generative deep learning for macromolecular structure and dynamics," *Curr. Opin. Struct. Biol.* **67**, 170–177 (2021).
- <sup>61</sup>X. Guo, Y. Du, S. Tadepalli, L. Zhao, and A. Shehu, "Generating tertiary protein structures via interpretable graph variational autoencoders," *Bioinf. Adv.* **1**, vbab036 (2021).
- <sup>62</sup>T. Rahman, Y. Du, L. Zhao, and A. Shehu, "Generative adversarial learning of protein tertiary structures," *Molecules* **26**, 1209 (2021).
- <sup>63</sup>M. T. Degiacomi, "Coupling molecular dynamics and deep learning to mine protein conformational space," *Structure* **27**, 1034–1040 (2019).
- <sup>64</sup>K. Moritsugu, "Multiscale enhanced sampling using machine learning," *Life* **11**, 1076 (2021).
- <sup>65</sup>K. Lindorff-Larsen and B. B. Kragelund, "On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins," *J. Mol. Biol.* **433**, 167196 (2021).
- <sup>66</sup>A. Ramanathan, H. Ma, A. Parvatikar, and S. C. Chennubhotla, "Artificial intelligence techniques for integrative structural biology of intrinsically disordered proteins," *Curr. Opin. Struct. Biol.* **66**, 216–224 (2021).
- <sup>67</sup>A. Gupta, S. Dey, A. Hicks, and H.-X. Zhou, "Artificial intelligence guided conformational mining of intrinsically disordered proteins," *Commun. Biol.* **5**, 610 (2022).
- <sup>68</sup>G. Janson, G. Valdes-Garcia, L. Heo, and M. Feig, "Direct generation of protein conformational ensembles via machine learning," *Nat. Commun.* **14**, 774 (2022).
- <sup>69</sup>M. Akdel, D. E. V. Pires, E. P. Pardo, J. Jänes, A. O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, A. Shenoy, W. Zhu, P. Kundrotas, V. R. Serra, C. H. M. Rodrigues, A. S. Dunham, D. Burke, N. Borkakoti, S. Velankar, A. Frost, J. Basquin, K. Lindorff-Larsen, A. Bateman, A. V. Kajava, A. Valencia, S. Ovchinnikov, J. Durairaj, D. B. Ascher, J. M. Thornton, N. E. Davey, A. Stein, A. Elofsson, T. I. Croll, and P. Beltrao, "A structural biology community assessment of AlphaFold2 applications," *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).
- <sup>70</sup>P. A. Raj, E. Marcus, and D. K. Sukumaran, "Structure of human salivary histatin 5 in aqueous and nonaqueous solutions," *Biopolymers* **45**, 51–67 (1998).
- <sup>71</sup>M. Goedert and M. G. Spillantini, "A century of Alzheimer's disease," *Science* **314**, 777–781 (2006).
- <sup>72</sup>F. Rahimi, A. Shanmugam, and G. Bitan, "Structure-function relationships of pre-fibrillar protein assemblies in Alzheimer's disease and related disorders," *Curr. Alzheimer Res.* **5**, 319–341 (2008).
- <sup>73</sup>M. Goedert, "Alpha-synuclein and neurodegenerative diseases," *Nat. Rev. Neurosci.* **2**, 492–501 (2001).
- <sup>74</sup>S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**, 1735–1780 (1997).

- <sup>75</sup>R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (The MIT Press, 2018).
- <sup>76</sup>M. Karplus, "Vicinal proton coupling in nuclear magnetic resonance," *J. Am. Chem. Soc.* **85**, 2870–2871 (1963).
- <sup>77</sup>G. W. Vuister, F. Delaglio, and A. Bax, "The use of  $^1J_{\text{CaHa}}$  coupling constants as a probe for protein backbone conformation," *J. Biomol. NMR* **3**, 67–80 (1993).
- <sup>78</sup>E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," [arXiv:1611.01144](https://arxiv.org/abs/1611.01144) (2016).
- <sup>79</sup>G. Tesei, J. M. Martins, M. B. A. Kunze, Y. Wang, R. Crehuet, and K. Lindorff-Larsen, "DEER-PREdict: Software for efficient calculation of spin-labeling EPR and NMR data from conformational ensembles," *PLoS Comput. Biol.* **17**, e1008551 (2021).
- <sup>80</sup>J. Iwahara, C. Tang, and G. Marius Clore, "Practical aspects of  $^1\text{H}$  transverse paramagnetic relaxation enhancement measurements on macromolecules," *J. Magn. Reson.* **184**, 185–195 (2007).
- <sup>81</sup>F. N. Newby, A. De Simone, M. Yagi-Utsumi, X. Salvatella, C. M. Dobson, and M. Vendruscolo, "Structure-free validation of residual dipolar coupling and paramagnetic relaxation enhancement measurements of disordered proteins," *Biochemistry* **54**, 6876–6886 (2015).
- <sup>82</sup>S. Sasmal, J. Lincoff, and T. Head-Gordon, "Effect of a paramagnetic spin label on the intrinsically disordered peptide ensemble of amyloid- $\beta$ ," *Biophys. J.* **113**, 1002–1011 (2017).
- <sup>83</sup>J. Roche, Y. Shen, J. H. Lee, J. Ying, and A. Bax, "Monomeric  $\text{A}\beta^{1-40}$  and  $\text{A}\beta^{1-42}$  peptides in solution adopt very similar ramachandran map distributions that closely resemble random coil," *Biochemistry* **55**, 762–775 (2016).
- <sup>84</sup>S. Vivekanandan, J. R. Brender, S. Y. Lee, and A. Ramamoorthy, "A partially folded structure of amyloid-beta (1–40) in an aqueous environment," *Biochem. Biophys. Res. Commun.* **411**, 312–316 (2011).
- <sup>85</sup>V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," *Proteins: Struct., Funct., Bioinf.* **65**, 712–725 (2006).
- <sup>86</sup>D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014).
- <sup>87</sup>M. AlQuraishi, "Parallelized natural extension reference frame: Parallelized conversion from internal to Cartesian coordinates," *J. Comput. Chem.* **40**, 885–892 (2019).
- <sup>88</sup>J. Li, K. C. Bennett, Y. Liu, M. V. Martin, and T. Head-Gordon, "Accurate prediction of chemical shifts for aqueous protein structure on 'real world' data," *Chem. Sci.* **11**, 3180–3191 (2020).
- <sup>89</sup>D. Svergun, C. Barberato, and M. H. J. Koch, "CRY SOL—A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates," *J. Appl. Crystallogr.* **28**, 768–773 (1995).
- <sup>90</sup>A. Sagar, C. M. Jeffries, M. V. Petoukhov, D. I. Svergun, and P. Bernadó, "Comment on the optimal parameters to derive intrinsically disordered protein conformational ensembles from small-angle x-ray scattering data using the ensemble optimization method," *J. Chem. Theory Comput.* **17**, 2014–2021 (2021).
- <sup>91</sup>F. Meng, M. M. J. Bellaiche, J.-Y. Kim, G. H. Zerze, R. B. Best, and H. S. Chung, "Highly disordered amyloid- $\beta$  monomer probed by single-molecule fret and MD simulation," *Biophys. J.* **114**, 870–884 (2018).
- <sup>92</sup>W.-Y. Choy, F. A. A. Mulder, K. A. Crowhurst, D. R. Muhandiram, I. S. Millett, S. Doniach, J. D. Forman-Kay, and L. E. Kay, "Distribution of molecular size within an unfolded state ensemble using small-angle X-ray scattering and pulse field gradient NMR techniques," *J. Mol. Biol.* **316**, 101–112 (2002).
- <sup>93</sup>A. Mazouchi, Z. Zhang, A. Bahram, G.-N. Gomes, H. Lin, J. Song, H. S. Chan, J. D. Forman-Kay, and C. C. Gradinaru, "Conformations of a metastable SH3 domain characterized by smFRET and an excluded-volume polymer model," *Biophys. J.* **110**, 1510–1522 (2016).
- <sup>94</sup>R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "MDTraj: A modern open library for the analysis of molecular dynamics trajectories," *Biophys. J.* **109**, 1528–1532 (2015).