JCTC Journal of Chemical Theory and Computation

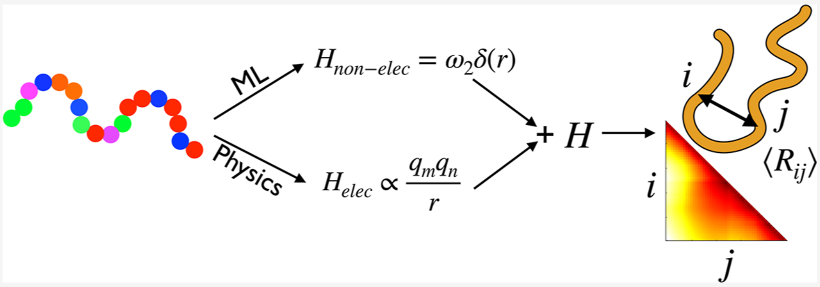
pubs.acs.org/JCTC Article

Physics-Based Machine Learning Trains Hamiltonians and Decodes the Sequence–Conformation Relation in the Disordered Proteome

Lilianna Houston,[§] Michael Phillips,[§] Andrew Torres, Kari Gaalswyk, and Kingshuk Ghosh*

Cite This: https://doi.org/10.1021/acs.jctc.4c01114 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information


ML
$H_{non-elec} = \omega_2\delta(r)$
Physics
$H_{elec} \propto \frac{q_m q_n}{r}$
$+ H \rightarrow$
i
j
$\langle R_{ij} \rangle$
i
j

ABSTRACT: Intrinsically disordered proteins and regions (IDPs) are involved in vital biological processes. To understand the IDP function, often controlled by conformation, we need to find the link between sequence and conformation. We decode this link by integrating theory, simulation, and machine learning (ML) where sequence-dependent electrostatics is modeled analytically while nonelectrostatic interaction is extracted from simulations for many sequences and subsequently trained using ML. The resulting Hamiltonian, combining physics-based electrostatics and machine-learned nonelectrostatics, accurately predicts sequence-specific global and local measures of conformations beyond the original observable used from the simulation. This is in contrast to traditional ML approaches that train and predict a specific observable, not a Hamiltonian. Our formalism reproduces experimental measurements, predicts multiple conformational features directly from sequence with high throughput that will give insights into IDP design and evolution, and illustrates the broad utility of using physics-based ML to train unknown parts of a Hamiltonian, rather than a specific observable, in combination with known physics.

1. INTRODUCTION

Machine learning (ML) can decipher complex patterns in data that are not typically amenable to closed-form mathematical equations. Problems in physical sciences are rife with such data sets and are routinely subjected to ML to ultimately make predictions. However, numerous problems are governed in part by known quantitative laws of physics, which are analytically tractable. In such a case, do we ignore that knowledge or leverage it to our advantage? This intriguing possibility presents itself in an important biophysical problem when modeling conformations of intrinsically disordered proteins and regions (collectively termed IDPs). Conformations of IDPs depend on both electrostatics and nonelectrostatic interactions. While physics of electrostatics is analytically tractable, nonelectrostatics is not amenable to closed-form mathematical relation. Hence, the approach we take builds and applies a ML model only to address the intractable nonelectrostatic contribution while preserving and leveraging analytically tractable electrostatic contribution. We showcase this hybrid usage of physics-based analytical theory and ML to model the IDP conformation that will be relevant not only in IDP biophysics but also in general problems of physical sciences.

IDPs participate in vital biological functions including transcriptional regulation, cellular differentiation, and the formation of membraneless organelles, to name a few. A growing body of data suggests that features of IDP sequence and conformation, which in turn depend on sequence,[1−6] provide insights into function.[7−17] Consistent with this, IDPs can modulate their conformation and function by biological regulators such as mutations and post-translational modifications (e.g., phosphorylation) that alter their sequence. Furthermore, IDPs can sensitively regulate their conformation by responding to environmental conditions such as changes in ionic strength, pH, or crowders (chemical regulators).[6,18−21] To advance IDP biophysics further, we need to decode the sequence–conformation relation and its response to chemical and biological regulators to ultimately decipher the sequence–conformation–function link.

Received: August 23, 2024
Revised: October 23, 2024
Accepted: October 24, 2024

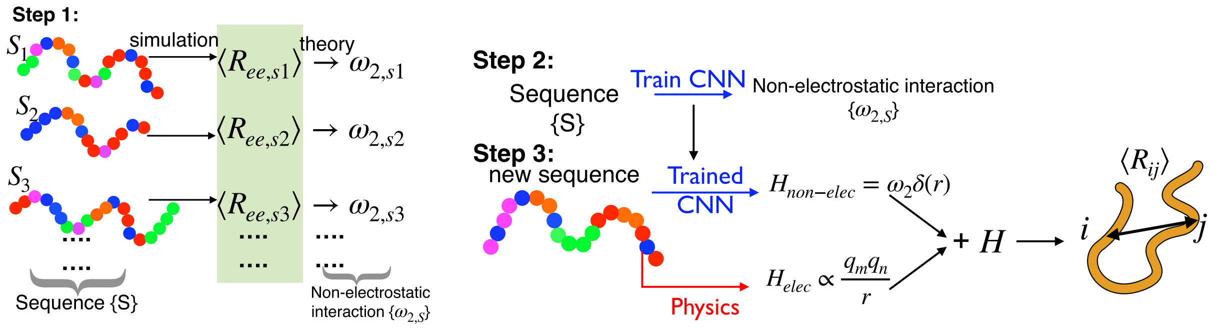

Step 1:
S_1
simulation
$\langle R_{ee,s1}\rangle$
theory
$\rightarrow \omega_{2,s1}$
S_2
$\langle R_{ee,s2}\rangle \rightarrow \omega_{2,s2}$
S_3
$\langle R_{ee,s3}\rangle \rightarrow \omega_{2,s3}$
Sequence {S}
Non-electrostatic interaction $\{\omega_{2,S}\}$
Step 2:
Sequence {S}
Train CNN
Non-electrostatic interaction $\{\omega_{2,S}\}$
Step 3:
new sequence
Trained CNN
$H_{non-elec} = \omega_2\delta(r)$
Physics
$H_{elec} \propto \frac{q_m q_n}{r}$
$+ H$
$\langle R_{ij}\rangle$
i
j

Figure 1. PML of IDP trains Hamiltonian and not a specific observable. (Left) Simulated values of end-to-end distance ($\langle R_{ee}\rangle$) for different sequences (S1, S2, S3...) are used to infer sequence-dependent nonelectrostatic interaction parameters (ω_2) from a physics-based theory in step 1. (Right) These sequence-dependent interaction parameters are trained with CNN in step 2. In step 3, for a new sequence, nonelectrostatic interaction Hamiltonian ($H_{\text{non-elec}}$) parametrized by ω_2 is predicted from the trained CNN model and combined with physics-based analytically tractable electrostatics (H_{elec}) to construct the overall Hamiltonian (H) from which several observables (such as distance $\langle R_{ij}\rangle$ between two residues i and j beyond $\langle R_{ee}\rangle$) are predicted. For the purpose of this figure, $\sqrt{\langle R_{i,j}^2\rangle}$ is denoted by $\langle R_{i,j}\rangle$ to avoid cumbersome notation.

IDPs are naturally abundant in the human proteome, and there is a need to understand their evolution across multiple proteomes. Advances in synthetic biology also require the design of IDPs with specific conformational features. Recognizing the demand from both IDP evolution and design that requires high-throughput modeling, it is timely to establish a mathematical link between IDP sequence and conformation. Different experimental tools have been developed to probe IDP configuration, from single-molecule FRET,[19,22,23] NMR,[24,25] to SAXS.[26,27] However, these specialized experiments on IDPs have yet to reach the necessary throughput. All-atom molecular dynamic simulations, despite their tremendous success in modeling folded proteins, are limited in IDPs due to their computational cost, sampling challenges,[28,29] and inaccuracies in force fields. More recently, coarse-grain simulations emerged as an alternate approach due to their computational efficiency and ability to accurately reproduce experimentally measured chain dimensions.[17,30−32] Coarse-grain simulations, despite their impressive throughput, are computationally costly for long chains. Simulation throughput is also limiting when designing an IDP with targeted conformational properties. Mutagenesis-based design often requires enumerating all possible mutations[6] and their effect on the conformation, causing a combinatorial explosion. ML or mathematical models based on theoretical physics can mitigate these challenges. Furthermore, mathematical models can give valuable insights and generate new hypotheses to be subsequently tested by detailed experiment and/or simulation.

ML models such as AlphaFold,[33] used in modeling folded proteins with unique structure, are inapplicable to IDPs[34] due to their ensemble nature and paucity of data. To overcome the lack of direct experimental data, coarse-grain simulations are used to generate ensemble average properties for a large set of sequences to train ML models.[17,32,35] These approaches train a large data set of a *specific observable* generated from the simulation of diverse sequences to predict the *same observable* for a new sequence. A separate approach has used machine learning to directly train the ensemble and its change upon mutation.[36] Naturally, these training schemes are limited to the prediction of the specific observable and/or an ensemble under a given environmental condition.

We provide a novel approach integrating physics-based theory, simulation, and ML that is not limited to one specific observable used to train the model and is versatile in its prediction. Our proposed formalism begins with a sequence-dependent analytically tractable Hamiltonian from which several ensemble average conformational properties can be derived with an approximate closed-form relation. These analytical models are scalable and capable of handling the aforementioned combinatorial explosion faced in IDP design and generating new hypotheses. A challenge, however, is to model sequence-dependent nonelectrostatic interaction parameters of the Hamiltonian. This parameter for a specific sequence is not known a priori. Sequence-dependent electrostatics, on the other hand, can be directly calculated using Coulomb's law and its screening based on laws of ionic equilibrium, amenable to analytical treatment using tools from theoretical polymer physics.[4,37] An appealing feature of our theory is its ability to infer interaction parameters of the Hamiltonian for a sequence if values of a specific observable is known. For example, the ensemble average end-to-end distance may be known from a simulation or experiment. Given this knowledge, the inference of the unknown parameter is straightforward with simple algebraic manipulation at no computational cost or optimization. We leverage this special feature of our formalism to infer sequence-dependent nonelectrostatic parameters from simulated values of a specific observable for a large set of sequences (see Figure 1 left panel). For this purpose, we utilize two recent large-scale simulation studies.[17,30] Since the underlying mapping between these nonelectrostatic parameters and their parent sequences is difficult to decipher analytically, we use ML to train these inferred parameters for a large set of sequences. For a new sequence, we then predict the nonelectrostatic patterning from the machine-learned model. The predicted nonelectrostatic interaction is further combined with our charge-patterning theory, not machine learned, derived from physicochemical laws (see Figure 1 right panel). This is the essence of physics-based ML (PML) in the present context. The resulting Hamiltonian can accurately reproduce multiple simulated conformational properties, e.g., ensemble average distance between any two residues and not just end-to-end distance for a given sequence.

Thus, PML integrates ML with a physics-based mathematical formalism and provides a powerful platform to predict both trained and untrained observables for a new sequence going beyond traditional approaches that are limited to a specific observable. Furthermore, this proof of concept principle shows the importance of training a Hamiltonian directly from experimental data, when such data of the IDP conformation becomes available at a large scale, even if measurements are limited to a specific observable. This will provide a high-throughput modeling of IDP conformation that will be useful in IDP design and applicable to predict conformations of disordered proteomes. Beyond IDP biophysics, our approach will also contribute to the emerging frontier at the interface of ML and physics.[38,39]

2. METHODS

2.1. Inferring Sequence-Dependent Nonelectrostatic Parameters from Simulations. The global dimension of an IDP can be characterized by ensemble average end-to-end distance squared ($\langle R_{ee}^2\rangle$) or a normalized swelling factor $x = \langle R_{ee}^2\rangle/(Nl^2)$, where N is the number of amino acids and $l = 5.5$ Å.[21] We have recently derived a sequence-dependent free energy ($F(x)$) as a function of the swelling factor (x) to determine the most likely x (by minimizing the free energy) for a given sequence using a model that accounts for (i) chain connectivity, (ii) electrostatic interaction and its screening by salt described by Debye−Huckel theory, (iii) residue-pair-specific nonelectrostatic interaction modeled by two-body delta function potential with unknown parameters, and (iv) a three-body mean-field repulsive interaction (with strength ω_3) as a product of two delta functions[21] needed to avoid collapse. Specifically, $F(x)$ is given by

$$\beta F(x) = \frac{3}{2}(x - \ln x) + \left(\frac{3}{2\pi}\right)^{3/2}\frac{\Omega}{x^{3/2}} + \omega_3\left(\frac{3}{2\pi}\right)^3\frac{B}{2x^3} + \frac{l_b}{b}\frac{Q}{x^{1/2}}\sqrt{\frac{6}{\pi}} \quad (1)$$

where $\beta = 1/(k_bT)$, k_b is Boltzmann constant, T is the absolute temperature, l_b is the Bjerrum length, assumed to be $l_b = 7.12$ Å $(293/T)$, and $b = 3.8$ Å is the bond length. Sequence and chain length dependence is embedded in three terms: Ω accounting for nonelectrostatic two-body interaction, B similarly capturing three-body interaction, and Q for two-body electrostatic interactions (at zero salt, see refs 21 and 40 for arbitrary salt conditions).

$$\Omega = \frac{1}{N}\sum_{m=2}^{N}\sum_{n=1}^{m-1}\omega_{m,n}(m-n)^{-1/2};$$
$$B = \frac{1}{N}\sum_{p=3}^{N}\sum_{m=2}^{p-1}\sum_{n=1}^{m-1}\frac{(p-n)}{[(p-m)(m-n)]^{3/2}}$$
$$Q = \frac{1}{N}\sum_{m=2}^{N}\sum_{n=1}^{m-1}q_mq_n(m-n)^{1/2} \quad (2)$$

Q is also known as sequence charge decoration (SCD)[4] and is calculated from sequence charge information by assigning $q = -1$ to glutamic, aspartic acids, and the C-terminal end, $q = +1$ to lysines, arginines, and the N-terminal end, and possibly $q = +0.5$ to histidines to be consistent with coarse grain simulation (see Supporting Information for more). The residue-pair-specific nonelectrostatic interaction between the two residues at m and n is $\omega_{m,n}$, related to the residue-pair-specific Lennard-Jones potential (or its variants) used in the simulation.[30] However, the mapping between the parameters used in simulation and $\omega_{m,n}$ used as a prefactor of the delta function potential in theory are unknown. We circumvent this by defining a two-body sequence-dependent effective parameter ω_2

$$\omega_2 = [\sum_{m=2}^{N}\sum_{n=1}^{m-1}\omega_{m,n}(m-n)^{-1/2}]/[\sum_{m=2}^{N}\sum_{n=1}^{m-1}(m-n)^{-1/2}] \quad (3)$$

Minimizing free energy, the equation for the most likely x becomes

$$x^{3/2}\left(1-\frac{1}{x}\right) = \left(\frac{3}{2\pi}\right)^{3/2}\omega_2(\sum_{m=2}^{N}\sum_{n=1}^{m-1}(m-n)^{-1/2})\frac{1}{Nx} + \frac{l_b}{l}\sqrt{\frac{2}{3\pi}}Q + \omega_3\left(\frac{3}{2\pi}\right)^3\frac{B}{x^{5/2}} \quad (4)$$

which can be used to predict x for a sequence if ω_2 and ω_3 are known with Q computed from the sequence. Conversely, for a given value of ω_3, the corresponding value of ω_2 for a sequence can be inferred if x is known. Ensemble average end-to-end distance obtained from coarse-grained simulation or experiment can be used to extract x for a given sequence. We use the latter approach to infer ω_2 values for different sequences using two separate large-scale coarse grain simulations performed by Zheng et al.[30] and Tesei et al.[17]

2.2. Training and Predicting ω_2 with Neural Networks. The inferred ω_2 values were trained against their sequence using a convolutional neural network (CNN$_Z$ for the Zheng set and CNN$_T$ for the Tesei set) model (see Supporting Information and Figure S1 for details of the CNN model). Tesei data set was further modified by removing highly homologous proteins to ensure that proteins in the test set are significantly different from the training set (see Supporting Information for details of detecting homologous sequences). Excluding homologous proteins from the original Tesei set resulted in a modified Tesei set with 27,060 proteins. We used k-folds to predict ω_2 values for the resulting Tesei data set and the Zheng set. For each set, we randomly shuffled and split the set into ten equally sized folds. We then trained ten different models, each using a different fold for testing and validation, and the remaining eight folds for training. Compiling the results of all test sets provides predicted ω_2 values for the full data set.

2.3. Predicting Conformational Properties from the Hamiltonian. For an unknown sequence, ω_2 is first predicted from the trained CNN model and used (in eqs 1 and 3) to predict most likely x and the ensemble average end-to-end distance. However, predictions are not limited to $\langle R_{ee}^2\rangle$ only. For example, ensemble average distances ($\langle R_{ij}^2\rangle$) between any two amino acid residues i and j, beyond just end-to-end distance, can be determined by minimizing the corresponding free energy $F(x_{ij})$

$$\beta F(x_{i,j}) = \frac{3}{2}(x_{i,j} - \ln x_{i,j}) + \left(\frac{3}{2\pi}\right)^{3/2}\frac{\text{SHDM}_{i,j}}{x_{i,j}^{3/2}} + \left(\frac{3}{2\pi}\right)^3\frac{\omega_3T_{i,j}}{2(i-j)x_{i,j}^3} + \frac{l_b}{b}\sqrt{\frac{6}{\pi}}\frac{\text{SCDM}_{i,j}}{x_{i,j}^{1/2}} \quad (5)$$

with $x_{ij}|i - j|l^2 = \langle R_{ij}^2\rangle$. The electrostatic contribution (in the zero salt limit) is calculated from sequence charge decoration matrix (SCDM); the details of which, along with its salt dependence, can be found in ref 21 and Supporting

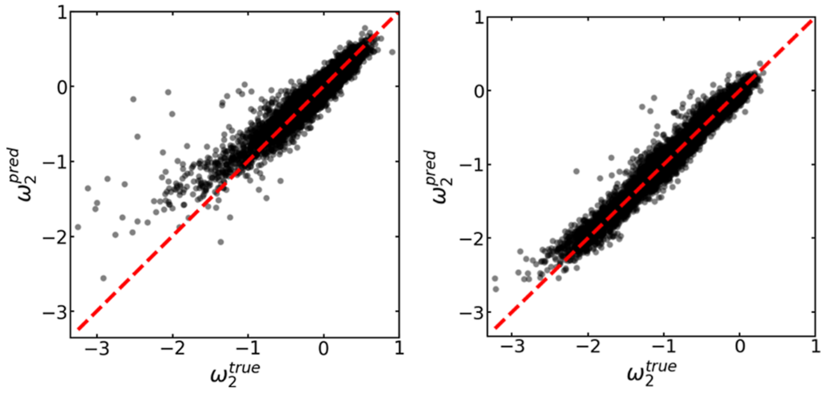


Figure 2. Nonelectrostatic effective potential can be mapped to a sequence-dependent parameter using a neural network. Sequence-dependent nonelectrostatic interaction strength ω_2 predicted from the neural network model well reproduces true ω_2 values extracted from two sets of simulations: (left) simulation set performed by Tesei et al. with 27,060 sequences (correlation $R = 0.94$ between the predicted and true values) and (right) simulation set performed by Zheng et al. with 15,390 sequences (correlation $R = 0.97$).

Information. The noncharge patterning contribution given by sequence hydropathy decoration matrix (SHDM) is defined as

$$\begin{aligned}\mathrm{SHDM}_{i,j} = \frac{1}{(i-j)}\Bigg[& \sum_{m=j}^{i}\sum_{n=1}^{j-1}\omega_{m,n}\frac{(m-j)^2}{(m-n)^{5/2}} \\ & + \sum_{m=j+1}^{i}\sum_{n=j}^{m-1}\omega_{m,n}(m-n)^{-1/2} \\ & + \sum_{m=i+1}^{N}\sum_{n=j}^{i}\omega_{m,n}\frac{(i-n)^2}{(m-n)^{5/2}} \\ & + \sum_{m=i+1}^{N}\sum_{n=1}^{j-1}\omega_{m,n}\frac{(i-j)^2}{(m-n)^{5/2}}\Bigg]\end{aligned} \tag{6}$$

For a specific sequence, SHDM_{ij} is estimated by replacing $\omega_{m,n}$ (in eq 6) by sequence-specific ω_2 from the CNN model. With this mean-field approximation, ω_3, and the definition of T_{ij} (see ref 21), the most likely x_{ij} and hence $\langle R_{ij}^2\rangle$ can be predicted for a given sequence. However, an improved (still approximate) model (SHDM_I) can be envisioned beyond the mean-field model. We note that there are four interaction terms (I1, I2, I3, and I4) in eq 6 with different summation ranges, corresponding to interactions between different segments of the parent sequence. These are I1, interaction between the segment bounding (j, i) and the N-terminal dangling end between $(1, j - 1)$; I2, interaction between amino acids in the segment bounded by (j, i); I3, interaction between segment (j, i) and the C-terminal dangling end $(i + 1, N)$; and I4, interaction between $(1, j - 1)$ and $(i + 1, N)$, the two dangling ends. These different sets of interactions can be modeled using ω_2 for specific sequence segments. For I1, $\omega_{m,n}$ can be replaced by ω_2 for the sequence bounded by $(1, i)$, similarly for I3, ω_2 could be estimated for the sequence bounded by (j, N). For I2 and I4, ω_2 can be estimated for the sequence bounded by (j, i) and the parent sequence $(1, N)$, respectively.

3. RESULTS

3.1. Simulated Hamiltonian Can Be Mapped to an Analytically Tractable Hamiltonian. We built two separate neural network models, $\mathrm{CNN_Z}$ and $\mathrm{CNN_T}$, to train sequence-dependent nonelectrostatic interaction parameters (ω_2) inferred from coarse-grain simulations of Zheng et al. and Tesei et al.,[17,30] respectively. Zheng and colleagues simulated 15,390 designed protein sequences and provided a linear regression model using two patterning metrics to describe ensemble average radius of gyration and scaling exponent.[30] More recently, Tesei et al. curated and simulated 28,058 disordered regions of the human proteome using a similar but different coarse-grain force field. Both $\mathrm{CNN_Z}$ and $\mathrm{CNN_T}$ models accurately reproduce sequence-specific ω_2 values in their respective test sets (Figure 2). We conclude that nonelectrostatic potentials used in the simulation can be effectively mapped to an analytically tractable delta function potential with a sequence-specific strength (ω_2). Furthermore, these effective parameters can be predicted by $\mathrm{CNN_Z}$ and $\mathrm{CNN_T}$ for any arbitrary sequence. The optimal value of the three-body interaction ($\omega_3 = 0.2$) was determined such that the predicted and true values (inferred from the simulation) of ω_2 correlate best for both Zheng and Tesei data sets simultaneously (see Supporting Information Figure S2 for details). While the CNN model accurately predicts ω_2 for most sequences, there are also a few outlier sequences with notable differences between the predicted and true ω_2 values. We compared the average fraction of 20 amino acids between the outlier and nonoutlier sets and found that outliers tend to be enriched in charges. Furthermore, outlier sequences tend to have a higher net charge compared to the nonoutlier sequences (see Supporting Information Figures S3−S5).

3.2. Nonelectrostatic Patterning Can Be Predicted Directly from the Sequence. Sequence-specific ω_2 predicted from the trained model yields a measure of nonelectrostatic patterning, analogous to charge patterning given by SCD metric.[4] We thus define a machine-learned sequence hydropathy decoration metric ($\mathrm{SHD_{ML}}$) as

$$\mathrm{SHD_{ML}} = \omega_2\Big[\sum_{m=2}^{N}\sum_{n=1}^{m-1}(m-n)^{-1/2}\Big]/N \approx \frac{4}{3}\omega_2 N^{1/2} \tag{7}$$

where in the second equality, we approximated the summation by an integral when N is sufficiently large. The primary nomenclature SHD was borrowed from the work of Zheng et al.[30] In the present work, we use ML using $\mathrm{CNN_T}$ or $\mathrm{CNN_Z}$ to compute $\mathrm{SHD_{ML}}$ so it can be used directly to predict size. Furthermore, $\mathrm{SHD_{ML}}$ (or ω_2 removing the chain length dependence) can be used in addition to the traditional

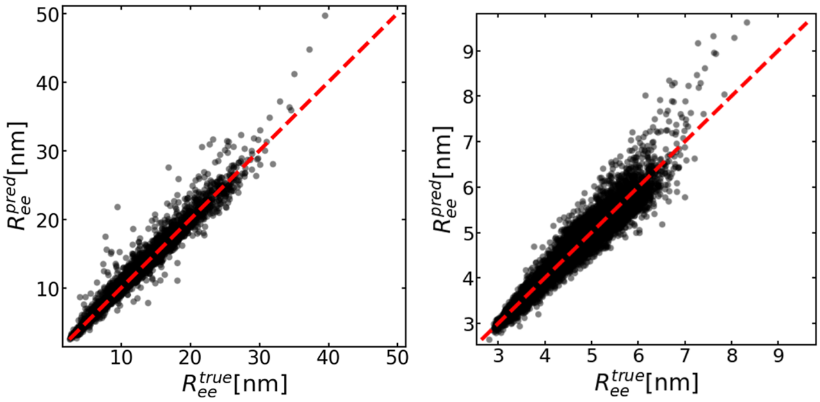


Figure 3. PML can predict chain dimensions for large sets of proteins consistent with simulation. Predicted end-to-end distance from PML reproduces simulation results for two sets of simulations: (left) Tesei set of 27,060 (correlation R = 0.99) and (right) Zheng set of 15,390 (correlation R = 0.97). Theoretical predictions were made with the conditions used in the simulation, i.e., charge assignments, ionic strength, temperature, etc., and ω_3 = 0.2.

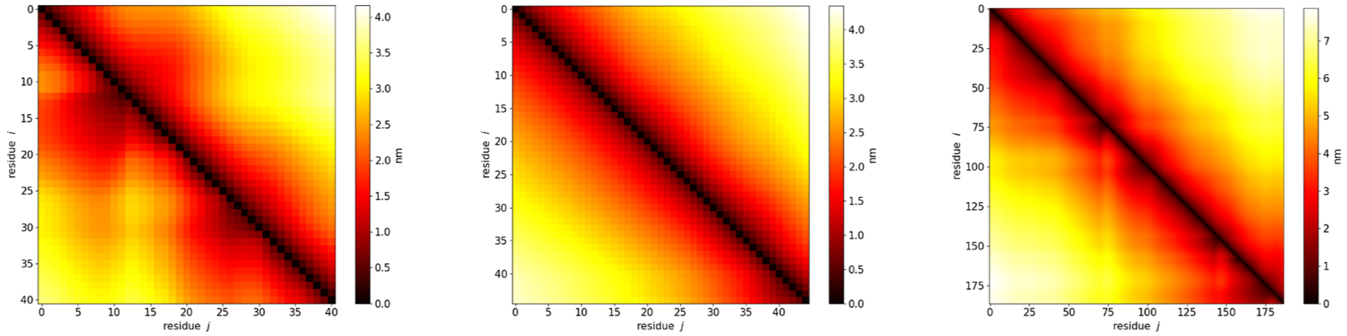


Figure 4. PML can predict local and global measures of chain conformation beyond end-to-end distance. Ensemble average inter-residue distances from Tesei set show close agreement between PML predicted (bottom triangle) and simulated maps (top triangle). Shown are the three most compact sequences out of the representative subset of 25 randomly drawn; see Supporting Information for 22 others and for correlations across all residue pairings for each sequence. Different colors show distances (in nanometers) between the two residues i and j shown in the two axes of the triangles. From left to right, sequences are $A0A1B0GVY4_{31-71}$, $A0A1W2PP97_{1-45}$, and $A6NHP3_{1-187}$.

patterning metric (hydropathy patterning[30] or charge patterning[3,4,41]) to build ML models for functions that are beginning to emerge.[17,42]

3.3. PML-Derived Hamiltonian Can Predict Size across the Disordered Proteome.

SCD[4] and similar metrics[3] so far have been used to qualitatively determine the relative trend of expansion/compaction between sequences of different charge content and patterning.[6,12,43] However, our theoretical formalism (eq 1) with machine-learned ω_2 values can predict IDP size (not just trend) for a sequence, accounting for both electrostatic and nonelectrostatic patterning. We reiterate that this is different from traditional ML approaches that train protein size and not the parameter of the Hamiltonian. We leverage this unique feature of our formalism and predict ensemble average end-to-end distance ($\langle R_{ee}^2 \rangle$) for a given sequence. Predicted chain dimensions compare well with the simulated values for the Zheng and Tesei data sets (see Figure 3). CNN$_Z$ was used to predict ω_2 values when comparing chain dimensions from Zheng simulations for 15,390 designed sequences. CNN$_T$ was used for the Tesei set to predict chain dimensions of 27,060 disordered proteins of the entire human proteome. PML-predicted R_{ee} values agree well with the true values for most sequences, with some outliers. We notice that some of these outliers were outliers in the predicted ω_2 values as well (see Supporting Information Figure S6). However, PML also predicts the dimension well for a significant fraction of the sequences identified as outliers when predicting ω_2. This finding suggests that modest inaccuracy in predicting ω_2 is masked in the prediction of the chain dimension, possibly due to the dominant role of electrostatics on the chain dimension for these sequences. To further investigate the nature of the outlier sequences, we divided the original set in seven bins based on the net charge of the proteins. We computed the correlation coefficient between the predicted and true values of R_{ee} within each bin (see Figure S7 in Supporting Information). While each bin has high correlation coefficient, there is a slight drop in correlation with increased net charge. While dimensions of these proteins were already predicted from these large-scale simulations, our goal here is not to derive further biological insights by reproducing these results. Instead, we want to highlight that our formalism provides a fast and accurate model to predict ensemble average end-to-end distances directly from the sequence. The end-to-end distance for the entire proteome could be predicted in about 16 h. Thus, if new sets of sequences—either for design or for another proteome—are constructed, our methodology will provide a quick prediction of conformations for further analysis without having to run lengthy simulations at first. From these preliminary predictions, new hypotheses can be generated and further tested with detailed but limited simulations afterward.

3.4. Ensemble Average Distance Maps Can Be Predicted Directly from the Sequence. Mathematical models grounded on Hamiltonian models can formally predict several ensemble average properties. Consistent with this expectation, our theoretical formalism can predict multiple ensemble average properties beyond just end-to-end distance, for example, ensemble average distance $\langle R_{ij}^2 \rangle$ between any two amino acid residues *i* and *j*. To showcase this broad predictive power of the PML approach, we have chosen 25 representative sequences from the Tesei list of the human proteome (see Supporting Information for details of the selection criteria). Predicted distance maps for each of the sequences compare well with the maps generated from the simulation. Following Tesei, the CALVADOS simulation package was used to generate simulated conformations to benchmark theoretical predictions (see Supporting Information for details of CALVADOS). The results for the three sequences with the most compact chain dimensions (quantified by *x*) are shown here for illustration (Figure 4); the remaining comparisons can be found in the Supporting Information (see Figures S8−S12 in the Supporting Information).

Simulated maps highlight regions of relative expansion and compaction even when sequence separation is fixed ($|i-j|$ is a constant), reflecting deviation from the homopolymer model. These detailed features are well reproduced in the predicted maps. Quantitative comparisons between predicted distances are also provided as correlations in the Supporting Information (see Figures S8−S12 in Supporting Information). Similar comparisons were performed for the Zheng set using 29 sequences, with reasonable agreement between simulation and prediction (see Figures S13−S17 in Supporting Information for details). For these predictions, we used CNN_Z. As an additional test of our PML prediction, we predicted distance maps of these 29 sequences from the Zheng list but using CNN_T and compared against simulations carried out using CALVADOS (different from Zheng simulation force field). Again, PML predictions agree well with these new simulations (Figures S18−S22 in Supporting Information). This additional analysis shows that simulated distance maps for the same sequence can be different depending on the force field. Not surprisingly, PML predictions made using different models of ω_2 (i.e., CNN_T and CNN_Z) are also different. Finally, we notice that specific sequences (13,550, 14,175, 14,782, and 15,339 in the Zheng list) can have noticeable differences (Pearson correlation below 0.95) between the predicted and simulated (using Zheng force field) distance maps. We hypothesized that the discrepancy is possibly due to the mean-field assumption $\omega_{m,n} \approx \omega_2$ used to estimate SHDM. We used the improved model SHDM_I to predict these distance maps for these specific sequences from the Zheng list. SHDM_I yields improved prediction of the distance maps (see Supporting Information Figure S23). However, it is important to note that SHDM_I is still an approximate model and further improvement would require building models that retain amino acid pair-specific interactions $\omega_{m,n}$. This would be needed when modeling an IDP without any charge residues where variations in the distance map will purely arise from nonelectrostatics, not accounted for in our mean-field model ($\omega_{m,n} \approx \omega_2$) at present.

3.5. PML Can Reproduce Experimental Data. How well do PML predictions compare against the experimental measurement of protein size? To answer this, we curated 64 protein sequences (including wild type and mutants) for which radius of gyration values were reported using SAXS and/or FRET.[30,31] PML predictions were made using the CNN_T model (Figure 5). Furthermore, we accounted for experimental salt, pH, and temperature conditions varying across experiments (see Table S1 in Supporting Information for different sequences and conditions). These conditions directly impact the electrostatic contributions in our model but ignore their impact on ω_2. CNN_T-based prediction is chosen here for comparison for two reasons. First, Tesei simulation was carried out with the force field that was originally adjusted to reproduce the majority of the proteins, although not all, in the benchmark set. No such adjustments were made for the force field used by Zheng. Furthermore, CNN_T was trained against sequences from human proteome that are expected to well represent the benchmark set consisting of natural sequences or their mutations. However, we ensured that none of these proteins in the benchmark set were used in the original Tesei set. In contrast, CNN_Z was trained against protein sequences that were designed and not naturally occurring. These could be the reasons why CNN_T performs better than CNN_Z (see Figure S24 in Supporting Information). We also notice minor deviations between PML prediction and experimental measurement, even when using CNN_T. The discrepancy could arise from two additional sources beyond training inaccuracies. First, the solution dependence of ω_2 is not included in the current model. Furthermore, we have estimated the radius of gyration from the ensemble average end-to-end distance by using a simple conversion based on homopolymer theory. Nevertheless, the close agreement between prediction and experiment (Figure 5) shows that PML with CNN_T offers

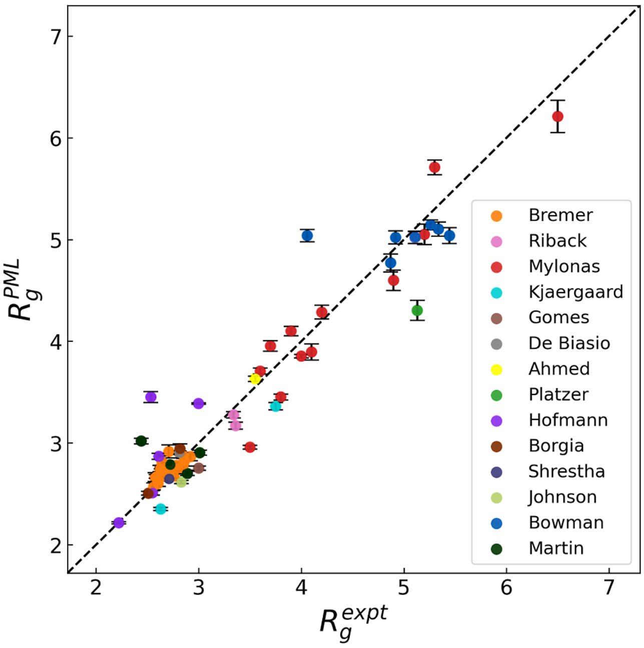


Figure 5. PML well reproduces experimental measurement of radius of gyration (R_g) across 64 different protein sequences. PML prediction was made assuming $\omega_3 = 0.2$ and using CNN_T to estimate ω_2. R_g^{PML} was estimated from $\sqrt{(\langle R_{ee}^2 \rangle/6)}$. We trained ten models, each with different training, validation, and test set. ω_2 and subsequently R_g values were calculated from these models and averaged to yield R_g^{PML}. Error bars are standard deviation of the ten R_g values. Different colors denote data reported by different groups; details of which can be found in Supporting Information.

a reliable and fast tool to predict protein size directly from the sequence.

4. DISCUSSION

We present a Hamiltonian-based mathematical formalism to determine conformational properties of IDPs accounting for sequence-specific charge and noncharge interactions simultaneously. While the effect of charge decoration can be analytically computed, noncharge patterning is difficult to describe mathematically. We overcome this difficulty by using the formalism in reverse: with the knowledge of the protein size such as ensemble average end-to-end distance, we infer the sequence-specific noncharge interaction parameter. In the absence of large-scale experimental data, we determine the protein size from recently performed high-throughput coarse-grain simulations originally benchmarked against experimental data. These simulated values serve as a proxy for experimental data and are used to extract noncharge parameters for a large set of sequences. Next, we trained these inferred sequence-dependent interaction parameters using a neural network to predict their values for a new sequence outside of the training set. The predicted values were combined with electrostatic contributions to reproduce a simulated observable (ensemble average end-to-end distance), demonstrating our ability to predict the ensemble average end-to-end distance directly from a sequence in an accurate and efficient manner. We further benchmarked our approach against pre-existing experimental data on multiple (64) protein sequences including wild type and mutants. Critically, our formalism, due to its reliance on Hamiltonian, can predict observables beyond end-to-end distance such as the ensemble average distance between any two arbitrary residues. These predictions provide a detailed knowledge of the sequence-dependent distance map, capturing both local and global measures of chain dimension. As a result, it can now predict conformational differences due to differences in both charge and noncharge patterning.[44,45] These broad-ranging predictions reiterate the power of an integrated approach that (i) deploys ML to selectively train part of the Hamiltonian that is not analytically tractable and (ii) combines with interactions that are known from the laws of physics and are analytically tractable. This integrated approach has been termed PML.

Although our theory has been applied to describe the equilibrium properties of a single IDP, it can extend far beyond. For example, it can serve to develop a theory of IDP dynamics under dilute conditions or even IDP solution when a single-chain Hamiltonian is extended to a multichain Hamiltonian. Modeling IDP solution will have implications in building a sequence-specific quantitative theory of liquid—liquid phase separation (LLPS). Analytical theories of LLPS are currently limited to predicting effects of charge patterning only due to their inability to quantify and model noncharge patterning. Consequently, qualitative trends are predicted when comparing two sequences with different charge patterning neglecting variations in noncharge interactions or, at best, phase diagrams are fitted.[46,47] Several models of noncharge interactions are being investigated to account for their effect on the phase diagrams.[48] A multichain Hamiltonian accounting for noncharge patterning, building on the PML formalism presented here, will be able to make quantitative predictions avoiding these approximations. It can also model differences in phase separation propensity between two sequences that only differ in noncharge patterning.[44] However, there are some limitations to our approach. The overall success of the proposed formalism to model IDP conformation, dynamics, and phase behavior will only be as accurate as the force fields from which interaction parameters were derived and trained. If better force fields for IDPs are generated, new simulations should be performed, and the neural network model should be retrained. The role of force field is apparent when comparing distance maps of 29 sequences (compare Figures S13−S17 and S18−S22 in the Supporting Information) generated from two different PMLs trained with different force fields used in Zheng and Tesei simulation. This is also evident in our comparison of the two models against known experimental data (see Figures 5 and S24 in Supporting Information). We have also ignored the presence of folded domains that may surround intrinsically disordered regions (IDRs) and alter the conformation due to additional interactions arising between IDR and the folded domain. However, inferred ω_2 would still serve as a measure of intrachain nonelectrostatic interaction. Similarly, these parameters can be further augmented to model salting out effect,[49] not included in the present model.

In summary, our approach makes four important points. First, it forms a bridge between simulation and theory. Specifically, it maps potentials used in coarse-grain simulations to effective interactions used in analytical theory. The ability to map these potentials allows us to predict numerous observables of multiple sequences without having to run new simulations. This would be critical when high-throughput calculation is needed in the design of IDPs and understanding conformations of multiple proteomes that would be computationally prohibitive, even for coarse-grain simulations. The gain in computational efficiency would be even more drastic if the formalism is extended to multichain problem or dynamics of IDPs that are typically plagued by associated computational costs of modeling multiple chains or hydrodynamic interactions. This is not to say that PML should completely replace simulation; rather, PML formalism should be used for preliminary analysis to generate insights and hypotheses that should be further tested with detailed but limited simulation and experiment. Second, the machine-learned noncharge patterning parameter will add a new metric to understand and train other models of protein function such as LLPS. Third, our approach shows a proof of concept principle to train interaction parameters and not observables when experimental data of IDP dimensions become available at a large scale. This is in contrast with traditional ML approaches. Finally, our PML approach shows how to integrate quantitative laws of physics and ML to problems not only in IDP biophysics but also in other areas in physical sciences where part of the problem is analytically solvable.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c01114.

> Details of the theory, CNN model, selection of training and test sets, comparison between outlier and nonoutlier sets, comparison graphs between PML prediction and simulation, and details of the proteins for which experimentally measured R_g values were compared against PML prediction (PDF)

AUTHOR INFORMATION

Corresponding Author

Kingshuk Ghosh − *Department of Physics and Astronomy, University of Denver, Denver, Colorado 80210, United States; Department of Molecular and Cellular Biophysics, University of Denver, Denver, Colorado 80210, United States;* orcid.org/0000-0003-4976-0986; Email: kghosh@du.edu

Authors

Lilianna Houston − *Department of Physics and Astronomy, University of Denver, Denver, Colorado 80210, United States*

Michael Phillips − *Department of Physics and Astronomy, University of Denver, Denver, Colorado 80210, United States;* orcid.org/0000-0003-0587-6880

Andrew Torres − *Department of Physics and Astronomy, University of Denver, Denver, Colorado 80210, United States*

Kari Gaalswyk − *Department of Physics and Astronomy, University of Denver, Denver, Colorado 80210, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c01114

Author Contributions

§L.H. and M.P. contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We acknowledge support from NIH (award number R01GM138901) and NSF (award number 2213103).

REFERENCES

(1) Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16155−16160.

(2) Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C. L.; Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 8183−8188.

(3) Das, R. K.; Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 13392−13397.

(4) Sawle, L.; Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.* **2015**, *143*, 085101.

(5) Sørensen, C. S.; Kjaergaard, M. Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 23124−23131.

(6) Firouzbakht, A.; Haider, A.; Gaalswyak, K.; Alaeen, S.; Ghosh, K.; Gruebele, M. HYPK: A marginally disordered protein sensitive to charge decoration. *Proc. Natl. Acad. Sci. U.S.A.* **2024**, *121*, No. e2316408121.

(7) Beh, L. Y.; Colwell, L. J.; Francis, N. J. A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E1063−E1071.

(8) Das, R. K.; Ruff, K. M.; Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2015**, *32*, 102−112.

(9) Zarin, T.; Tsai, C. N.; Nguyen Ba, A. N.; Moses, A. M. Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E1450−E1459.

(10) Sherry, K. P.; Das, R. K.; Pappu, R. V.; Barrick, D. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E9243−E9252.

(11) Portz, B.; Lu, F.; Gibbs, E. B.; Mayfield, J. E.; Rachel Mehaffey, M.; Zhang, Y. J.; Brodbelt, J. S.; Showalter, S. A.; Gilmour, D. S. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. *Nat. Commun.* **2017**, *8*, 15231.

(12) Lin, Y. H.; Chan, H. S. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophys. J.* **2017**, *112*, 2043−2046.

(13) Cohan, M.; Ruff, K.; Pappu, R. Information theoretic measures for quantifying sequence-ensemble relationships of intrinsically disordered proteins. *Protein Eng. Des. Sel.* **2019**, *32*, 191−202.

(14) Zarin, T.; Strome, B.; Nguyen Ba, A. N.; Alberti, S.; Forman-Kay, J.; Moses, A. M. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife* **2019**, *8*, No. e46883.

(15) Huihui, J.; Ghosh, K. Intrachain interaction topology can identify functionally similar intrinsically disordered proteins. *Biophys. J.* **2021**, *120*, 1860−1868.

(16) Cohan, M.; Shinn, M.; Lalmansingh, J.; Pappu, R. Uncovering Non-random Binary Patterns Within Sequences of Intrinsically Disordered Proteins. *J. Mol. Biol.* **2022**, *434*, 167373.

(17) Tesei, G.; Trolle, A.; Jonsson, N.; Betz, J.; Knudsen, F.; Pesce, F.; Johansson, K.; Lindorff-Larsen, K. Conformational ensembles of the human intrinsically disordered proteome. *Nature* **2024**, *626*, 897−904.

(18) Dedmon, M.; Patel, C.; Young, G.; Pielak, G. FlgM gains structure in living cells. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12681−12684.

(19) Soranno, A.; Koenig, I.; Borgia, M.; Hofmann, H.; Zosel, F.; Nettels, D.; Schuler, B. Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 4874−4879.

(20) Sizemore, S. M.; Cope, S. M.; Roy, A.; Ghirlanda, G.; Vaiana, S. M. Slow internal dynamics and charge expansion in the disordered protein CGRP: A comparison with Amylin. *Biophys. J.* **2015**, *109*, 1038−1048.

(21) Ghosh, K.; Huihui, J.; Phillips, M.; Haider, A. Rules of Physical Mathematics Govern Intrinsically Disordered Proteins. *Annu. Rev. Biophys.* **2022**, *51*, 355−376.

(22) Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16155−16160.

(23) Schuler, B.; Soranno, A.; Hofmann, H.; Nettels, D. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annu. Rev. Biophys.* **2016**, *45*, 207−231.

(24) Showalter, S. A. Intrinsically Disordered Proteins: Methods for Structure and Dynamics Studies. *eMagRes* **2014**, 3, 181−190.

(25) Jensen, M. R.; Zweckstetter, M.; Huang, J.; Blackledge, M. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem. Rev.* **2014**, *114*, 6632−6660.

(26) Bernado, P.; Svergun, D. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst.* **2012**, *8*, 151−167.

(27) Kachala, M.; Valentini, E.; Svergun, D. Application of SAXS for the structural characterization of IDPs. *Adv. Exp. Med. Biol.* **2015**, *870*, 261−289.

(28) Strodel, B. Energy Landscapes of Protein Aggregation and Conformation Switching in Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433*, 167182.

(29) Gaalswyk, K.; Haider, A.; Ghosh, K. Critical Assessment of Self-Consistency Checks in the All-Atom Molecular Dynamics Simulation of Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **2023**, *19*, 2973−2984.

(30) Zheng, W.; Dignon, G.; Brown, M.; Kim, Y. C.; Mittal, J. Hydropathy patterning complements charge patterning to describe conformational preferences of disordered proteins. *J. Phys. Chem. Lett.* **2020**, *11*, 3408−3415.

(31) Tesei, G.; Schulze, T. K.; Crehuet, R.; Lindorff-Larsen, K. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2111696118.

(32) Lotthammer, J. M.; Ginell, G.; Griffith, D.; Emenecker, R.; Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods* **2024**, *21*, 465−476.

(33) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(34) Ruff, K.; Pappu, R. AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.* **2021**, *433*, 167208.

(35) Gupta, A.; Dey, S.; Hicks, A.; Zhou, H. Artificial intelligence guided conformational mining of intrinsically disordered proteins. *Commun. Biol.* **2022**, *5*, 610.

(36) Taneja, I.; Lasker, K. Machine-learning-based methods to generate conformational ensembles of disordered proteins. *Biophys. J.* **2024**, *123*, 101−113.

(37) Muthukumar, M. 50th Anniversary Perspective: A Perspective on Polyelectrolyte Solutions. *Macromolecules* **2017**, *50*, 9528−9560.

(38) Noe, F.; Tkatchenko, A.; Muller, K.-R.; Clementi, C. Machine learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.

(39) Levine, H.; Tu, Y. Machine learning meets physics: a two-way street. *Proc. Natl. Acad. Sci. U.S.A.* **2024**, *121*, No. e2403580121.

(40) Huihui, J.; Firman, T.; Ghosh, K. Modulating charge patterning and ionic strength as a strategy to induce conformational changes in intrinsically disordered proteins. *J. Chem. Phys.* **2018**, *149*, 085101.

(41) Pal, T.; Wessen, J.; Das, S.; Chan, H. Differential Effects of Sequence-Local versus Nonlocal Charge Patterns on Phase Separation and Conformational Dimensions of Polyampholytes as Model Intrinsically Disordered Proteins. *arXiv* **2024**, arXiv:2407.07226.

(42) von Bülow, S.; Tessei, G.; Lindorff-Larsen, K. Prediction of phase separation propensities of disordered proteins from sequence. *bioRxiv* **2024**, 2024.06.03.597109.

(43) McCarty, J.; Delaney, K. T.; Danielsen, S. P. O.; Fredrickson, G. H.; Shea, J. E. Complete Phase Diagram for Liquid-Liquid Phase Separation of Intrinsically Disordered Proteins. *J. Phys. Chem. Lett.* **2019**, *10*, 1644−1652.

(44) Martin, E.; Holehouse, A.; Peran, I.; Farag, M.; Incicco, J.; Bremer, A.; Grace, C. R.; Soranno, A.; Pappu, R.; Mittag, T. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **2020**, *367*, 694−699.

(45) Bowman, M.; Riback, J.; Rodriguez, A.; Guo, H.; Li, J.; Sosnick, T.; Clark, P. Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 23356−23364.

(46) Lin, Y.; Brady, J.; Chan, H. S.; Ghosh, K. A unified analytical theory of heteropolymers for sequence specific phase behaviors of polyelectrolytes and polyampholytes. *J. Chem. Phys.* **2020**, *152*, 045102.

(47) Lin, Y.-H.; Kim, T.; Das, S.; Pal, T.; Wessen, J.; Rangadurai, A.; Kay, L.; Kay, J.; Forman-Kay; Chan, H. Electrostatics of Salt-Dependent Reentrant Phase Behaviors Highlights Diverse Roles of ATP in Biomolecular Condensates. *arXiv* **2024**, arXiv:2401.04873.

(48) Das, S.; Lin, Y.-H.; Vernon, R.; Forman-Kay, J.; Chan, H. Comparative roles of charge, *π*, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 28795−28805.

(49) Wohl, S.; Jakubowski, M.; Zheng, W. Salt-dependent conformational changes of intrinsically disordered proteins. *J. Phys. Chem. Lett.* **2021**, *12*, 6684−6691.