Accelerated Missense Mutation Identification in Intrinsically Disordered Proteins using Deep Learning

Swarnadeep Seth and Aniket Bhattacharya*
Department of Physics, University of Central Florida, Orlando, Florida 32816-2385, USA
(Dated: July 7, 2024)

ABSTRACT

We use a combination of Brownian dynamics (BD) simulation studies and Deep Learning (DL) strategies for rapid identification of large structural changes caused by missense mutations in intrinsically disordered proteins (IDPs). We generate the training set for the DL algorithm from the simulation results obtained from a coarse-grained model (HPS model) of protein whose parameters have reproduced experimental gyration radii for a variety of IDPs with different charge-hydropathy ratios [Dignon *et al.*, PLOS Comp. Biology, 14, 2018, Tesei *et al.* PNAS, 118, 2021, Seth *et al.* J. Chem. Phys. **160**, 014902 (2024)]. We use one of these well parameterized models (HPS2 model) to simulate 2000 random IDPs of length ranging between 30 to 400 amino acids obtained from the DisProt database. Using the gyration radii of the simulated IDPs as the training set, we developed a multi-layer perception architecture that predicts the gyration radii of 33 known IDPs with 95% accuracy from the sequence itself. We now utilize the neural network to predict the gyration radii of every permutation of missense mutations of a particular IDP. Our approach successfully identified the most susceptible amino acid responsible for the mutations and pinpointed mutation-prone regions that induced significant alterations in the radius of gyration when compared to the wild-type IDP structures. We later validated the prediction by running BD simulations on the subset of identified mutants. The neural network yields a 10,000-fold reduction in the search space for potentially harmful mutations. Furthermore, we use SHAP (SHapley Additive exPlanations) analysis to determine the relative importance of the input features and their contributions to the predictability of our neural network. Our findings have substantial implications for understanding diseases related to IDP mutation and for the development of potential therapeutic interventions. We believe our study will promote further research on more efficient and accurate predictions of mutation effects in disordered proteins.

I. Introduction

Any change in the amino acid sequence in a protein may result in non-trivial changes in its physical properties, such as, local stiffness which will affect subchain dynamics and proximity effects, its overall conformational ensemble, and dynamics that will ultimately impact its assigned biological tasks [1]. The smallest unit of that change is a missense mutation, where an amino acid in a given protein is substituted by another [2, 3]. In particular, missense mutations in intrinsically disordered proteins (IDPs) have been implicated in playing crucial roles in many diseases, such as Alzheimer's disease (AD), Parkinson's disease (PD), Type II Diabetes and many others [4–6]. Half of the human cancer is due to the mutation in p53. A common feature of ADs and PDs is the growth of protein droplets, resulting in liquid-liquid phase separation in IDPs [7, 8].

For an IDP of length L there are approximately $L \times$ (# of Amino Acids ~ 20) $\simeq 20L$ possible missense mutations. One of the significant challenges in the field is to narrow down and identify those missense mutations that are lethal, as these mutations are key to understanding disease mechanisms and developing targeted therapies [3]. Unlike folded proteins an increasing number of IDPs listed in the databases [9, 10] have amply demonstrated violation of the structure-function dogma as these flexible proteins sample and dynamically evolve in their vast conformational ensemble making it difficult to study experimentally. High-field NMR [12, 13] and single molecule FRET [14] have been used to study IDPs. It is evident that traditional experimental and computational methods are impractical due to the structural flexibility of IDPs and the vast number of potential mutations [15, 16]. Thus, there is a need for a novel approach to rapidly identify lethal mutations, understand their structural implications, and develop potential therapies. Since missense mutations may adversely affect both the conformations and dynamically evolving structures of the IDPs, a practical and direct approach to develop fundamental understanding of the physical processes involved is to study not only naturally occurring but also artificial missense mutations experimentally and using computer simulation studies [16, 17]. The available experimental results have been instrumental in validating various coarse-grained models to study IDPs [11, 18, 19]. These CG model of IDPs have been used to study single molecule properties, liquid-liquid phase separation (LLPS) in IDPs, formation of IDP-complexes [20, 21] and successfully extracted some key aspects of the IDPs [22]. A major advantage of these CG models is that there is not only no cost for carrying out expensive experiments, the calculations are reasonably fast so that one can study

* Author to whom the correspondence should be addressed; Aniket.Bhattacharya@ucf.edu

more IDPs and available results may narrow down and direct further experiments [23]. In a previous paper, we made a detailed investigation and comparison of two CG models of IDPs (HPS1 and HPS2) [24] and calculated the mean square error (MSE) from each model. A comparison to the most recent experimental data revealed that despite use of two different hydropathy scales, both the models converge well with the experimental gyration radii values [19]. Therefore, in this article we use HPS2 model to to carry out BD simulation for a random selection of 2000 IDPs out of 7000 IDPs from DisProt database with disorder score $\geq$ 99%) whose statistical properties are shown Fig. 1). We use the gyration radii from the BD simulation as the *supervised training set* for an Artificial Neural Net (ANN) to predict missense mutation affects the physical properties of the proteins. Our studies show (i) that most often missense mutations are benign but for a few the effect of the mutation is quite radical. (ii) The ANN is several order of magnitude faster to reproduce the same results with 96% accuracy showing great potential for rapidly detecting missense mutations in IDPs implicated in human disease.

II. CG Model for the IDPs & ANN architecture

We provide essentials both for the HPS2 model with Ashbaugh-Hatch potential [19] and the neuralnet architecture used to study and predict missense mutations in IDPs. Details for the HPS2 model are given in our previous paper [19].

A. HPS2 model

In HPS2 model the amino acid residues are represented as single CG beads those interact among themselves by a modified Van der Waals interaction potential, first introduced by Ashbaugh and Hatch [25]

$$U_{AH}\left(r_{ij}\right)=\begin{cases}U_{LJ}\left(r_{ij}\right)+(1-\lambda_{ij})\epsilon_{ij}, & r_j \leq 2^{\frac{1}{6}}\sigma_{ij}\\ \lambda_{ij}U_{LJ}\left(r_{ij}\right), & \text{otherwise}\end{cases} \quad (1)$$

where U_{LJ} is the Lennard-Jones (LJ) potential,

$$U_{LJ}\left(r_{ij}\right)=4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]. \quad (2)$$

Here, $r_{ij}=|\vec{r}_i-\vec{r}_j|$ is the distance between the amino acid beads with indices i and j positioned at $\vec{r}_i$ and $\vec{r}_j$, $\epsilon_{ij}=\frac{1}{2}\left(\epsilon_i+\epsilon_j\right)$ and $\lambda_{ij}=\frac{1}{2}\left(\lambda_i+\lambda_j\right)$ are the strength of the van der Waal interaction and average hydropathy factor between any two amino acids with indices i and j.

The hydropathy factor which is the key ingredient to differentiate interactions among IDPs. The HPS2 hydropathy scale is introduced by Tesei *et al.* using the Bayesian parameter-learning procedure to optimize the hydropathy values drawn from many different hydropathy scales [18]. Here, we use HPS2 scale with optimized interaction parameter $\epsilon = 0.2$ kcal/mol to run BD simulations to obtain the gyration radii ($\langle R_g\rangle \equiv \sqrt{\langle R_g^2\rangle}$ for brevity) of several thousands of IDPs to train a neural network architecture as discussed in the next sections. In our recent publication we demonstrated the equivalence of the HPS2 scale (with $\epsilon = 0.2$) and the HPS1 scale (with $\epsilon = 0.1$) due to Dignon *et al.* those produce nearly identical results with same MSE [19].

A harmonic bond potential with spring constant k_b = 8033 kJ/(mol·nm^2) = 1920 kcal/(mol·nm^2 (Eqn. 3)

$$U_b\left(r_{ij}\right)=\frac{k_b}{2}\left(r_{ij}-r_{ij}^0\right)^2. \quad (3)$$

acts between two consecutive amino acid residues i and $j = i \pm 1$. The equilibrium bond length is set to $r_{ij}^0 = r_0 = 0.38$ nm, the distance between α Carbon atoms for the successive amino acids. This necessitates to exclude the EV interaction among the bonded neighbors.

A screened-Coulomb (SC) interaction (Eqn. 4) acts between any two charged amino acids.

$$U_{SC}\left(r_{\alpha\beta}\right)=\frac{q_\alpha q_\beta e^2}{4\pi\epsilon_0\epsilon_r}\left(\frac{e^{-\kappa r_{\alpha\beta}}}{r_{\alpha\beta}}\right) \quad (4)$$

where the indices α and β refer to the subset of the indices i and j for the charged amino acids, ϵ_r is the dielectric constant of water, and κ is the inverse Debye screening length [26]. The inverse Debye length κ^{-1} is dependent on the ionic concentration (I) and expressed as

$$\kappa^{-1}=\sqrt{8\pi l_B I N_A\times 10^{-24}} \quad (5)$$

where N_A is the Avogadro's number and l_B is the Bjerrum length,

$$l_B=\frac{e^2}{4\pi\epsilon_0\epsilon_r k_B T}. \quad (6)$$

At higher temperatures, the dielectric constant typically decreases, which affects the strength of the electrostatic interactions. If the dielectric constant does not account for temperature effects, the electrostatic interactions may be overestimated, leading to unrealistic protein conformations or interactions. Hence, we implement the temperature-dependent dielectric constant of water as expressed by the empirical relation [27]

$$\epsilon_r(T)=\frac{5321}{T}+233.76-0.9297T$$
$$+1.147\times10^{-3}T^2-8.292\times10^{-7}T^3. \quad (7)$$

B. Statistical properties of IDPs

We gathered approximately 7000 IDP FASTA sequences from the MobiDB database with the predicted disordered score exceeding 99%. The length of the sequences ranges between $N = 20-150$ with almost 2300 IDPs with of length between 40 to 60 as shown

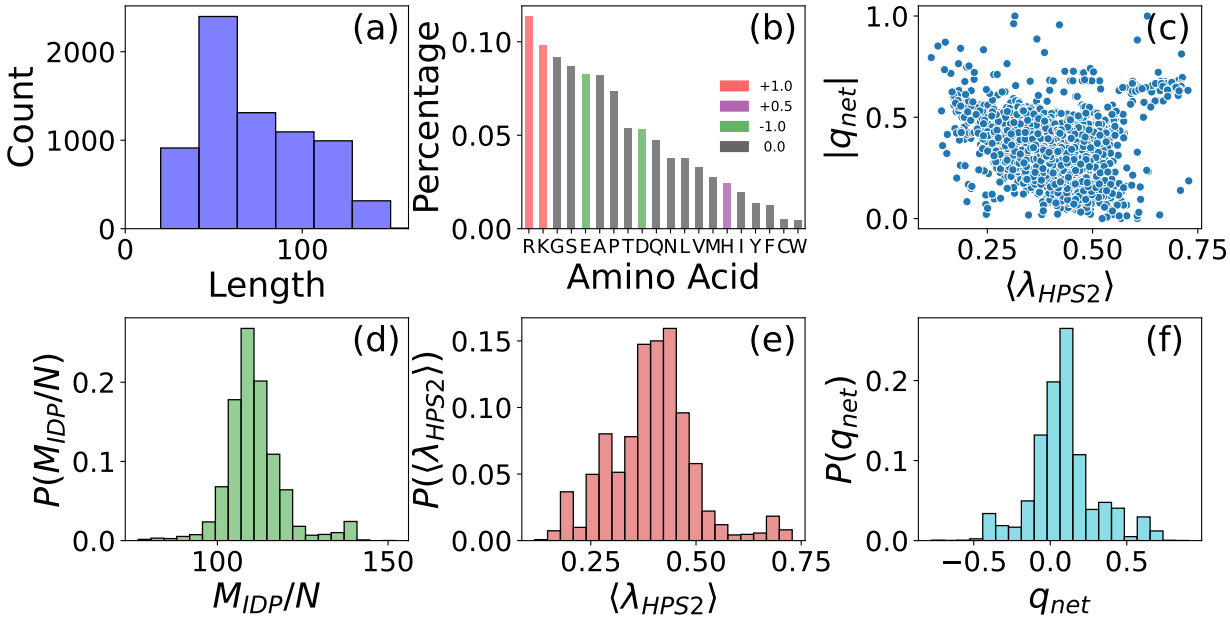

FIG. 1. Characteristics of 7000 IDPs with $\geq 99\%$ disorder predictor score obtained from the MobiDB database. (a) Length distribution of the IDPs. (b) The distribution of the propensity of occurrence of the individual amino acids. The positive, negative, neutral amino acids are colored in red, green, and grey respectively. (c) Average hydropathy $\langle\lambda_{HPS2}\rangle$ versus average net charge per residue $\langle|q_{net}|\rangle$. (d)-(f) The distributions of mass per residue M_{IDP}/N, λ_{HPS2} and q_{net} respectively.

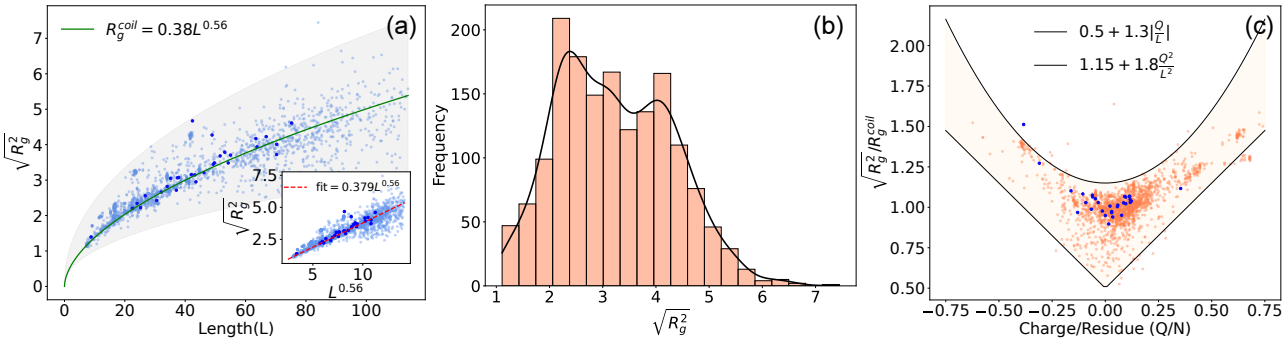

FIG. 2. (a) Radius of gyrations of IDPs as a function of chain length. The lighter blue shaded circles show the gyration radii of simulated IDPs using HPS2 model and the darker blue circles show the 33 IDPs from our previous study [19]. The green line $R_g^{coil} = 0.38L^{0.56}$ confirms the gyration radii scaling and the red dotted line in the inset plot shows gyration radii fit as a function of scaled length $L^{0.56}$. (b) The distribution of $\sqrt{\langle R_g^2\rangle}$ for the 2000 IDPs obtained from BD simulation. (c) The scaled gyration radii as a function of charge par residue N/L. The parabola $1.15 + 1.8\frac{Q^2}{N^2}$ and the line $0.5 + 1.3|\frac{Q}{N}|$ centered at zero serve as the empirical upper and lower bounds of scaled gyration radii.

in Fig. 1(a). The IDPs are well known for containing a larger fraction of disorder promoting residues such as Lysine, Glutamine, Serine, Proline, Glycine, Arginine, and Glutamic acid. From the 7000 datasets of IDPs we calculate the propensity of occurrence of the individual amino acids and their relative probability. Fig. 1(b) confirms that positively charged Arginine and Lysine are the two most abundant residues followed by negatively charged Glutamic acid and Aspartic acid on fifth and ninth locations. To make a diverse dataset of larger length IDPs up to $N = 300$, we randomly repeat the some of the sequences until we reach the desired length. We verify that on those repeated artificial IDP sequences the relative occurrences of the amino acid residues are fairly maintained. We also note that for large sampling of the IDPs the distributions for the mass, hydropathy and net charge per residue (Fig. 1(d)-(f)) follow a Gaussian distribution indicating pure randomness of the samples.

To conduct a detailed analysis of the statistical properties of IDPs, we carry out BD simulation on 2000 random sequences out of the 7000 sequences using the HPS2 hydropathy parameters to obtain the equilibrium bond length ~ 0.382 nm so that the length of an IDP containing N Amino Acid beads $L = 0.382N$ nm. The equilib-

rium radius of gyration $\langle R_g \rangle$ has been used as the key parameter to compare with experimental results. The $\langle R_g \rangle$s obtained from BD simulation on 2000 IDPs are shown in Fig. 2. A best fit produces

$$\langle R_g \rangle \approxeq 0.38 L^{0.56}, \tag{8}$$

the same scaling relation that we obtained on a smaller subset of data for 33 chosen IDPs [19]. We also show two empirical bounds on the scaled gyration radii as a function of charge per residue $\frac{Q}{N}$ shown in Fig. 2(b). The modulus of charge/residue serves as the lower bound while the parabolic function as square of charge/residue bounds the maximum range of the scaled gyration radii. These empirical ranges will help to determine the most probable value of radius of gyration of an unknown IDPs from the net charge and number of residues.

C. Neural Network

Artificial neural networks (ANN) are designed to emulate the human brain's functionality by interconnecting artificial neurons. In an ANN, several layers are included in building the network architecture, with each layer consists of multiple neurons. The input layer matches the dimension of the input data, while the output layer neurons contain the information intended for learning. The intermediate layers add complexity and expressiveness to the NN, aiding in the learning process. Each neuron in every layer is connected via a weight vector to neurons in the subsequent layer. The value in each neuron is computed as a weighted sum of the values from the previous layer's neurons, then passed through a non-linear, differentiable transformation known as the activation function. To optimize the neural network, adjustments are made to the weights assigned to all neurons, minimizing the difference between the NN's predicted output (i.e., the output layer) and the actual output.

We train a fully connected multilayer perceptron (MLP) neural network [33] for the gyration radii prediction of disordered proteins based on a diverse set of input features based on amino acid sequence only. For the input layer of the MLP neural network we choose a subset of 18 physics-based features from a larger set those quantitatively affect and determine the gyration shape of an IDP (Fig. 3). For this supervised ANN-model training, we split the 2000 IDP dataset into 80-20 training and testing dataset. We carry out MD simulation using the HPS2 model on these 2000 IDPs to obtain the radius of gyration data which serves as the true values for neural network loss calculation.

1. *Physics based features as input parameters for the ANN*

IDPs are mostly polyampholytes and polyelectrolytes. In addition, each amino acid bead has different mass and hydropathy indices. Therefore, it is expected the length of the chain and the distribution of the positive and negative charge resides as well as their hydropathy values will affect the conformational properties. We have used the following characteristics that depend on the sequence length, masses, charges and hydropathy indices of the amino acids making a given IDP. We define

$$M = \sum_{i=1}^{N} m_i, \quad \bar{m} = M/N \tag{9a}$$

$$Q_{abs} = \sum_{i=1}^{N} |q_i|, \quad q_{abs} = Q_{abs}/N \tag{9b}$$

$$Q_{net} = \sum_{i=1}^{N} q_i, \quad q_{net} = Q_{net}/N \tag{9c}$$

$$\lambda_{sum} = \sum_{i=1}^{N} \lambda_i \quad \langle \lambda \rangle = \lambda_{sum}/N \tag{9d}$$

Here m_i, q_i, and λ_i are the mass, charge, and hydropathy index of the ith amino acid. We include average mass per unit length $\bar{m}$ and average hydropathy $\langle \lambda \rangle$ in the training set of ANN and the quantities as itemized below.

Drawing analogy with the physical center of mass of a chain we define three quantities $M_{CM}, Q_{CM}, \lambda_{CM}$, with respect to the center of the AA index $(N/2)$ as follows:

$$M_{CM} = \frac{1}{M} \sum_{i}^{N} \frac{i - i_{CM}}{i_{CM}} m_i, \tag{10a}$$

$$Q_{CM} = \frac{1}{Q_{abs}} \sum_{i}^{N} \frac{i - i_{CM}}{i_{CM}} q_i. \tag{10b}$$

$$\lambda_{CM} = \frac{1}{\lambda_{sum}} \sum_{i}^{N} \frac{i - i_{CM}}{i_{CM}} \lambda_i. \tag{10c}$$

The value of i_{CM} is the midpoint of N, either $N/2$ or $(N-1)/2$ depending upon the even and odd number of residues. Under this construction, the range $M_{CM}, Q_{CM}, \lambda_{CM}$ can vary between $(-1, 1)$ that captures the asymmetry in the residue mass, charge, and hydropathy distributions regardless of the IDP length.

Likewise, we define three RMS fluctuations in mass, charge, and hydropathy

$$\delta M = \sqrt{\frac{1}{N} \sum_{i}^{N} (m_i - \bar{m})^2} \tag{11a}$$

$$\delta Q = \sqrt{\frac{1}{N} \sum_{i}^{N} (q_i - q_{abs})^2} \tag{11b}$$

$$\delta \lambda = \sqrt{\frac{1}{N} \sum_{i}^{N} (\lambda_i - \langle \lambda \rangle)^2}. \tag{11c}$$

Entropy in the sequence space plays an essential role to capture the sequence dependent properties of the IDPs. We calculate the dimensionless Shannon entropy S^{all}_{IDP} and the Shannon entropy for the charged amino acids in the sequence S^{q}_{IDP} (in units of Boltzmann constant k_B) of the FASTA sequence defined as follows

$$S^{all}_{IDP} = -\sum_i p_i \log_2(p_i). \tag{12a}$$

$$S^{q}_{IDP} = -\sum_\alpha p_\alpha \log_2(p_\alpha). \tag{12b}$$

Here $p_i = n_i/N$, where the n_i is the number of times a particular amino acid with the index $i \in 20$ occurs in the sequence. Likewise, for S^{q}_{IDP}, $p_\alpha = n_\alpha/N_q$, and the index α runs for the charged amino acids only with $\sum_\alpha n_\alpha = N_q$. The repetitive appearance of a particular amino acid will reduce the number of permutation and its contribution in dynamical heterogeneity will be captured by the corresponding Shannon entropy.

The other relevant quantities used to train the ANN are as follows:

$$f^{+1}_q = N^{+1}_q/N; \quad f^{1/2}_q = N^{1/2}_q/N \tag{13a}$$

$$f^{-1}_q = N^{-1}_q/N; \quad f^{0}_q = 1 - f^{+1}_q - f^{1/2}_q - f^{-1}_q \tag{13b}$$

$$f^* = \frac{(f^+ - f^-)^2}{f^+ + f^-}, \tag{13c}$$

$$SCD = \frac{1}{N}\sum_{i=2}^{N}\sum_{j=1}^{i-1} q_i q_j \cdot |i-j|^{0.5} \tag{13d}$$

Here N^{+1}_q, $N^{1/2}_q, and N^{-1}_q/N$ are the number of IDPs with $+1$, $+1/2$, and -1 charges, so that f^{+1}_q, $f^{1/2}_q$, f^{-1}_q are the corresponding fractions, and $f^0_q = \left(N - N^{+1}_q - N^{+1/2}_q - N^{-1}_q\right)/N = N^0_q/N$ is the fraction of neutral amino acids. The quantities $f^+ = f^{+1}_q + f^{1/2}_q$ and $f^- = f^{-1}_q$ are the net positive and negative charge per residue of an IDP. The importance of charge asymmetry parameter f^* [34, 35] and the sequence charge decoration parameter SCD parameter have been discussed in the literature [36].

Finally, the quantities

$$l^{(+1)}_{patch}, \quad l^{(+\frac{1}{2})}_{patch}, \quad l^{(-1)}_{patch} \tag{14}$$

are the contiguous charge patches consisting of unit positive, half-positive, and unit negative charges respectively (please note that the IDPs have ± 1 and $+\frac{1}{2}$ charges). $l^{(q)}_{patch} \neq 0$, for $q = \pm 1, \frac{1}{2}$ if at least two consecutive amino acids have the same charge within the charge sequence (else $l^{(q)}_{patch} = 0$). The parameters as described in Eqns. 9(a), 9(c), 9(d), 10(a)-(c), 11(a)-(c), 12(a)-(b), 13(a)-(c) along with the length L of the amino acid constitute twentyone important linearly independent and uncorrelated physics based features derived from the IDP sequence information as input vector to train the MLP network as described in the next section. We have further investigated the relative weights these quantities through SHAP in Fig. 5 [37, 38].

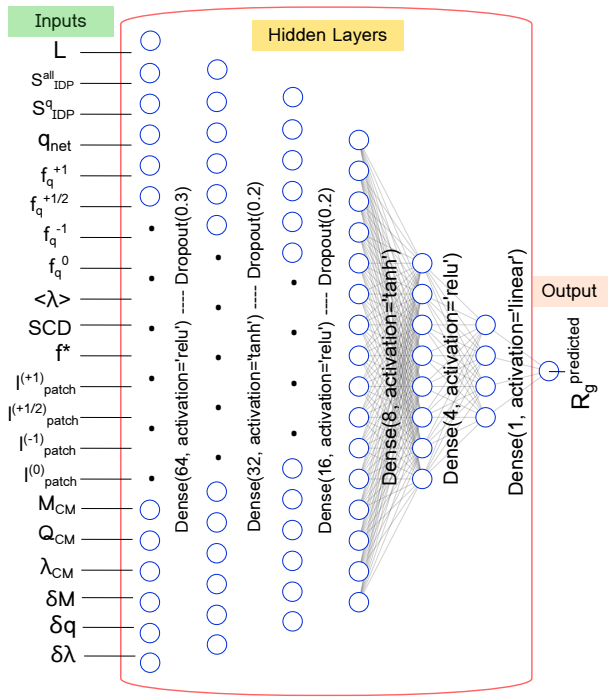

FIG. 3. The fully connected neural network architecture to predict the gyration radii of a IDP from its sequence description. The 21 input features are obtained from the IDP FASTA sequence (please see text for the explanation of the input parameters) and are fed into an artificial neural network consists of five hidden layers. The final layer is a single neuron that outputs the gyration radii prediction in nm.

2. *Neural Network Architecture*

The hidden layers of the neural network play a crucial role in extracting intricate patterns and relationships within the input data shown in Fig. 3. The first hidden layer consists of 64 units employing the Rectified Linear Unit (ReLU) activation function, promoting nonlinear transformations. Subsequently, the second, third, and fourth hidden layers comprise 16, 8 and 4 units, respectively, utilizing the Hyperbolic Tangent (Tanh) and ReLU activation function alternatively. This configuration is designed to facilitate the extraction of complex representations from the input features, allowing the neural network to discern relationships between the input states. The output layer of the neural network is composed of a single neuron, employing a linear activation

function. This choice of activation function is tailored for regression tasks, aiming to predict a continuous output of the radius of gyration.

3. *Neural Network Training*

The training process involves the optimization of model parameters over 1200 epochs, utilizing a batch size of 50. We employ Adam optimizer [39] with a learning and decay rate of 0.0021 and 0.0005 respectively. We jointly use Mean Squared Logarithmic Error Loss (MSLE) and Huber loss as the total loss function. The main advantage of MSLE is that it penalizes large errors more severely than Mean Squared Error (MSE) by taking the logarithm of the predicted and target values and encourages the model to predict relative errors accurately. We run the MLP network several times and incorporate dropout layers to confirm that both training and testing accuracy stay within $90-96\%$ range indicating robustness of the network training within these 1200 epochs. The three dropout layers with $30\%, 20\%$, and 20% dropout rate (shown in Fig. 3) and early stopping criteria reduce the chance of model over-fitting on the training data.

III. Results

A. Neural Network Predictions on Known IDPs

We validate our ANN performance by comparing the MSE error of the gyration radii values obtained from the BD simulation with those from the neural network prediction. The Table I denotes the 33 IDPs used to evaluate the model performance. The IDPs listed in Table I were studied in great detail using two HPS models and produced nearly identical results [19]. We ensured that these IDP sequences are never seen by the ANN. The gyration radii values from the BD simulation using the HPS2 model and the ANN predicted values are listed in the third and 4th column, respectively. The ANN predicts the gyration radii with very high accuracy with a 0.50% margin of error. The Fig. 4 shows the simulation vs. predicted radius of gyration plot for these 33 IDPs. Most of the data points fall closely to the unit slope straight line except two proteins CspTm and ProTa-N with larger deviation. In our previous studies [19] we observed this off-trend for the same two proteins that demands more careful attention. Both of these proteins are highly charged and it is likely the electrostatics need to be perfected.

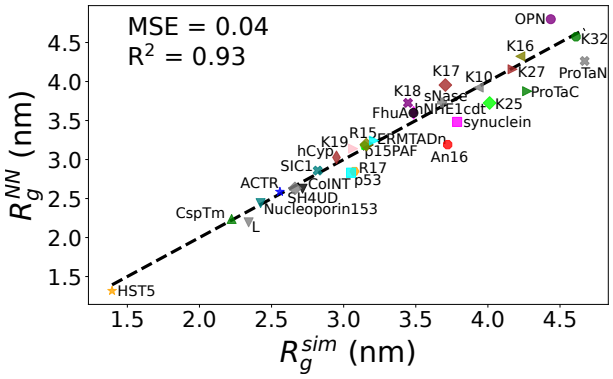

FIG. 4. Gyration radii of each of the 33 IDPs is obtained from the BD simulation and from the neural network. The IDPs are represented with symbols in different colors. The black dashed line denotes the line of unit slope. The mean square error (MSE) is calculated from the average squared deviation of NN-predicted gyration values from the BD simulations. R^2 is used for calculating the correlation coefficient.

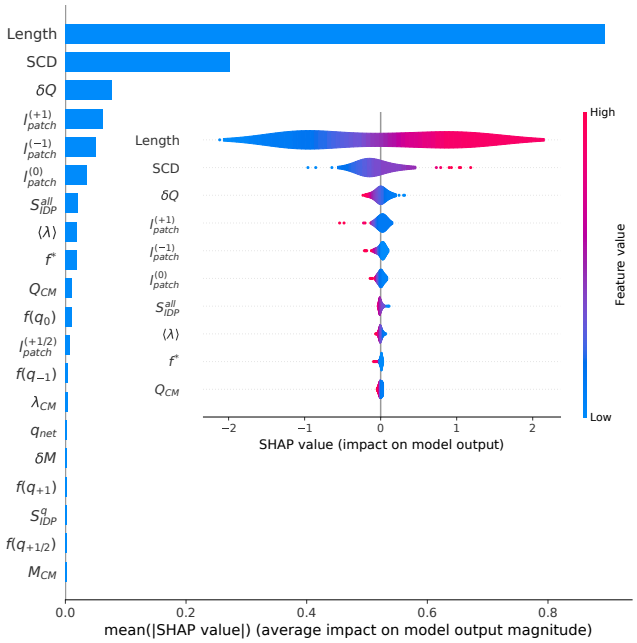

FIG. 5. SHAP value analysis shows the feature importance in descending order. The inset (violin) plots quantitatively describe the individual feature impact on gyration radii.

B. Identifying the feature importance using Explainable AI

Neural network models are often described as black-box models due to their use of nonlinear activation functions and multiple hidden layers, which make it almost impossible to determine how the input features contribute to meaningful outcomes, even if the model's accuracy is high. To interpret how the model arrives at its predictions, we borrow the idea of Shapley values [37] from game theory to explain the workings of our neural network model. Shapley values (named after Lloyd Shapley) provide a method to fairly distribute the "payout" (prediction) among the features (players) based on their contributions. We utilize SHAP analysis [38] to understand our neural network's decision-making process

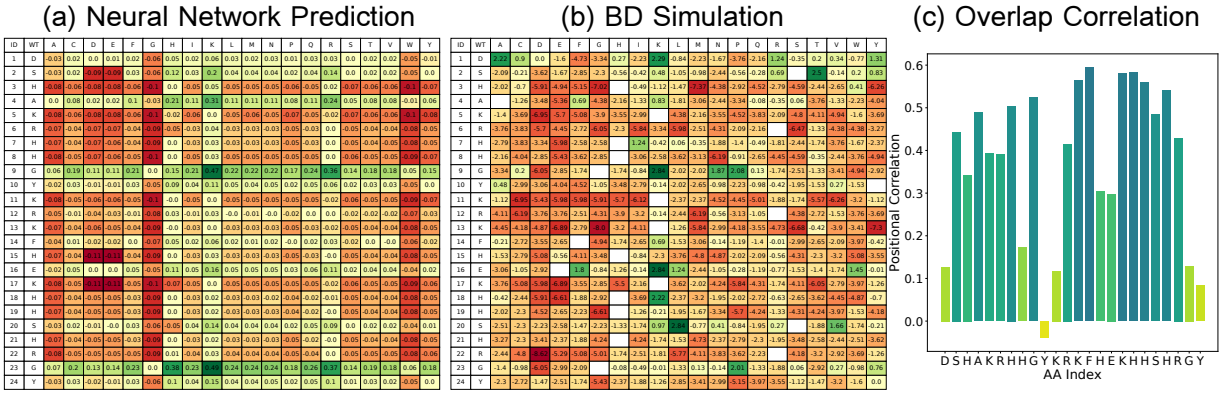

FIG. 6. (a) The checker-board heat-map of the percentage change in the gyration radii of HST5 with the point mutation from neural network prediction. The first column of the map denotes the wild type sequence and the point mutations are listed in the first row. The corresponding percentage change in the gyration radii is shown in red to green colors. The dark red denotes the mutant gyration radii is greater than the wild-type sequence and vice versa for the green boxes. (b) denotes the same changes obtained from the BD simulation. (c) compares the correlation between the neural network prediction and the BD simulations results.

by quantifying the importance of each input feature and their interactions.

Fig. 5 shows the SHAP analysis of relative weights of 21 input features on the predicted gyration radii using ANN. Each point on the violin plots (inset) represents a SHAP value for one of 21 features corresponding to a single data point, with a color (red/blue for high/low values). As expected, since IDPs are described as $\langle R_g \rangle \approxeq 0.38L^{0.56}$ (Eqn. 8), SHAP predicts that length L of the IDP is the most important feature which commensurate to polymer physics principle with high values of L (indicated by red) correspond to positive SHAP values. SHAP also brings out the relative wights of various quantities pertaining to the charges on the gyration radii which is not easily interpretable or less obvious. For example, SHAP predicts that after the IDP length L, the next most important contribution comes from SCD. which has a positive impact on the $\langle Rg \rangle$, with higher values increasing the $\langle Rg \rangle$ while lower SCD slightly decreases the Rg. Other features such as δQ, l^{+1}_{patch}, and l^{-1}_{patch} also affect the gyration radii but to a lesser extent. The important fact is that the SHAP analysis will promote further physics based research to come up with a justification why δQ has less weight compared to the SCD parameter.

C. Identifying the change in gyration radii associated with Missense Mutations

IDPs are frequently subject to missense mutations, wherein a singular amino acid is substituted by one of the remaining 19 amino acids. Pathogenic missense variants disrupt protein function and reduce organismal fitness, while benign missense variants have limited effects. Consequently, a comprehensive exploration of all potential mutant sequences for a given IDP is imperative. Traditional Molecular Dynamics (MD) simulations for calculating gyration radii across the entire spectrum of mutant sequences are inherently time-consuming. To circumvent this computational bottleneck, we employ our trained ANN model, which significantly expedites calculation of gyration radii for all conceivable mutants. This innovative approach enables a significant reduction in computational time, rendering our methodology highly efficient for exhaustive mutational analyses within the realm of IDPs.

Finally, we to demonstrate the efficacy of the ANN prediction we make a one-to-one comparison of the ΔR_g^{Sim} with ΔR_g^{ANN} in a color map for the HST5, IDP of 24 amino acid residues shown side by side in Fig. 6 and note that the corresponding overlap integral for each missense mutation is about 60%. This opens the scope that the neural network performance can be improved by training with a high amount of simulation/experimental data if available.

D. Summary and Outlook

To summarize, we have provided a straight forward, reasonably accurate and computationally efficient algorithm to identify missense mutations in IDPs using a combination of simulation results from a well tested CG model of IDPs and feeding this information to a ANN. We have shown that the ANN predicts the same with 96% accuracy and three orders of magnitude faster than the BD algorithm. The algorithm is open for significant and continuous improvement as more and more data from the BD simulation data backed up by experiments become available to train the ANN. The heat map of the results shows that a large fraction of the mutations are less significant while only a few undergo a major physical change. Thus the use of ANN to rapidly identify this small subset is a key and important element in our findings. This will

IDP Name	N	R_g^{sim} (nm)	R_g^{NN} (nm)	ΔRg (%)
HST5	24	1.394	1.317	5.55
IN	60	2.203	2.160	1.96
L	64	2.341	2.195	6.24
CspTm	67	2.223	2.238	-0.66
ACTR	71	2.558	2.587	-1.14
Nucleoporin153	81	2.424	2.441	-0.69
SH4UD	85	2.662	2.630	1.21
SIC1	90	2.820	2.858	-1.36
p53	93	3.051	2.833	7.13
CoINT	98	2.715	2.626	3.27
K19	99	3.065	3.140	-2.46
R17	100	3.075	2.849	7.34
p15PAF	111	3.157	3.191	-1.09
ProTaN	112	4.671	4.263	8.74
R15	114	3.145	3.176	-0.99
hCyp	115	2.950	3.028	-2.65
ERMTADn	122	3.210	3.243	-1.03
ProTaC	129	4.272	3.876	9.26
K18	130	3.446	3.726	-8.11
hNHE1cdt	131	3.485	3.600	-3.31
sNase	136	3.682	3.730	-1.30
synuclein	140	3.787	3.483	8.03
FhuA	143	3.484	3.598	-3.27
K17	145	3.705	3.955	-6.74
K27	167	4.171	4.157	0.32
K10	168	3.936	3.923	0.32
K16	176	4.225	4.323	-2.32
K25	185	4.012	3.725	7.15
An16	185	3.721	3.191	14.25
K32	198	4.611	4.573	0.83
OPN	219	4.436	4.799	-8.18

TABLE I. Gyration Radii comparison between the simulation radius of gyration values and the neural network predicted value. The 1st and 2nd columns list the name and the number of amino acid residues. The simulation radius of gyration R_g^{sim} is listed in the 3rd and neural network predicted values R_g^{NN} are written in the 4th columns respectively. The last column denotes the percentage error.

allow to go back to these few mutated species and carry out BD simulation to gain a fundamental understanding of the accompanying changes not attainable by machine learning algorithm only.

Though, the neural network traditionally viewed as a black-box, we utilize SHAP analysis to understand the degree of impact of individual physics based feature on gyration radii prediction and in turns sheds lights on the inner workings of our neural network model. We showed that the length of IDPs is the most significant predictor of the radius of gyration, followed by the SCD parameter. This transparency not only aligns with polymer physics principles but also enhances the interpretability and trustworthiness of our NN model. This highlight the importance of explainable AI in understanding and validating complex neural network models with more confidence.

We want to conclude with some remark on recent related publications where the use of deep learning has been reviewed [40, 41] and explored in studying IDPs [42–44]. A fundamental challenge for the IDPs is to predict the conformational properties directly from the primary sequence. Thus, in principle there can be multple approaches in terms of deep learning strategies, choice of training sets, use of simulation and experimental data, etc., to increase the convergence. Chao et al. [42] used various regression models, *e.g.*, L2-Linear regularized regression (LRR), Kernel ridge regression (KRR), Gaussian process regression (GPR), and free forward neural network (FNN) to compare MSEs from different regression methods. LRR, KRR, and GPR produce MSE in the range 0.93-0.94, comparable to ours reported here. Janson *et al.* [43] used adversarial network to generate conformational ensembles of IDPs. In a similar spirit Lotthammer *et al.* introduced a deep learning package called ALBATROSS [44] using a bidirectional recurrent neural network with long short-term memory cells (LSTM-BRNN). Google DeepMind recently reported prediction of pathogenic mutations in α-synuclein [3]. The final prediction of α-Missense is a pathogenicity score, which reflect the likelihoods of mutations to cause disease, rather than predicted changes in protein structures [45]. While the algorithm may well be very efficient, this provides a probability score without providing any physical justification. Our approach on the contrary based on physical models of IDPs which have been used to disease progression and liquid-liquid phase separation in living systems with the parameters validated by experiments. Thus, the predictions from our model can be used for further inference and analysis of pathogenic mutations responsible for AD, PD and other disease caused by missense mutations. Whether the physical properties of the IDPs and their conformational ensemble can be useful for predicting biophysical properties such as protein stability and interactions, they have generally been of little benefit for clinical variant effect prediction Thus identifying missense mutations as demonstrated here is of broad general application in rapidly identifying harmful mutations in proteins.

IV. Supplementary Materials

We include the heat-maps of the percentage change in the gyration radii due to missence mutations for IN, CspTm, ProTaN, ProTaC, α-synuclein, and p53 IDPs as PDF files.

V. Acknowledgments

All computations were carried out using STOKES High-Performance Computing Cluster at UCF.

[1] Vladimir Vacic, Phineus R. L. Markwick, Christopher J. Oldfield, Xiaoyue Zhao, Chad Haynes, Vladimir N. Uversky, Lilia M. Iakoucheva, *Disease-Associated Muta-*

tions Disrupt Functionally Important Regions of Intrinsic Protein Disorder, PLoS Comput Biol 8(10): e1002709. doi:10.1371/journal.pcbi.1002709 Comp. Biol.

[2] Shannon Stefl, Hafumi Nishi, Marharyta Petukh, Anna R. Panchenko and Emil Alexov, *Molecular Mechanisms of Disease-Causing Missense Mutations*, J. Mol. Biol. **425**, 3919 (2013).

[3] Cheng et al.,*Accurate proteome-wide missense variant effect prediction with AlphaMissense*, Science 381, eadg7492 (2023)

[4] Vladimir D. Manyilov *et al.*, *Chaotic aging: intrinsically disordered proteins in aging-related processes*, Cellular and Molecular Life Sciences (2023) 80:269.

[5] Orkid Coskuner-Weber and Vladimir N. Uversky, *Insights into the Molecular Mechanisms of Alzheimer's and Parkinson's Diseases with Molecular Simulations: Understanding the Roles of Artificial and Pathological Missense Mutations in Intrinsically Disordered Proteins Related to Pathology*, Int. J. Mol. Sci. **19**, 336 (2018) ; doi:10.3390/ijms19020336

[6] Coskuner-Weber, O., Mirzanli, O. & Uversky, V. N. Intrinsically disordered proteins and proteins with intrinsically disordered regions in neurodegenerative diseases. Biophysical Reviews 14, 679-707 (2022).

[7] Simon Alberti, Amy Gladfelter, and Tanja Mittag, *Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates*, Cell **176**, 419 (2019)

[8] Salman F. Banani, Hyun O. Lee, Anthony A. Hyman and Michael K. Rosen, *Biomolecular condensates: organizers of cellular biochemistry*, Nature Review **18** 285 (2017).

[9] A database of intrinsically disordered proteins: *https://mobidb.bio.unipd.it/*

[10] Sickmeier, M. *et al.* DisProt: the Database of Disordered Proteins. Nucleic Acids Research 35, D786-D793 (2007).

[11] Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B. & Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. PLOS Computational Biology **14**, e1005941 (2018).

[12] Carlos W. Bertoncini, Young-Sang Jung, Claudio O. Fernandez, Wolfgang Hoyer, Christian Griesinger, Thomas M. Jovin, and Markus Zweckstetter, *Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein*, PNAS **102**, 1430 (2005).

[13] Marco D. Mukrasch, Stefan Bibow, Jegannath Korukottu, Sadasivam Jeganathan, Jacek Biernat, Christian Griesinger, Eckhard Mandelkow, Markus Zweckstetter, *Structural Polymorphism of 441-Residue Tau at Single Residue Resolution*

[14] Benjamin Schuler, Andrea Soranno, Hagen Hofmann, and Daniel Nettels, *Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins], Annu. Rev. Biophys. 2016. 45:207–31.*

[15] Douglas M Fowler & Stanley Fields *Deep mutational scanning: a new style of protein science*

[16] Lillian T. Chong, Jed W. Pitera, William C. Swope, and Vijay S. Pande, *"Comparison of computational approaches for predicting the effects of missense mutations on p53 function"*, J. Mol. Graph. Model. **27**, 978–982 (2009).

[17] Ganne A, Balasubramaniam M, Ayyadevara S and Shmookler Reis RJ, *Machine-learning analysis of intrinsically disordered proteins identifies key factors that contribute to neurodegeneration-related aggregation*, Front. Aging Neurosci. **14**, 938117 (2022); doi: 10.3389/fnagi.2022.938117

[18] 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. Proceedings of the National Academy of Sciences 118, (2021).

[19] S. Seth, B. Stine, and A. Bhattacharya, *Fine structures of Intrinsically disordered proteins*, J. chem. Phys. **160**, 014902 (2024).

[20] Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. Nature 555, 61-66 (2018).

[21] Sottini, A. *et al.* Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. Nature Communications 11, 5736 (2020).

[22] Gregory-Neal W. Gomes, Mickal Krzeminski, Ashley Namini, Erik W. Martin, Tanja Mittag, Teresa Head-Gordon, Julie D. Forman-Kay, and Claudiu C. Gradinaru, *Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET*, J. Am. Chem. Soc. **142**, 15697-15710 (2020).

[23] Dannenhoffer-Lafage, T., & Best, R, B. A Data-Driven Hydrophobicity Scale for Predicting Liquid-Liquid Phase Separation of Proteins. J. Phys. Chem. B 125, 4046-4056 (2021).

[24] HPS1 is the same as the scale introduced by Dignon *et al.*, excepting the. Likewsie, HPS2 scale is the same as in M3 introduced by Tesei excepting the end groups are charged.

[25] Ashbaugh, H. S. & Hatch, H. W. Natively Unfolded Protein Stability as a Coil-to-Globule Transition in Charge/Hydropathy Space. Journal of the American Chemical Society 130, 9536-9542 (2008).

[26] J. N. Israelachvili, *Intermolecular and Surface forces*, 3rd edition, Elsevier (2011).

[27] Akerlof, G. C. & Oshry, H. I. The Dielectric Constant of Water at High Temperatures and in Equilibrium with its Vapor. Journal of the American Chemical Society 72, 2844-2847 (1950).

[28] Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology 157, 105-132 (1982).

[29] Engelman, D. M., Steitz, T. A. & Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annual Review of Biophysics and Biophysical Chemistry 15, 321-353 (1986).

[30] Hopp, T. P. & Woods, K. R. A computer program for predicting protein antigenic determinants. Molecular Immunology 20, 483-489 (1983).

[31] Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. Journal of Molecular Biology 179, 125-142 (1984).

[32] Cornette, J. L. *et al.* Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. Journal of Molecular Biology 195, 659-685 (1987).

[33] *Neural networks from scratch in Python*, Harrison Kinsley & Daniel Kukieła (2020).

[34] Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. & Pappu, R. V. Net charge per residue modulates confor-

mational ensembles of intrinsically disordered proteins. Proceedings of the National Academy of Sciences 107, 8183-8188 (2010).

[35] Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proceedings of the National Academy of Sciences 110, 13392-13397 (2013).

[36] Ghosh, K., Huihui, J., Phillips, M. & Haider, A. Rules of Physical Mathematics Govern Intrinsically Disordered Proteins. Annual Review of Biophysics 51, 355-376 (2022).

[37] Shapley, Lloyd S. A value for n-person games. Contributions to the Theory of Games 2.28, 307-317 (1953).

[38] Lundberg, Scott M., & Lee, Su-In, A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30, 4765-4774 (2017).

[39] D. P. Kingma and J. L. Ba, *ADAM: A Methof for Stochastic Optimization*, ICLR 2015.

[40] Arvind Ramanathan, Heng Ma, Akash Parvatikar, and S Chakra Chennubhotla, Artificial intelligence techniques for integrative structural biology of intrinsically disordered proteins, Current Opinion in Structural Biology 2021, 66:216–224

[41] Kresten Lindorff-Larsen and Birthe B. Kragelund, On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins, J. Mol. Bio 433 (2021) 167196.

[42] Tzu-Hsuan Chao, Shiv Rekhi, Jeetain Mittal, and Daniel P. Tabo, Data-driven models for predicting intrinsically disordered protein polymer physics directly from composition or sequence, Mol. Syst. Des. Eng., 2023, **8**, 1146–1155.

[43] Giacomo Janson, Gilberto Valdes-Garcia, LimHeo, & Michael Feig, Direct generation of protein conformational ensembles via machine learning, Nature Communications, **14**, 774(2023).

[44] Lotthamer, J.M., Ginell, G.M.,Griffith, D., Emenecker, R.J., Holehouse, A.S. Direct prediction of intrinsically disordered protein conformational properties from sequence, Nature Methods, **21**, 465-476 (2024).

[45] Joseph A. Marsh and Sarah A. Teichmann, *Predicting pathogenic protein variants*, Science **381**, 1284 (2023).