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: March 5, 2025)

ABSTRACT

We use a combination of Brownian dynamics (BD) simulation results and Deep Learning (DL) strategies for rapid identification of large structural changes caused by missense mutations in intrinsically disordered proteins (IDPs). We used ~ 6500 IDP sequences from MobiDB database of length $20 - 300$ to obtain gyration radii from BD simulation on a coarse-grained single bead amino acid model (HPS2 model) used by us and others [Seth *et al.* J. Chem. Phys. **160**, 014902 (2024), Dignon *et al.* PLOS Comp. Biology, 14, 2018, Tesei *et al.* PNAS, 118, 2021] to generate the training sets for the DL algorithm. Using the gyration radii $\langle R_g \rangle$ of the simulated IDPs as the training set, we develop a multilayer perceptron neural net (NN) architecture that predicts the gyration radii of 33 IDPs previously studied using BD simulation with 97% accuracy from the sequence and the corresponding parameters from the HPS model. We now utilize this NN to predict gyration radii of every permutation of missense mutations in IDPs. Our approach successfully identifies mutation-prone regions that induce significant alterations in the radius of gyration when compared to the wild-type IDP sequence. We further validate the prediction by running BD simulations on the subset of identified mutants. The neural network yields a $(10^4 - 10^6)$-fold faster computation in the search space for potentially harmful mutations. Our findings have substantial implications for rapid identification and understanding diseases related to missense mutations in IDPs and for the development of potential therapeutic interventions. The method can be extended to accurate predictions of other 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 [7–11], a subject which has gained considerable attention in last decade as they offer significant challenges in understanding how universal aspects of the IDPs are altered by sequence specific properties.

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 [12, 13] 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. Small angle X-Ray scattering (SAXS) [14–17], high-field NMR [18, 19] and single molecule Förster resonance energy transfer (sm-FRET) [20–24] 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 [25, 26]. 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 [26, 27]. The available experimental results have been instrumental in validating various coarse-grained models to study IDPs [28–32]. For the simplest CG models each amino acid has been lumped into a single bead [28–30]. These CG model of IDPs have been used to study single molecule properties, liquid-liquid phase separation (LLPS) in IDPs, formation of IDP-complexes [33, 43–48] and successfully extracted some key aspects of the IDPs [24]. In addition to these single-bead amino acid models those use different hydropathy scales (please refer to the next section), the two-bead representation of IDPs (SOP-IDP model) by

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

Baul *et al.* [31] has converged with the single-bead model and SAXS results [17]. Recently the two-bead SOP-IDP model has been extended to account for multi-domains and IDPs [32]. 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 more IDPs and available results may narrow down and direct further experiments [33]. In a previous paper, we made a detailed investigation and comparison of two CG models of IDPs (HPS1 and HPS2) [34] 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 [30]. Therefore, in this article we use HPS2 model to to carry out BD simulation for a random selection of approximately 6500 IDPs from MobiDB database with disorder score $\geq 99\%$ whose statistical properties are shown Fig. 1. Despite the fact that many properties of the Single IDPs have been investigated, we use the gyration radii from the BD simulation as the *supervised training set* for an Artificial Neural Net (ANN) to predict how missense mutation affects the physical properties of the proteins, as it is the gyration radii data that has been tabulated the most in previous studies [35, 36]. 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 97% 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 and the Neural Net architecture used to study and predict missense mutations in IDPs. Details for the HPS2 model are given in our previous paper [30].

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 [37]

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

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 the same as CALVADOS introduced by Tesei *et al.* [29] using the Bayesian parameter-learning procedure to optimize the hydropathy values drawn from many different hydropathy scales [38–42] excepting the end charges are kept as they are. 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 [30].

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 [49]. 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 [50]

$$\epsilon_r(T)=\frac{5321}{T}+233.76-0.9297T$$
$$+1.147\times 10^{-3}T^2-8.292\times 10^{-7}T^3. \quad (7)$$

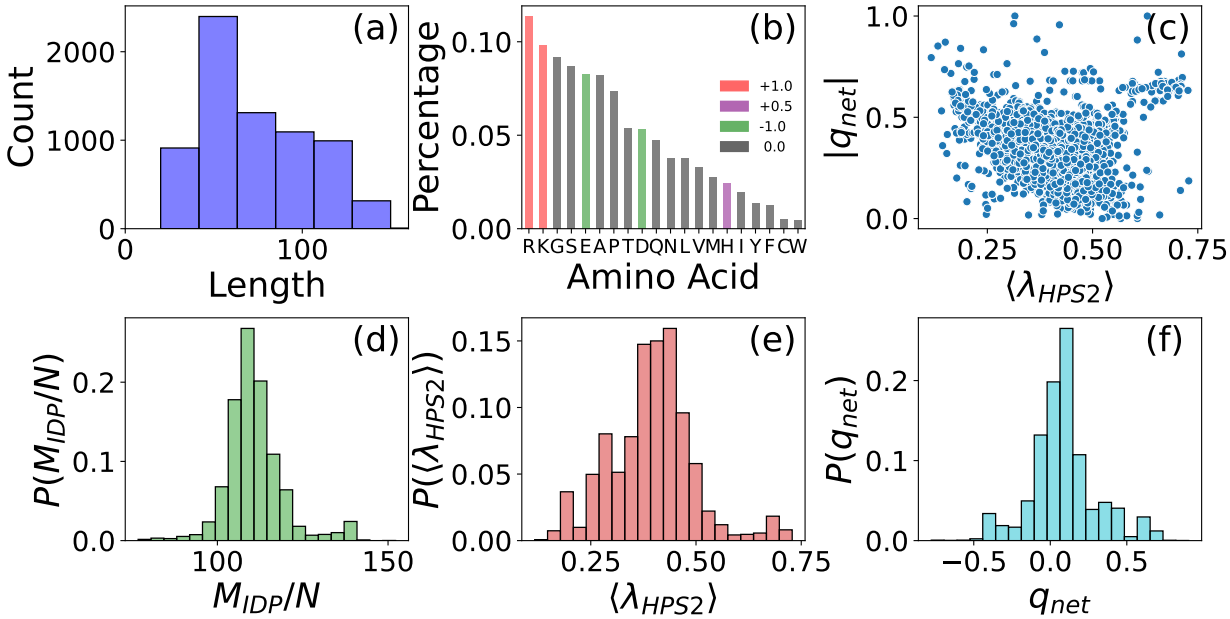

FIG. 1. Characteristics of $\sim$ 6500 IDPs with $\geq$ 99% disorder predictor score obtained from the MobiDB [12] database. (a) Length distribution of the IDPs. (b) The distribution of the propensity of occurrence of the individual amino acids. The positive, negative, neutral amino acids are colored 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.

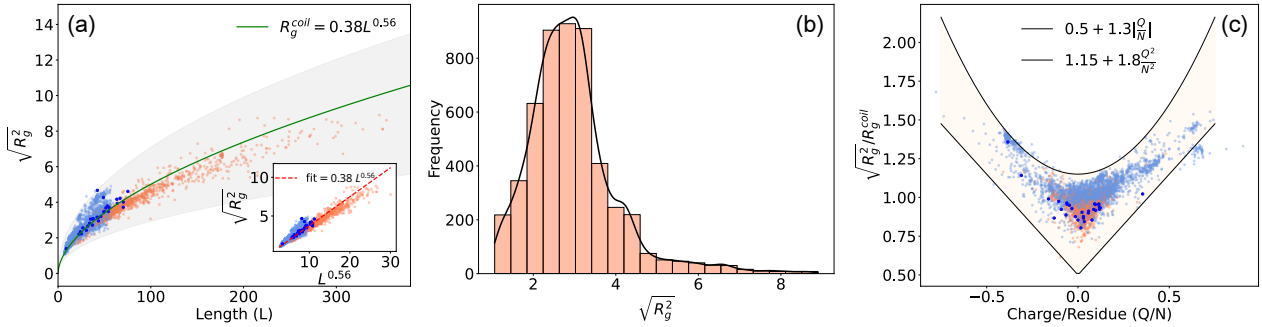

FIG. 2. (a) Radius of gyrations of IDPs as a function of chain length. The lighter blue-shaded circles show the gyration radii of simulated IDPs using HPS2 model and the darker blue circles show the 33 IDPs from our previous study [30]. The coral-shaded circles represent 1400 human disordered proteome IDPs from [36]. The green line $R_g^{coil} = 0.38N^{0.56}$ 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 all the $\approx$ 8000 IDPs. (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.

B. Statistical properties of IDPs

We gathered approximately 6500 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 3000 IDPs with of length between 40 to 60 as shown 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. We gather 1400 human disordered proteome IDPs [36] and the 6500 datasets of IDPs form MobiDB dataset to 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. We 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 for the 6500 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 equilibrium 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

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

the same scaling relation that we obtained on a smaller subset of data for 33 chosen IDPs [30]. 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 [51] 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 23 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 ≈ 8000 IDP dataset into 80-20 training and testing dataset. We carry out MD simulation using the HPS2 model on 6500 IDPs (and use additional $\approx$1450 gyration values from reference [36]) 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_{IDP}^{all} and the Shannon entropy for the charged amino acids in the sequence S_{IDP}^{q} (in units of Boltzmann constant k_B) of the FASTA sequence defined as follows

$$S_{IDP}^{all} = -\sum_{i} p_i \log_2(p_i). \tag{12a}$$

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

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_{IDP}^{q}, $p_\alpha = n_\alpha/N_q$, 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:

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

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

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

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

$$SHD = \frac{1}{N}\sum_{i=2}^{N}\sum_{j=1}^{i-1} (\lambda_i + \lambda_j) \cdot |i-j|^{\beta} \tag{13e}$$

Here N_q^{+1}, $N_q^{1/2}, and N_q^{-1}/N$ are the number of IDPs with $+1$, $+1/2$, and -1 charges, so that f_q^{+1}, $f_q^{1/2}$, f_q^{-1} are the corresponding fractions, and $f_q^0 = \left(N - N_q^{+1} - N_q^{+1/2} - N_q^{-1}\right)/N = N_q^0/N$ is the fraction of neutral amino acids. The quantities $f^+ = f_q^{+1} + f_q^{1/2}$ and $f^- = f_q^{-1}$ are the net positive and negative charge per residue of an IDP. The importance of charge asymmetry parameter f^* [43, 44] and the sequence charge decoration parameter SCD parameter (Eqn. 13d) introduced by Sawle and Ghosh [52, 53] have been discussed in the literature. In a similar spirit the Zheng *et al.* introduced hydropathy decoration parameter SHD (Eqn. 13e) [54] and provided physical arguments to show $0.95 \leq \beta \leq 1.1$. They further showed that $\beta = -1.0$ describes scaling behavior of IDPs well. We have used $\beta = -1$ in our analysis.

The quantities

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

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). $l_{patch}^{(q)} \neq 0$, for $q = \pm 1, \frac{1}{2}$ if at least two consecutive amino acids have the same charge within the charge sequence (else $l_{patch}^{(q)} = 0$). Finally, we introduce the Sequence-Based Compactness Score (SBCS) metric to evaluate the compactness of the IDP based on the distribution and clustering of high-λ residues. It is defined as

$$\text{SBCS} = \langle\lambda\rangle \sum_{j=1}^{k-1} \frac{1}{d_{j,j+1}} \tag{15}$$

where, $d_{j,j+1}$ is the is the distance between consecutive high-λ (hydrophobic) residues, indexed by j and $j+1$, whose $\lambda > 0.5$ in the sequence, and k is the total number those residues. The SBCS provides insights into how high-λ residues are distributed within a sequence, and their potential role in influencing the aggregation behavior. It combines both the hydrophobicity of residues and their spatial organization to measure the overall compactness of the sequence The parameters as described in Eqns. 9(a)-9(d), 10(a)-(c), 11(a)-(c), 12(a)-(b), 13(a)-(e), 14, and 15 along with the length L of the amino acid constitute 23 important physics based features derived from the IDP sequence information as input vectors to train the MLP network as described in the next section. Since not all of them aren't linearly independent, we have further investigated the relative weights these quantities through SHAP in Fig. 5 [55, 56].

2. *Neural Network Architecture*

The eight 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 non-linear transformations. Subsequently, the $2^{nd}, 3^{rd}, 4^{th} \cdot\cdot 8^{th}$ hidden layers comprise $128, 64, 32, 16, 8$ and 4 units, respectively, utilizing the Hyperbolic Tangent 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

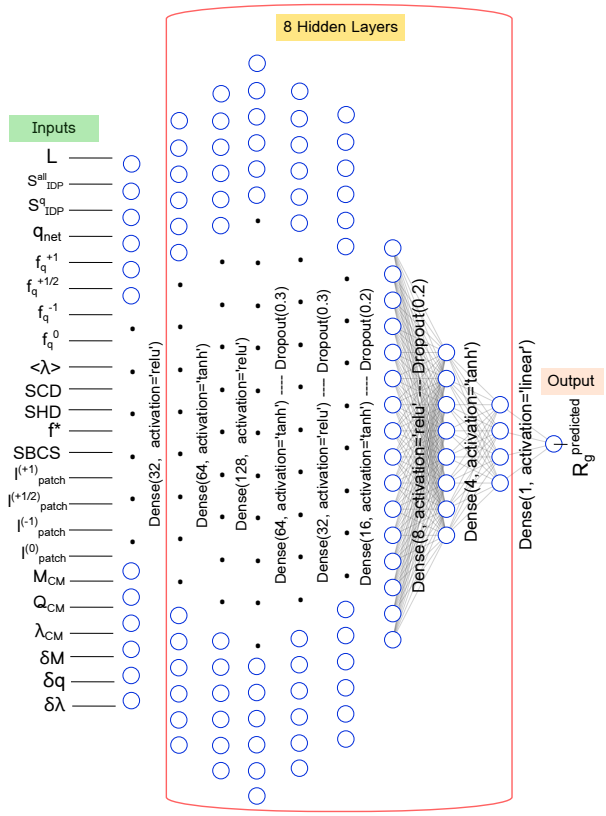

FIG. 3. The fully connected neural network architecture to predict the gyration radii of a IDP from its sequence description. The 23 input features are obtained from the IDP FASTA sequence (please see text for the explanation of the input parameters) and are fed into an artificial neural network consists of eight hidden layers. The final layer is a single neuron that outputs the gyration radii prediction in nm.

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 [57] 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 $90-97\%$ range indicating robustness of the network training within these 1200 epochs. The four dropout layers with $30\%, 30\%, 20\%$, 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 [30]. 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. For most of the IDPs the ANN predicts the gyration radii based on the physical characteristics of the HPS model only with an accuracy with an error $\lesssim 5\%$ with an exception of four IDPs (hNHE1cdt, synuclein, An16, and K32). We find that the use more hidden layers and the newly introduced SBSC parameter have significant effects on increasing the accuracy. The Fig. 4 shows the simulation vs. predicted radius of gyration plot for these 33 IDPs.

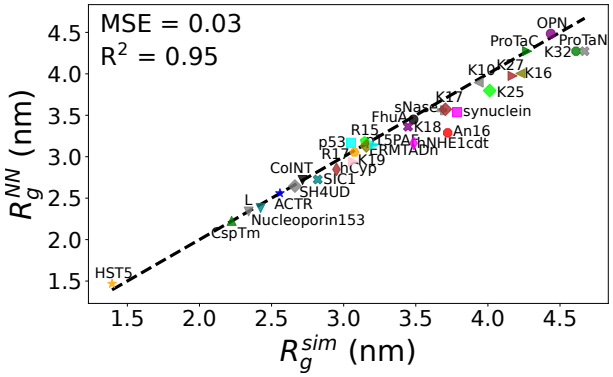

FIG. 4. Gyration radii of each of the 33 IDPs is obtained from the BD simulation and from the neural network. The IDPs are represented with symbols in different colors. The black dashed line denotes the line of unit slope. The mean square error (MSE) is calculated from the average squared deviation of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

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 [55]

IDP Name	N	R_g^{sim} (nm)	R_g^{NN} (nm)	ΔRg (%)
HST5	24	1.394	1.468	-5.30
IN	60	2.203	2.235	-1.46
L	64	2.341	2.339	0.09
CspTm	67	2.223	2.226	-0.13
ACTR	71	2.558	2.560	-0.07
Nucleoporin153	81	2.424	2.382	1.73
SH4UD	85	2.662	2.646	0.59
SIC1	90	2.820	2.725	3.38
p53	93	3.051	3.169	-3.86
CoINT	98	2.715	2.719	-0.14
K19	99	3.065	2.941	4.05
R17	100	3.075	3.054	0.69
p15PAF	111	3.157	3.130	0.86
ProTaN	112	4.671	4.273	8.52
R15	114	3.145	3.189	-1.40
hCyp	115	2.950	2.847	3.50
ERMTADn	122	3.210	3.142	2.11
ProTaC	129	4.272	4.273	-0.03
K18	130	3.446	3.362	2.45
hNHE1cdt	131	3.485	3.164	9.21
sNase	136	3.682	3.560	3.30
synuclein	140	3.787	3.537	6.61
FhuA	143	3.484	3.449	1.00
K17	145	3.705	3.572	3.60
K27	167	4.171	3.975	4.70
K10	168	3.936	3.894	1.06
K16	176	4.225	4.005	5.21
K25	185	4.012	3.798	5.33
An16	185	3.721	3.288	11.64
K32	198	4.611	4.273	7.33
OPN	219	4.436	4.484	-1.07

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.

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 [56] to understand our neural network's decision-making process by quantifying the importance of each input feature and their interactions.

Fig. 5 shows the SHAP analysis of relative weights of 23 input features on the predicted gyration radii using ANN. Each point on the violin plots (inset) represents a SHAP value for one of 23 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

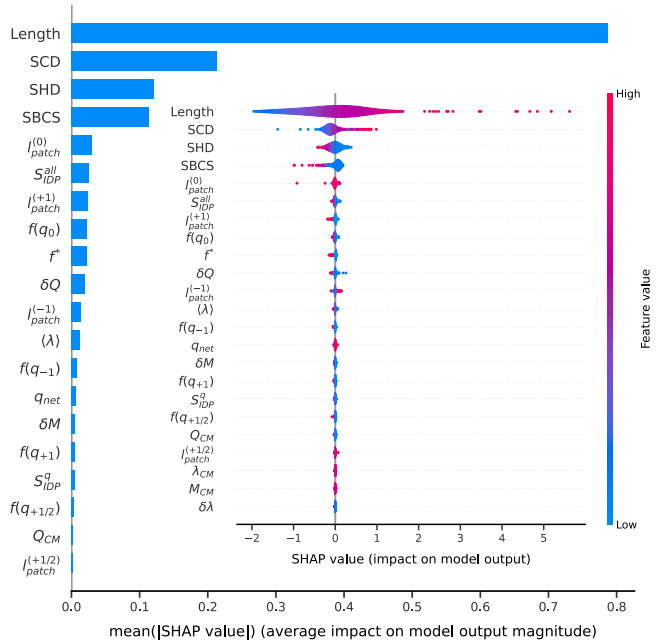

FIG. 5. SHAP value analysis shows the feature importance in descending order. The inset (violin) plots quantitatively describe the individual feature impact on gyration radii.

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 and SHD. SCD has a positive impact on the $\langle R_g \rangle$, with higher values increasing the $\langle R_g \rangle$ while lower SCD slightly decreases the $\langle R_g \rangle$. For the case of SHD this trend it opposite to that of SCD. The 4th significant parameter we find is the SBCS introduced by us. This parameter has comparable to that of SHD with similar trend. Other features such as δQ, l_{patch}^{+1}, and l_{patch}^{-1} 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 a particular quantity is more/less important than another one.

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

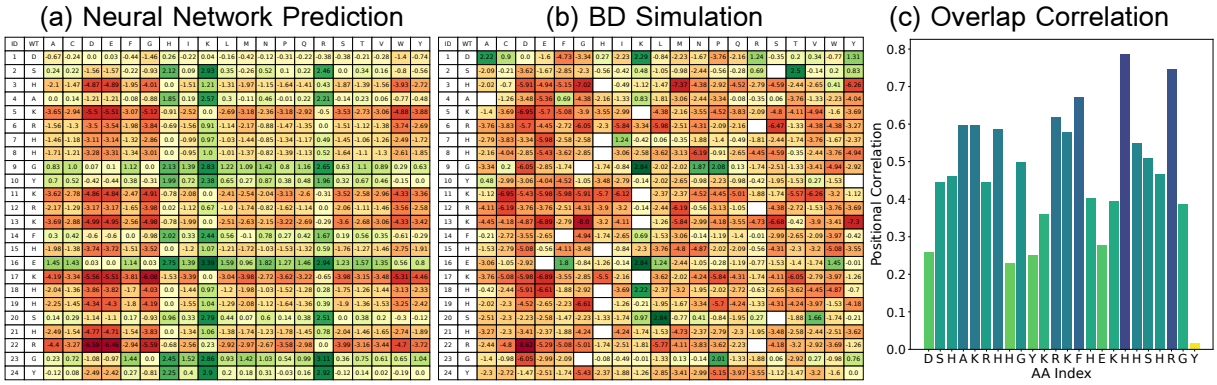

FIG. 6. (a) The checker-board heat-map of the percentage change in the gyration radii of HST5 with the point mutation from neural network prediction. The first column of the map denotes the wild type sequence and the point mutations are listed in the first row. The corresponding percentage change in 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.

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. It is worth noting that a large variation in $\langle R_g \rangle$ may not necessarily imply a pathogenic variant and worthy of further studies.

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 as large as 80%. This opens the scope that the neural network performance can be improved by training with a high amount of simulation/experimental data if available along with new sequence based functions, such as SBSC introduced in this paper.

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 97% accuracy and six orders of magnitude faster than the BD algorithm. The algorithm is open for significant and continuous improvement as (i) more and more data from the BD simulation data backed up by experiments become available to train the ANN and (ii) choice of new parameters for the training sets. The heat map of the results shows that a large fraction of the mutations are less significant while only a few undergo a large variation in sizes. Thus, the use of ANN to rapidly identify this small subset is a key and important element in our findings. This will 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, SHD, and the SBSC parameter introduced in this paper. 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 making connection to this work with other work where Physics based model along with ML algorithms have been used to explore IDPs [35, 36, 58–61]. It is immensely practical to increase the accuracy of predictions the conformational properties of the IDPs as much as possible directly from the sequence itself, although other physical features will be needed for a more comprehensive physical understanding. From the ML perspective there can be multiple approaches in terms of different deep learning strategies, choice of training sets, use of simulation and experimental data, *etc.*. As an example, Chao *et al.* [60] 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 but slightly lower than ours reported here. Janson *et al.* [61] used the well es-

tablished Generative Adversarial Network (GAN) [62] to generate conformational ensembles of IDPs which they call idpGAN. idpGAN can generate physically realistic conformational ensembles of proteins without encountering kinetic barriers in an actual MD simulation and therefore, it is extremely fast and complements ours and other ML methods used by Chao *et al.* A combination of different ML algorithms along with the data generated from simulation of different level of granurality will be a practical approach to speed up the investigations [58]. Lotthammer *et al.* introduced a deep learning package called ALBATROSS [35] using a bidirectional recurrent neural network with long short-term memory cells (LSTM-BRNN) to study a total of 41,202 IDPs and IDRs (19,075 naturally occurring IDRs and 22,127 rationally designed IDRs). However, Lotthammer *et al.* used ABSINTH [63] but not the HPS2 and therefore, a comparison of the results will reinforce the findings to the IDP community. In a related study Tesei *et al.* used CALVADOS and Machine learning model to study 28,058 occurring in human proteome to predict apparent scaling exponents [36]. Tesei *et al.* found only a small shift in the histogram of the entropy in benign and pathogenic variants. By calculating the effective Flory exponents for proteins from Clinvar database [64] they found compactness of benign and pathogenic variants are statistically insignificant. 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 [65]. While the algorithm may well be very efficient, this provides a probability score without providing any physical justification. Our approach (an similar approaches recently used by others) 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. Admittedly, predictions solely based on sequence-gazing and ML algorithms have their limitations, will not be able to capture the physical scenario which will require further knowledge of the dynamics. For example, visualization of free energy landscape has revealed important information about fibril formation in Amyloid-beta [66, 67]. However, collective explainable ML methods without carrying out expensive simulation with the SHAP analysis is extremely useful in narrowing down the key parameters for vast number of IDPs in different physical environment for time consuming simulation studies. Our studies of missense mutations is a practical step in that direction which we believe will promote further work for broad general application in rapidly identifying harmful mutations in proteins and ultimately clinical variant effect predictions.

IV. Supplementary Materials

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

V. Acknowledgments

The research at UCF has been supported by a DARPA AI-BTO (HR00112530047) award. All computations were carried out using STOKES High-Performance Computing Cluster at UCF. We thank both the anonymous reviewers for a thorough reading of the original manuscript and comments.

[1] Vacic, V., Markwick, P. R. L., Oldfield, C. J., Zhao, X., Haynes, C., Uversky, V. N., & Iakoucheva, L. M. (2012). Disease-Associated Mutations Disrupt Functionally Important Regions of Intrinsic Protein Disorder. PLoS Computational Biology, 8(10), e1002709. https://doi.org/10.1371/journal.pcbi.1002709

[2] Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R., & Alexov, E. (2013). Molecular Mechanisms of Disease-Causing Missense Mutations. Journal of Molecular Biology, 425(21), 3919-3936. https://doi.org/10.1016/j.jmb.2013.07.014

[3] Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science, 381(6664). https://doi.org/10.1126/science.adg7492

[4] Manyilov, V. D., Ilyinsky, N. S., Nesterov, S. V., Saqr, B. M. G. A., Dayhoff, G. W., Zinovev, E. V., Matrenok, S. S., Fonin, A. V., Kuznetsova, I. M., Turoverov, K. K., Ivanovich, V., & Uversky, V. N. (2023). Chaotic aging: intrinsically disordered proteins in aging-related processes. Cellular and Molecular Life Sciences, 80(9). https://doi.org/10.1007/s00018-023-04897-3

[5] Coskuner-Weber, O., & Uversky, V. (2018). 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. International Journal of Molecular Sciences, 19(2), 336. https://doi.org/10.3390/ijms19020336

[6] Coskuner-Weber, O., Mirzanli, O., & Uversky, V. N. (2022). Intrinsically disordered proteins and proteins with intrinsically disordered regions in neurodegenerative diseases. Biophysical Reviews, 14(3), 679-707. https://doi.org/10.1007/s12551-022-00968-0

[7] Shin, Y., & Brangwynne, C. P. (2017). Liquid phase condensation in cell physiology and disease. Science, 357(6357). https://doi.org/10.1126/science.aaf4382

[8] Alberti, S., Gladfelter, A., & Mittag, T. (2019). Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. Cell, 176(3), 419-434. https://doi.org/10.1016/j.cell.2018.12.035

[9] Banani, S. F., Lee, H. O., Hyman, A. A., & Rosen, M. K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nature Reviews Molecular Cell Biology, 18(5), 285-298. https://doi.org/10.1038/nrm.2017.7

[10] Alberti, S., & Dormann, D. (2019). Liquid–Liquid Phase Separation in Disease. Annual Review of Genetics, 53(1), 171-194. https://doi.org/10.1146/annurev-genet-112618-043527

[11] Martin, E. W., & Mittag, T. (2018). Relationship of Sequence and Phase Separation in Protein Low-Complexity Regions. Biochemistry, 57(17), 2478-2487. https://doi.org/10.1021/acs.biochem.8b00008

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

[13] Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., Obradovic, Z., & Dunker, A. K. (2007). DisProt: the Database of Disordered Proteins. Nucleic Acids Research, 35(Database), D786-D793. https://doi.org/10.1093/nar/gkl893

[14] Bernadó, P. & Svergun, D. I. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Mol. BioSyst. 8, 151-167 (2012).

[15] Bernadó, P., & Svergun, D. I. (2012). Analysis of Intrinsically Disordered Proteins by Small-Angle X-ray Scattering. Intrinsically Disordered Protein Analysis, 107-122. *Intrinsically Disordered Protein Analysis*: Volume 2, Methods and Experimental Tools, Methods in Molecular Biology, vol. 896, Vladimir N. Uversky and A. Keith Dunker (eds.)

[16] Riback, J. A., Bowman, M. A., Zmyslowski, A. M., Knoverek, C. R., Jumper, J. M., Hinshaw, J. R., Kaye, E. B., Freed, K. F., Clark, P. L., & Sosnick, T. R. (2017). Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. Science, 358(6360), 238-241. https://doi.org/10.1126/science.aan5774

[17] Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J., & Holehouse, A. S. (2024). Direct prediction of intrinsically disordered protein conformational properties from sequence. Nature Methods, 21(3), 465-476. https://doi.org/10.1038/s41592-023-02159-5

[18] Bertoncini, C. W., Jung, Y.-S., Fernandez, C. O., Hoyer, W., Griesinger, C., Jovin, T. M., & Zweckstetter, M. (2005). Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein. Proceedings of the National Academy of Sciences, 102(5), 1430-1435. https://doi.org/10.1073/pnas.0407146102

[19] Mukrasch, M. D., Bibow, S., Korukottu, J., Jeganathan, S., Biernat, J., Griesinger, C., Mandelkow, E., & Zweckstetter, M. (2009). Structural Polymorphism of 441-Residue Tau at Single Residue Resolution. PLoS Biology, 7(2), e1000034. https://doi.org/10.1371/journal.pbio.1000034

[20] Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D., & Schuler, B. (2012). Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. Proceedings of the National Academy of Sciences, 109(40), 16155-16160. https://doi.org/10.1073/pnas.1207719109

[21] Schuler, B., Soranno, A., Hofmann, H., & Nettels, D. (2016). Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. Annual Review of Biophysics, 45(1), 207-231. https://doi.org/10.1146/annurev-biophys-062215-010915

[22] Schuler, B. Perspective: Chain dynamics of unfolded and intrinsically disordered proteins from nanosecond fluorescence correlation spectroscopy combined with single-molecule FRET. The Journal of Chemical Physics 149, 010901 (2018).

[23] Merchant, K. A., Best, R. B., Louis, J. M., Gopich, I. V. & Eaton, W. A. Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. Proceedings of the National Academy of Sciences 104, 1528-1533 (2007).

[24] Gomes, G.-N. W., Krzeminski, M., Namini, A., Martin, E. W., Mittag, T., Head-Gordon, T., Forman-Kay, J. D., & Gradinaru, C. C. (2020). Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. Journal of the American Chemical Society, 142(37), 15697–15710. https://doi.org/10.1021/jacs.0c02088

[25] Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature Methods, 11(8), 801-807. https://doi.org/10.1038/nmeth.3027

[26] Chong, L. T., Pitera, J. W., Swope, W. C., & Pande, V. S. (2009). Comparison of computational approaches for predicting the effects of missense mutations on p53 function. Journal of Molecular Graphics and Modelling, 27(8), 978-982. https://doi.org/10.1016/j.jmgm.2008.12.006

[27] 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

[28] 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).

[29] 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).

[30] Seth, S., Stine, B., & Bhattacharya, A. (2024). Fine structures of intrinsically disordered proteins. The Journal of Chemical Physics, 160(1). https://doi.org/10.1063/5.0176306

[31] Baul, U., Chakraborty, D., Mugnai, M. L., Straub, J. E. & Thirumalai, D. Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins. The Journal of Physical Chemistry B 123, 3462-3474 (2019).

[32] Baratam, K., & Srivastava, A. (2024). SOP-MULTI: A Self-Organized Polymer-Based Coarse-Grained Model for Multidomain and Intrinsically Disordered Proteins with Conformation Ensemble Consistent with Experimental Scattering Data. Journal of Chemical Theory and Computation, 20(22), 10179-10198. https://doi.org/10.1021/acs.jctc.4c00579

[33] 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).

[34] 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.

[35] Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J., & Holehouse, A. S. (2024). Direct prediction of intrinsically disordered protein conformational properties from sequence. Nature Methods, 21(3), 465-476. https://doi.org/10.1038/s41592-023-02159-5

[36] Tesei, G., Trolle, A. I., Jonsson, N., Betz, J., Knudsen, F. E., Pesce, F., Johansson, K. E., & Lindorff-Larsen, K. (2024). Conformational ensembles of the human intrinsically disordered proteome. Nature, 626(8000), 897-904. https://doi.org/10.1038/s41586-023-07004-5

[37] 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).

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

[39] 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).

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

[41] 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).

[42] Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., & DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. Journal of Molecular Biology, **195**(3), 659-685. https://doi.org/10.1016/0022-2836(87)90189-6

[43] 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. Proceedings of the National Academy of Sciences 107, 8183-8188 (2010).

[44] Das, R. K., & Pappu, R. V. (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proceedings of the National Academy of Sciences, 110(33), 13392-13397. https://doi.org/10.1073/pnas.1304749110

[45] Joseph, J. A., Reinhardt, A., Aguirre, A., Chew, P. Y., Russell, K. O., Espinosa, J. R., Garaizar, A., & Collepardo-Guevara, R. (2021). Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. Nature Computational Science, **1**(11), 732-743. https://doi.org/10.1038/s43588-021-00155-3

[46] Borgia, A., Borgia, M. B., Bugge, K., Kissling, V. M., Heidarsson, P. O., Fernandes, C. B., Sottini, A., Soranno, A., Buholzer, K. J., Nettels, D., Kragelund, B. B., Best, R. B., & Schuler, B. (2018). Extreme disorder in an ultrahigh-affinity protein complex. Nature, **555**(7694), 61-66. https://doi.org/10.1038/nature25762

[47] Sottini, A. *et al.* Sottini, A., Borgia, A., Borgia, M. B., Bugge, K., Nettels, D., Chowdhury, A., Heidarsson, P. O., Zosel, F., Best, R. B., Kragelund, B. B., & Schuler, B. (2020). Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. Nature Communications, bf 11(1). https://doi.org/10.1038/s41467-020-18859-x

[48] Farag, M., Borcherds, W. M., Bremer, A., Mittag, T., & Pappu, R. V. (2023). Phase separation of protein mixtures is driven by the interplay of homotypic and heterotypic interactions. Nature Communications, **14**(1). https://doi.org/10.1038/s41467-023-41274-x

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

[50] 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).

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

[52] Sawle, L. and Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. J. Chem. Phys. 2015, **143**, 085101.

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

[54] Wenwei Zheng, Gregory Dignon, Matthew Brown, Young C. Kim, and Jeetain Mittal, J. Phys. Chem. Lett. 2020, **11**, 3408-3415.

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

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

[57] Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization (Version 9). arXiv. https://doi.org/10.48550/ARXIV.1412.6980

[58] 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

[59] 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.

[60] Chao, T.-H., Rekhi, S., Mittal, J., & Tabor, D. P. (2023). Data-driven models for predicting intrinsically disordered protein polymer physics directly from composition or sequence. Molecular Systems Design & Engineering, **8**(9), 1146-1155. https://doi.org/10.1039/d3me00053b

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

[62] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., (2014). Generative adversarial nets: Proceedings of the 28th International Conference on Neural Information Processing Systems - volume 2. Guide Proceedings. https://dl.acm.org/doi/10.5555/2969033.2969125

[63] Vitalis, A., & Pappu, R. V. (2008). ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. Journal of

Computational Chemistry, **30**(5), 673-699. Portico. https://doi.org/10.1002/jcc.21005

[64] https://www.ncbi.nlm.nih.gov/clinvar/

[65] Marsh, J. A., & Teichmann, S. A. (2023). Predicting pathogenic protein variants. Science, **381** (6664), 1284-1285. https://doi.org/10.1126/science.adj8672

[66] Chakraborty, D., Straub, J. E., & Thirumalai, D. (2020). Differences in the free energies between the excited states of Aβ40 and Aβ42 monomers encode their aggregation propensities. Proceedings of the National Academy of Sciences, **117** (33), 19926-19937. https://doi.org/10.1073/pnas.200257011

[67] Kulkarni, P., Leite, V. B. P., Roy, S., Bhattacharyya, S., Mohanty, A., Achuthan, S., Singh, D., Appadurai, R., Rangarajan, G., Weninger, K., Orban, J., Srivastava, A., Jolly, M. K., Onuchic, J. N., Uversky, V. N., & Salgia, R. (2022). Intrinsically disordered proteins: Ensembles at the limits of Anfinsen's dogma. Biophysics Reviews, **3**(1). https://doi.org/10.1063/5.0080512