SCIENTIFIC REPORTS
natureresearch

OPEN

The Order-Disorder Continuum: Linking Predictions of Protein Structure and Disorder through Molecular Simulation

Claire C. Hsu[1], Markus J. Buehler[2] & Anna Tarakanova[3,4*]

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions within proteins (IDRs) serve an increasingly expansive list of biological functions, including regulation of transcription and translation, protein phosphorylation, cellular signal transduction, as well as mechanical roles. The strong link between protein function and disorder motivates a deeper fundamental characterization of IDPs and IDRs for discovering new functions and relevant mechanisms. We review recent advances in experimental techniques that have improved identification of disordered regions in proteins. Yet, experimentally curated disorder information still does not currently scale to the level of experimentally determined structural information in folded protein databases, and disorder predictors rely on several different binary definitions of disorder. To link secondary structure prediction algorithms developed for folded proteins and protein disorder predictors, we conduct molecular dynamics simulations on representative proteins from the Protein Data Bank, comparing secondary structure and disorder predictions with simulation results. We find that structure predictor performance from neural networks can be leveraged for the identification of highly dynamic regions within molecules, linked to disorder. Low accuracy structure predictions suggest a lack of static structure for regions that disorder predictors fail to identify. While disorder databases continue to expand, secondary structure predictors and molecular simulations can improve disorder predictor performance, which aids discovery of novel functions of IDPs and IDRs. These observations provide a platform for the development of new, integrated structural databases and fusion of prediction tools toward protein disorder characterization in health and disease.

Intrinsically disordered proteins (IDPs) make up 35 to 45% of proteins contained within eukaryotes, and sequences with an IDR (intrinsically disordered region) longer than 30 residues occur twice as frequently in eukaryotic proteins than in sets of randomly selected proteins[1,2]. Disordered regions fulfill a variety of functions: short linear motifs play a role in targeting for post-translational modifications or cell signaling[3–5] and longer regions promote molecular recognition and protein-protein interactions[6,7], among others. IDPs and IDRs can serve as flexible linkers between structured regions or as flexible binding sites for ligands[6]. Some IDPs undergo a disorder-order transition upon binding to other proteins through molecular recognition features (MoRFs), amphipathic regions within longer disordered regions[6,8]. The nature of the disordered regions is key to the resulting function of the protein: the length of the disordered region, the amount of disorder, and the specific location of the disordered regions all influence the functional role of the protein[6].

Biological implications of IDPs and IDRs range from cell signaling to cell cycle control[6,9]. IDPs and IDRs play a role in numerous diseases, examples including the tau protein in Alzheimer's[10], aggregate proteins in Parkinson's disease[11], and several driver proteins and prion-like regions in neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS)[12–14]. Recent studies have also connected structural disorder to drug design applications[15], as characterization of the dynamics of a disease-associated IDP may guide ligand selection during

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Laboratory for Atomistic and Molecular Mechanics, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Mechanical Engineering, University of Connecticut, Storrs, CT, USA. [4]Department of Biomedical Engineering, University of Connecticut, Storrs, CT, USA. *email: anna.tarakanova@uconn.edu

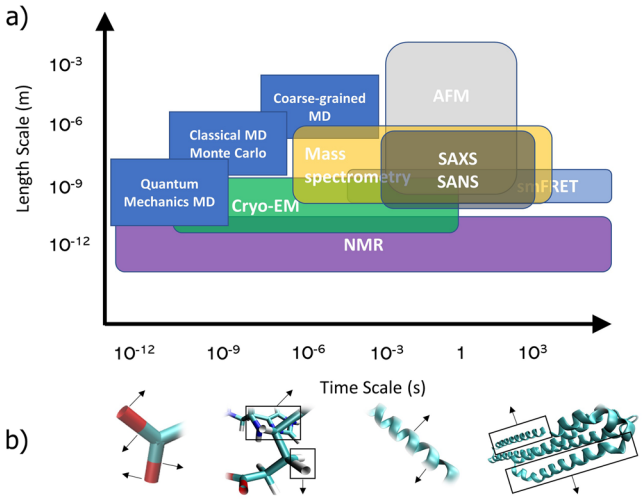

Figure 1. (**a**) Experimental and simulation techniques used to define protein structure and dynamics at different time and length-scales. ("MD" – molecular dynamics; "AFM" – atomic force microscopy; "SAXS" – small angle X-ray scattering; "SANS" – small angle neutron scattering; "EM" – electron microscopy; "NMR" – nuclear magnetic resonance spectroscopy; "smFRET" – single molecule fluorescence resonance energy transfer) (**b**) Movement at different length-scales (bonds, side chains, residues, and domains) that can be characterized. Visualization with VMD[103].

drug development, and have identified the role of disorder in enzymic function[16]. The strong link between protein function and protein disorder motivates a deeper and more fundamental characterization of IDPs and IDRs for discovering new functions and relevant mechanisms. However, the structure of IDPs and IDRs and associated functions remain hard to detect - these proteins tend to evolve faster than structured proteins at the sequence level, so there is less functional information to derive from homologues[6,17–20]. In addition, many current experimental techniques fail to accurately characterize IDPs and IDRs due to their dynamic nature[21,22]. Some techniques also have resolution or timescale constraints (as illustrated in Fig. 1), which can affect the ability to capture disorder on residue length scales or longer timescales[23,24].

Characterizing protein structure and disorder. A variety of experimental techniques (as detailed in Fig. 1) are used to characterize the structure of proteins, with variable applicability to rigid and flexible proteins: some methods can capture conformational transitions of IDPs and IDRs while others fail to describe dynamics at all. Some methods used to characterize protein structure include X-ray crystallography, NMR spectroscopy, mass spectrometry (MS) techniques, electron microscopy, and small-angle X-ray scattering (SAXS).

X-ray crystallography is one of the most commonly used techniques for structural characterization of proteins found in the PDB (Protein Data Bank)[25], suitable for proteins that can be successfully bound to ordered crystals. However, X-ray crystallography generally fails to determine the structure of dynamic regions[22], which leads to regions of missing electron density in resolved protein structures.

Recent advances in NMR spectroscopy have contributed to the characterization of protein ensembles with increasing resolution[26–28]. NMR spectroscopy can successfully capture the dynamics of protein structures[28], and integrative models now combine different techniques with NMR to more accurately characterize dynamic features of a protein[29–33]. Recent advances have proposed kinetic protein crystallography, combining high resolution static imagery from X-ray crystallography with lower resolution 3-D structural ensembles from NMR spectroscopy, to provide an improved description of protein structure than either method individually[29,30]. Other methods propose coupling NMR with molecular dynamics simulations to capture conformational heterogeneity of proteins[31–33].

Mass spectrometry can be used to capture conformational intermediates[34]. For example, ion-mobility mass spectrometry (IM-MS) with electrospray ionization uses the resulting charge state distribution to determine conformations and disorder[34–36]. Hydrogen/deuterium-exchange mass spectrometry (HDX-MS) captures the dynamics of IDPs well[37–39], as protein conformation affects rates of exchange, especially in cases when other methods fail to characterize highly disordered regions[22]. HDX-MS is notably useful in describing regions of protein-protein interaction, where IDPs may undergo binding[40].

Recent advances in single-particle electron cryo-microscopy (CryoEM), where multiple protein conformations can be isolated, have generated images of disordered proteins with up to 4 Å resolution[41–43]. Small-angle X-ray scattering (SAXS) and small-angle neutron scattering (SANS) can characterize flexible IDPs and IDRs and determine protein compactness[44,45], and combined with other high-resolution techniques like NMR, small-angle scattering techniques can derive structural information on multiple length-scales[24].

Single molecule fluorescence techniques such as single molecule Förster resonance (smFRET) have helped to describe protein ensembles by capturing long-range transitions between IDP and IDR configurations[23,46]. While newer methods may contribute accurate characterizations of multiple IDP and IDR conformations, models may

still conflict with one another, which has led to a growing number of studies adopting integrative methods, utilizing multiple techniques to generate models at multiple resolutions[47].

Cataloguing disorder. Functional proteins may exist in numerous conformations: the Protein Quartet model proposes solid and ordered, liquid-like and disordered, gas-like extended disordered, and pre-molten globule disordered states, which suggests different levels of disorder linked to the various IDP and IDR functions[48,49]. Yet many existing definitions of disorder used in training modern disorder predictors or classifying regions in disorder databases still utilize a polarized ordered or disordered designation[50–56]. Proteins generally lie on an order-disorder continuum[21], a description which recognizes that there may be significant intermediate stages that can have functional implications. To capture this, structural ensembles of proteins are often characterized, capturing different folded or unfolded states of a protein and its dynamic motion, through a growing arsenal of NMR techniques[21]. However, large structure databases such as the PDB still lack an extensive number of such ensembles, and its experimental data often fails to report multiple protein conformations. NMR, while better suited for IDPs and IDRs, still sometimes fails to assign quickly fluctuating disordered regions[22]. Newer databases specific to storing protein ensembles, such as the Protein Ensemble Database (PED)[57] and the Protein Order and Disorder Database (PODD)[21], and databases specific to storing protein disorder annotations, such as DisProt[58], are also much smaller in size than the PDB, on an order of 100 to 1,000 times – for instance, the PED contains around 24 protein ensembles and the PODD contains over 5000 two-dimensional protein ensembles. DisProt catalogs over 2,000 disordered regions. Thus, there exists a disconnect between the state of protein databases and the order-disorder continuum that can capture the full spectrum of protein structure and dynamics.

To better address the state of disorder prediction, below we evaluate the strengths and limitations of currently available structure prediction methodology. Secondary structure prediction has evolved with the growth of machine learning and predictive algorithms, whose varying performances may be leveraged to complement disorder predictions.

A brief history of protein secondary structure prediction. The solution to predicting protein secondary structure has evolved quickly over the past few decades, making strides in optimizing input features and model architecture. Early predictors utilized neural networks[59–62] and support vector machines[63], with Rost and Sander[64] and Zvelebil *et al.*[65] as some of the first to use multiple sequence alignments as an input feature into their neural network. Jones then introduced PSI-BLAST output matrices, which contain sequence conservation information based on similar amino acid sequences, as a new input feature in PSI-PRED[66], setting the precedent for nearly all future predictors. Cuff and Barton then combined Hidden Markov Model (HMM) profiles with the PSI-BLAST profile to create the Jnet prediction method, which uses two artificial neural networks[67]. Other recent work used machine learning to capture structural features of proteins and applied it to the design of new protein sequences[68].

These early predictors formed the foundation for many modern predictors, as research has accelerated in recent years. Input had largely been considered on a residue-by-residue basis, until the idea of utilizing local and global contexts within protein sequence arose. Convolutional neural networks (CNNs) provide more information on surrounding sequence and structure by using a sliding window to capture a fuller local context[69]. Compared to unmodified position-specific scoring matrix (PSSM) profiles, CNN features as inputs have improved prediction accuracy by up to 4%[70].

Recurrent neural networks (RNNs) have also improved prediction accuracy by taking global sequence context and nonlocal interactions into account. Other models attempt to capture more information around each residue by using sliding windows containing surrounding regions for each residue, but RNNs can retain information from any part of the previously seen sequence. To solve the disappearing or exploding gradient problem often found with RNNs, models have adopted gate and memory structures, such as GRUs[71] and long short-term memory (LSTM) units[72,73]. Many models built to predict protein structure are also bidirectional recurrent neural networks (BRNNs), traversing the sequence in the forward and backward direction. Examples include SSPRO[74], an ensemble of 100 BRNNs, and SPIDER3[73], a combination of two LSTM BRNNs and two fully connected layers.

To address dependencies on adjacent secondary structure labels, many models also infer secondary structure from nearby secondary structures, in addition to local or global PSI-BLAST profile patterns. Baldi *et al.*[74] introduced a template-based method, which uses secondary structure of homologous proteins as a template for prediction of other structures. Conditional neural fields[75] can also take advantage of surrounding labels, or surrounding secondary structure, to influence prediction of other labels.

Since each method provides a unique piece of information (e.g. local context, global context, nearby structure dependencies), many models also combine different methods. For example, DeepCNF[76] utilizes deep convolutional neural networks instead of shallow neural networks in its conditional neural field. Li and Yu combined multiscale CNNs with stacked BGRUs to learn both multiscale local contexts and nonlocal interactions[71].

Most predictors use PSSM profiles as input features derived from a nonredundant subset of the PDB, such that the subset maintains some level of sequence dissimilarity in the form of percentage identity cutoff. Additional inputs include the raw amino acid sequence (in one-hot encoding), Hidden Markov Model (HMM) profiles, and other physio-chemical annotations or properties of the protein. Common benchmark protein sets include CASP[77] sets or CB513[67], which also remain under some sequence identity cutoff from the training set.

In 2001, Rost theorized a limit on prediction accuracy of 88%[78], accounting for state-of-the-art predictor performance (PSIPRED, JPred2) at the time. Reasons for such a limit include limitations and ambiguities on structure determination through X-ray or NMR methods, limitations and hard-coded threshold values in assignment algorithms (such as in the DSSP[79] algorithm), and, of particular relevance to this study, the dynamic nature of protein structure.

The above predictors utilize static structure information from the PDB, which creates a disconnect between structure predictors and protein disorder. Such a disconnect may still be leveraged to provide additional disorder information, especially in the case of varying disorder definitions and predictive techniques, as we will demonstrate.

Predicting Protein Disorder. To address the challenge of protein disorder prediction, several disorder predictors have been created, all of which take advantage of different sources of data. Disorder has long been characterized as the absence of atomic coordinate information in native structures determined by X-ray crystallography due to the flexibility of the protein in that region, deeming it invisible in crystallographic electron density maps. As a result, many predictors of disorder in IDPs and IDRs focus on labeling regions of missing electron density as regions of disorder[50,51,80]. As discussed above, a number of techniques has been effective in detecting disorder to develop structural ensemble datasets, thereby capturing conformational variability and disorder in flexible proteins. More recently, annotations taken from disorder databases accumulate such experimental data. These varying definitions of disorder, however, affect the performance and transitivity of predictors. As a result, the current individual definitions of disorder may have to be expanded or combined to account for the diverse functionalities of IDPs and IDRs and the order-disorder continuum.

Disorder predictor development ranges from deterministic biophysical models to trained machine learning algorithms. Predictors such as IUPred[81] utilize existing structural data for a rule-based system that predicts disorder in a novel protein given its sequence, usually by aiming to minimize energy or favoring specific amino acid pairs over others. IUPred specifically uses an energy estimation method, where pairwise potentials are determined between pairs of amino acids, and the model parameters are derived only from globular proteins from the PDB. As in the case of secondary structure predictors, machine learning methods have become increasingly popular, as seen with the development of predictors such as DISOPRED3 (neural network)[51], DisEMBL (neural network)[82], PONDR-VSL2B (SVM)[54], and DeepCNF-D (conditional neural field)[50]. Combining different types of methods results in meta-predictors, such as PONDR-FIT[83] or GSmetaDisorder3D[84], which take advantage of multiple predictors specialized in identifying different types of disordered regions (e.g. length) or trained using different methods (e.g. energy functions, machine learning).

In addition to inconsistent definitions of disorder, the difficulty of building a disorder predictor also lies in the intrinsic relationship between disordered region length and protein function[85]. Specifically, short IDR identification remains an integral part of disorder prediction, as these regions can act as motifs or serve as linkers. However, there has often been a disparity between short disordered region prediction ($<$10 AA) and long region prediction using general disorder predictors. Due to biases in training data or feature selection[54], predictors trained on longer regions of disorder tended to perform more poorly while predicting short disordered regions[86], so to compensate, some predictors train separately on datasets of different disordered region length. For instance, the PONDR predictor family contains separate predictors for long ($>$30AA) and short regions (VSL2-L and VSL2-S, respectively). Both achieve an overall accuracy of over 80% for their respective length regions[54]. However, for general length-independent predictors, shorter regions tend to yield higher prediction accuracy than longer regions[55,87].

Databases of protein disorder. Recent interest in IDPs and IDRs as important functional proteins has sparked the development of disordered protein databases, which often combine multiple experimental techniques. As mentioned previously, DisProt, the PED, and the PODD all catalog a number of disordered regions or protein ensembles, leveraging experimental techniques such as small-angle X-ray scattering, NMR, and SAXS.

With the increase in disorder predictor performance, some databases now combine predicted disorder annotations with any available experimental information. For instance, MobiDB 3.0 accumulates information from DisProt, UniProtKB and FuzDB for general disorder annotation, as well as disorder predictions from IUPred, VSL2b, DisEMBL, and other disorder predictors in its three-layered annotation scheme[53]. MobiDB contains disorder information from predictions for over 80 million proteins. D^2P^2[2,52] similarly utilizes multiple predictors, including IUPred, VSL2b, and Espritz predictors, and combines them into a disorder agreement metric. It contains prediction information on over 10 million proteins. A comprehensive review of disorder predictors and databases can be found in He (2009)[9] and Meng (2017)[1].

As these databases were more recently developed (D^2P^2 first developed in 2012, MobiDB in 2012 and DisProt in 2006) compared to protein databases for traditional folded protein structure, the number of proteins with experimental, manually curated information in these databases (rather than predicted data) is not yet on the scale of databases such as the PDB or UniProt, which contain over 100,000 structures. This disparity leads to a lack of catalogued disorder information which could provide insight into novel protein functions.

Leveraging structural data for disorder predictions. Currently, static structural data has limited use for disorder prediction because disorder inherently relies on the dynamic nature of proteins. Here we explore novel approaches to leverage static structural data and structure prediction algorithms for extracting additional disorder information that existing disorder predictors fail to capture, through molecular simulation and structure prediction error.

Disordered regions within folded proteins can be used to train disorder databases. The structure predictor itself can give us information about disorder within the static structure. Because structure predictors are often trained on evolutionary data to recognize structure from homologous proteins, disordered regions will inherently perform worse because disordered regions are known to evolve at a faster pace. In addition, structure predictors are also generally trained on static structural information, which fails to capture dynamics and flexibility of a protein, especially in disordered regions.

A recent study evaluated 26 different disorder predictors, which demonstrated large variability in disorder predictions[56]. This variability could be attributed to the different inherent definitions of disorder, different training datasets, or different predictor specialties, as detailed earlier. More concerning, however, is the discovery of the under-prediction of disorder in many disorder predictors – for instance, DISOPRED3, a predictor used in this study, tended to bias towards ordered labels. This disparity calls for additional sources of data that can be leveraged for disorder information in the case where disorder predictors fail.

Recent work has studied the link between structure and disorder[21,28], through the development of the s2D method, for example, which concurrently predicts disorder and structure using NMR spectroscopy. In this work, we look at the link between secondary structure and disorder predictor results, exploring these results in comparison to dynamic protein fluctuation signatures determined from molecular simulation of sample proteins. We highlight regions of high flexibility revealed by molecular simulations that disorder predictors fail to capture. We find, consistently across all types of prediction methods, that areas of poor structure predictor performance may suggest high flexibility or disorder.

Methods

Experiments were conducted using CullPDB[88], a set of 11,154 proteins sharing no more than 25% sequence identity. This dataset was split into training and test sets by isolating 15% (1673 proteins) of the original set as the test set. The CullPDB derived training set was further filtered to remove any sequences sharing more than 40% identity with any protein in the test sets. 8-class labels to represent protein secondary structure were generated using DSSP[79], with missing DSSP labels assigned as coiled residues.

To look at samples with a varied structure content, a non-redundant test set of 1673 proteins was split into 5 bins of increasing helix (including 3_{10}, α, π helices) and beta (beta strand and bridges) content. The helix/beta content was quantified by counting chains of consecutive amino acids of either helical or beta secondary structure based on DSSP, exponentially increasing with chain length. Normalization of the score was performed by dividing total content score by overall sequence length. Chain lengths were kept within 200 ± 20 residues.

We considered three disorder predictors with different input data and training methods: IUPred (long)[81], DISOPred[51] and DISOclust[89]. IUPred uses physical properties of amino acid pairs in the sequence to determine order/disorder. DISOPred3 uses a consensus method between DISOPred2 and an additional SVM classifier for protein binding. DISOclust utilizes variability in predictions from ModFOLD2clust, which compares 3D models of a protein, to judge disorder.

Four secondary structure predictors are also considered: DeepCNF[76], SPIDER3[73], SSPRO8[74], and 2D-CNN[70]. DeepCNF and the 2D-CNN are both machine learning-based models, using convolutional neural nets trained on PSSM matrices, while SPIDER3 uses a bidirectional recurrent neural net with LSTM cells trained on PSSM matrices, in addition to the HMM profile and physio-chemical properties of the amino acid sequence. SSPRO8 also utilizes a bidirectional recurrent neural network but takes in structural similarity as an additional parameter. These models were trained on culled versions of the PDB, employing a cutoff sequence identity to eliminate redundancy in training sets. To evaluate sampled proteins, the web servers for the first three predictors were used, and 2D-CNN was implemented locally as described in Supplemental Information online. A summary of the secondary structure and disorder predictors used in the study is included in Supplementary Table S1.

Molecular models of five representative protein structures with PDB IDs 3PLW[90], 2R6V[91], 1DZF[92], 3HZ8[93], 3UMH[94] were considered. It should be noted that despite the use of MD for identifying regions of disorder, there are weaknesses associated with the method, including incomplete sampling and dependence on starting structure. Residues missing from the PDB file were excluded from the simulation and results. All simulations were carried out using GROMACS version 5.1.2[95]. Each structure was placed into a rectangular water box with periodic boundary conditions. The CHARMM27 force field was used, which includes CHARMM22 and CMAP for proteins[96]. While CHARMM36m could have been used to provide improved correlation of the data, the less computationally expensive CHARMM22/CMAP force field was sufficient to demonstrate a proof of concept, as we are able to capture fluctuations and correlated (short) disordered regions. Each molecule was fully solvated using the TIP3P water model[97] and neutralized by adding the appropriate number of chloride counter ions. Frames were saved every 2 ps for analysis. Each structure was first minimized through the steepest descent algorithm to ensure no steric clashes. Then, each structure was simulated for 1 μs in an NVT ensemble at 310 K. The time step used was 2 fs. The Berendsen thermostat[98] was used for temperature coupling. The LINCS[99] algorithm was used to constrain covalent bonds with hydrogen atoms. The short-range electrostatic interactions and Lennard-Jones interactions were evaluated with a cutoff of 10 Å. Particle-mesh Ewald summation[100] was used to calculate long-range electrostatic interactions with a grid spacing of 1.6 Å and a fourth order interpolation.

A total of 1500 frames was extracted out of each 1 μs simulation. The last 300 ns of each simulation was sampled to extract a molecular longevity metric. For each residue in each sample, structural longevity was measured as the average duration during which the secondary structure DSSP assignment remained constant, with a score of 1 equivalent to a constant structure through the 300 ns, and a score nearing 0 indicating structural fluctuation within every 200 ps range. Intermediate values were determined by considering stability of a DSSP assignment at a given residue, increasing exponentially with longer durations of consistent DSSP label. The longer the assignment remained consistent, the higher the 0–1 value assigned.

$$\text{Longevity at residue} = \frac{avg(\textit{number frames with consistent DSSP label})}{\textit{total frames}} \quad (1)$$

We used in-house TCL and Matlab scripts to perform all analysis. All simulations were completed using the Extreme Science and Engineering Discovery Environment (XSEDE).

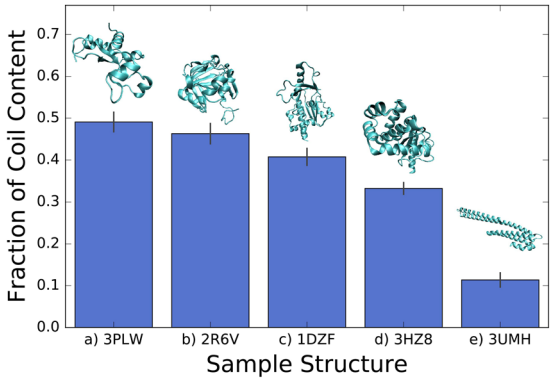

Figure 2. % coil content vs order counter for sampled molecules, 3PLW[90] (**b**) 2R6V[91] (**c**) 1DZF[92] (**d**) 3HZ8[93] (**e**) 3UMH[94]. The five proteins are ordered from greatest to least coil content.

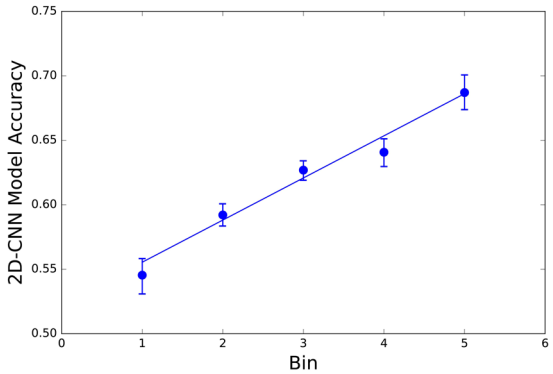

Figure 3. Prediction accuracy for all proteins in each of five bins (the test set split into bins of differing helix/beta content) in increasing order, using 2D-CNN[70] as implemented in Supplemental Information. Confidence interval = 0.95.

Results

In this study, we first split protein samples – a non-redundant test set of 1673 proteins – content into bins of increasing disorder, where ordered structure is defined as helix (including 3_{10}, α, π helices) and beta structures, and evaluate different predictive model performance on each bin. We then display the predictor performance alongside disorder predictions and molecular longevity data, identifying key regions of correlation and disagreement. Figure 2 evaluates content of sampled representative proteins from each test set, Figures 3 and 4 study structure predictor performance, and Figures 5–7 closely examine the structure and disorder of each sampled protein across its sequence. We first define the level of disorder in a protein by its secondary structure content. When dividing the test set into bins, we separate protein sets based on a decreasing coil content and corresponding increasing helix/beta content. 1672 proteins in the test set were divided into five bins of size 214, 452, 508, 233, and 265 proteins, in increasing helix/beta content. In general, increased coil content tends to align with increased disorder and associated flexibility in a protein[50,101], so our test set separation corresponds to an increasing degree of disorder. Figure 2 captures the increasing helix and beta content and corresponding decreasing coil content across the five sampled proteins, one from each test bin. Bin level is assigned according to secondary structure content – bin 1 is represented by structure 3PLW, bin 2 by structure 2R6V, bin 3 by structure 1DZF, bin 4 by 3HZ8, and bin 5 by 3UMH.

Structure prediction model performance was then evaluated for all proteins in each of the five bins. Cumulatively, the convolutional neural net model proposed by Li *et al.*[71] predicts secondary structure for proteins in the last bin of protein samples (highest order) with 15% higher accuracy than proteins in the first bin (least order) (bin 5 compared to bin 1 in Fig. 3). Prediction accuracy is directly correlated to the degree of disorder in the molecule.

Across another four published methods with varying machine learning techniques, more accurate predictions are found for increasingly ordered proteins (Fig. 4). Note that Fig. 4 depicts prediction accuracy for sampled proteins from each bin while Fig. 3 shows average model (2D-CNN) accuracy for all proteins in each bin. We find that accuracy increases from the most disordered to least disordered proteins in both cases. JPred4 and SPIDER3 both utilize Q3 (3-state secondary structure, with states coil, beta, and helix) DSSP labels, which could account for higher performance than the more specified Q8 (8-state secondary structure, with states 3_{10} helix, α helix, π helix, beta bridge, extended strand, turn, bend, and loop) labels used by 2D-CNN and DeepCNF. We note that SSPRO8

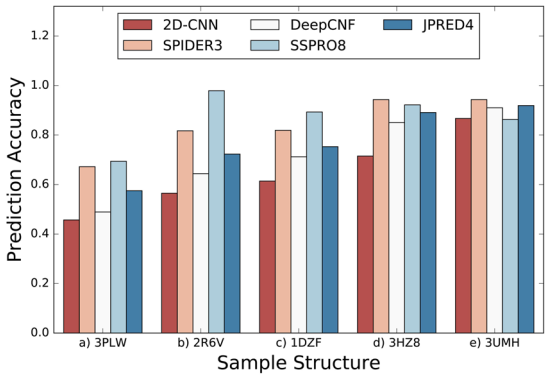

Figure 4. Prediction Accuracy of 5 samples (**a**) 3PLW[90] (**b**) 2R6V[91] (**c**) 1DZF[92] (**d**) 3HZ8[93] (**e**) 3UMH[94]), with increasing order (sampled proteins).

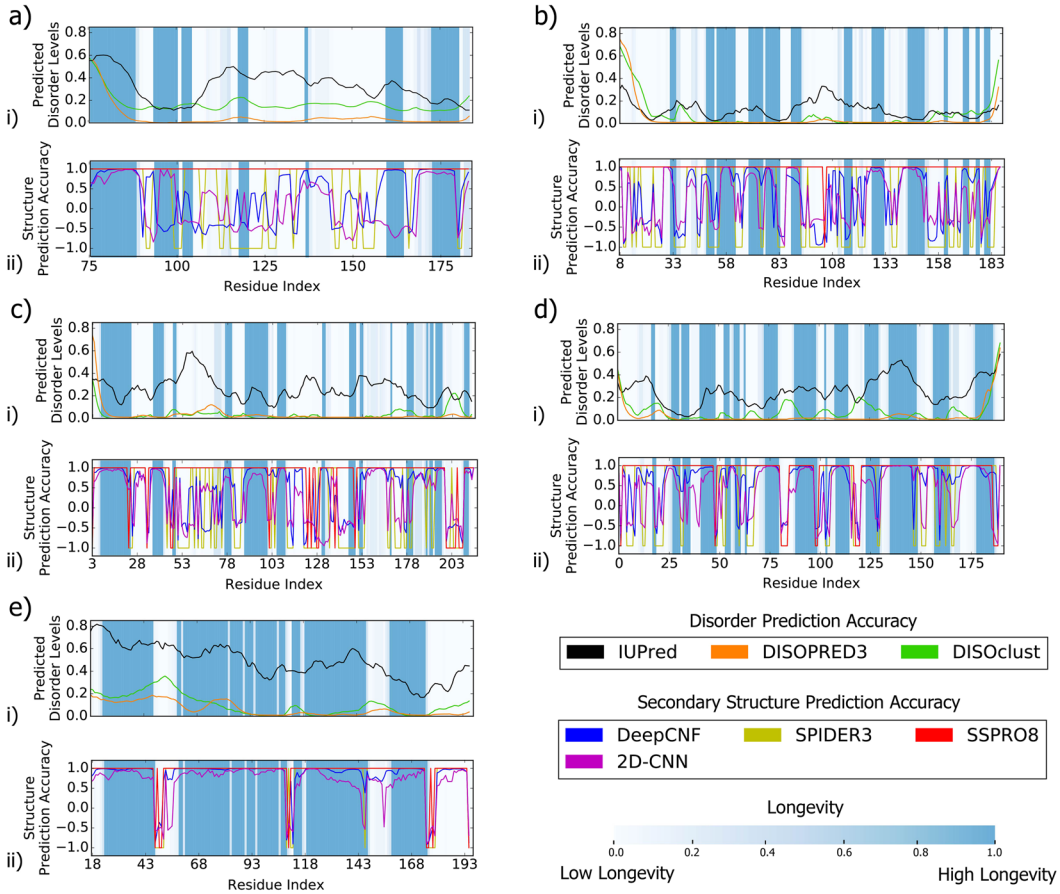

Figure 5. Predicted disorder with IUPred[81], DisoPRED[51], and DISOclust[89] predictors (i) and secondary structure prediction accuracy based on SPIDER3[73], DeepCNF[76], 2D-CNN[70], and SSPRO8[74] predictors (ii) highlighted with molecular structure longevity through molecular dynamics simulation (i, ii) for (**a**) 3PLW[90] (**b**) 2R6V[91] (**c**) 1DZF[92] (**d**) 3HZ8[93] (**e**) 3UMH[94]. For longevity, blue regions indicate higher longevity regions while white regions indicate lower longevity regions.

deviates from a linear trend as the model takes sequence-based structural similarity into account. However, all five predictors generally increase from a prediction accuracy of 60% to 80% from the most disordered to least disordered sampled proteins. We also note a direct correlation between prediction accuracy in Fig. 4 and the sample helix and beta content in Fig. 2.

We compare per-residue results from four secondary structure prediction algorithms: SPIDER3[73], DeepCNF[76], 2D-CNN[70], and SSPRO8[74] for five representative protein structures with PDB IDs a) 3PLW[90] b) 2R6V[91] c) 1DZF[92]

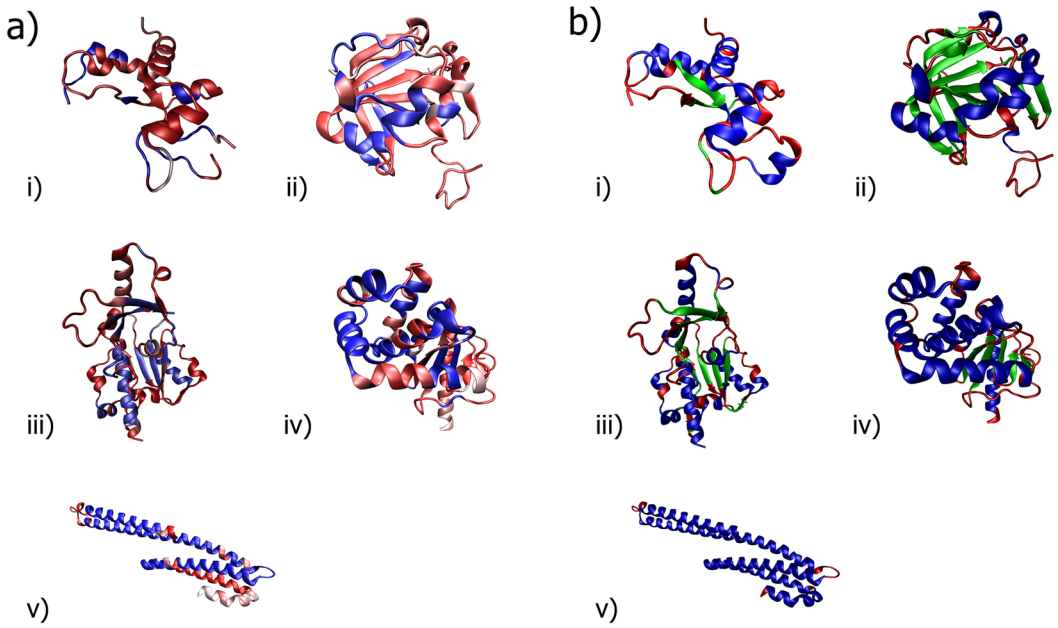

Figure 6. Models of the five sampled proteins in this study: (i) 3PLW[90] (ii) 2R6V[91] (iii) 1DZF[92] (iv) 3HZ8[93] (v) 3UMH[94]. (**a**) Molecular longevity on a red/blue scale for low/high structural longevity. (**b**) DSSP assignments with red (coiled), green (beta), and blue (helix) structures.

d) 3HZ8[93] e) 3UMH[94]. The top panel in Fig. 5(a–e) displays the per-residue predictions from three disorder predictors: IUPred (long), DisoPRED, and DISOclust for comparison. The bottom panel in Fig. 5(a–e) corresponds to the results of the four secondary structure predictors. The blue highlight in both panels in Fig. 5(a–e) displays the longevity of secondary structure based on molecular dynamics simulation results. Longevity is defined as average number of frames of consistent DSSP label divided by total number of frames, as in Eq. (1).

Accuracy of secondary structure prediction and degree of disorder is found to be consistently inversely correlated. The three disorder predictors considered show some correlation among themselves, but disagree in key regions (Fig. 5(a–e)(i)). DISOclust peaks tend to be more pronounced compared to more subtle peaks in IUPred or DISOPRED. For example, in Fig. 5(a)(i), the DISOclust peaks mirror those of DIOSPRED at residues 110–120 and 155–165 but have larger amplitudes, and are much more defined than the less clear peaks in IUPred predictions. As noted in Nielsen 2019[56], DISOPRED tends to under-predict disorder, which may explain its comparatively lower overall disorder predictions. At protein regions of high disorder, sample proteins demonstrate dips in secondary structure prediction accuracy and confidence based on all four structure predictors considered here (Fig. 5(a–e)(ii)). Model accuracy is represented by positive (correct) and negative (incorrect) values. Model per-residue confidence is determined using maximum class probability derived from n-class output as values and is represented on a 0 to +/−1 scale (farther from 0 suggesting higher confidence).

In regions of increased predicted disorder, especially around peaks predicted by DISOclust and DISOPRED3 (whose predictions have generally lower values among the three disorder predictors considered), SPIDER3 tends to predict less accurately (Fig. 5(a–e)(i-ii)). These peaks in disorder are likely significant, as those regions often also display significant peaks in IUPred, which suggests consensus. For instance, in Fig. 5(a)(i), peaks in disorder from residue 110 to 120 translate to a band of low secondary structure prediction accuracy (Fig. 5(a)(ii)). Similar trends are found in Fig. 5(b)(i-ii) at residue 90 to 100 and Fig. 5(c)(i-ii) at residue 45 to 70. These regions are also consensus regions of disorder for three disorder predictors considered (Fig. 5(a-e)(i)). We also observe that molecular longevity (visualized in Fig. 6(a–b)(i–v)) tends to correlate well with structure predictor performance across all five sampled proteins. However, molecular longevity results highlight regions not well-identified by disorder predictors but better correlated with structure predictor accuracy. For instance, Fig. 5(c)(i-ii) at residue 110 to 128 displays a band of poor SPIDER3 performance and low longevity, but no corresponding peak in disorder.

DeepCNF predictions display similar correlations to disorder predictions as do the SPIDER3 predictions (Fig. 5(a–e)(i-ii). Figure 5(b)(i-ii) at residue 133 to 145 displays weak DeepCNF prediction confidence (light red and blue colors suggest high uncertainty in the predicted DSSP label at these regions) and low longevity. Figure 5(c)(i-ii), from residue 110 to 140, also displays low structure predictor accuracy and longevity, but no significant peak in predicted disorder, other than a minor peak from IUPred. However, at consensus regions of peak disorder (as before), we find weak DeepCNF performance (e.g. Figure 5(b)(i-ii) res. 90–100, Fig. 5(c)(i-ii) res. 45–70, Fig. 5(e)(i-ii) res. 45–55). Interestingly, longevity and DeepCNF predictor accuracy tend to align with DISOclust predictions, even when there is no consensus disorder prediction at these regions.

2D-CNN prediction accuracy also aligns well with molecular longevity results, and peaks in disorder match with poor 2D-CNN performance and low longevity regions (Fig. 5(a–e)(i-ii)). We again observe key regions of poor 2D-CNN confidence or inaccuracy and low molecular longevity that do not have a corresponding disorder

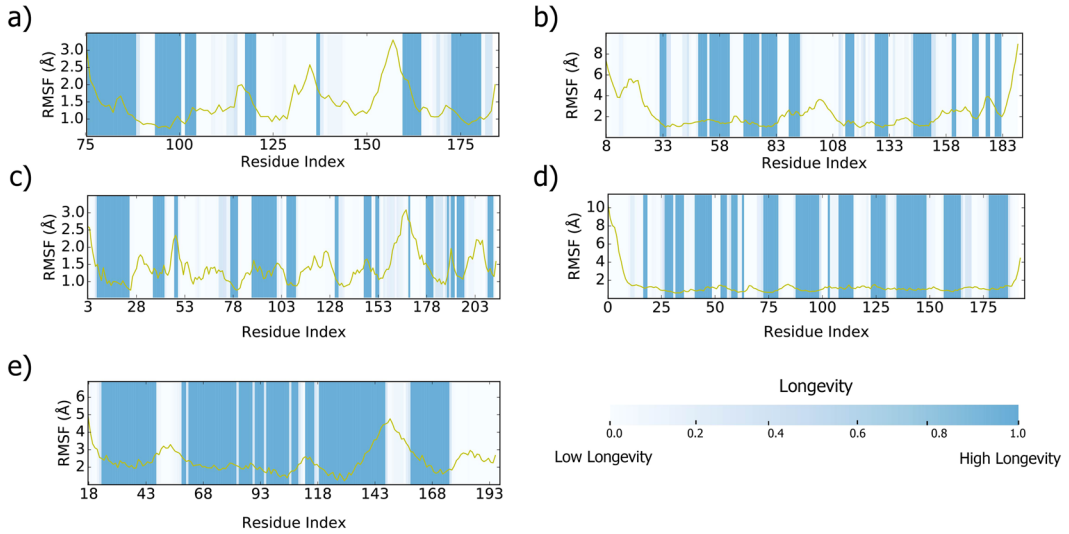

Figure 7. Root mean square fluctuation (RMSF) per-residue plots highlighted with molecular structure longevity through molecular dynamics simulation for (**a**) 3PLW[90] (**b**) 2R6V[91] (**c**) 1DZF[92] (**d**) 3HZ8[93] (**e**) 3UMH[94]. For longevity, blue regions indicate higher longevity regions while white regions indicate lower longevity regions.

consensus (e.g. Figure 5(c)(i-ii) res. 103–150, Fig. 5(a)(i-ii) res. 125–150). In addition, while 2D-CNN performance again aligns well with DISOclust and most DISOPRED results, it has mixed alignment with IUPred predictions. For instance, at some peaks, there is good correlation between IUPred and poor predictor performance (Fig. 5(b)(i-ii) res. 58–70), but at other regions, an IUPred peak with no consensus from the other two disorder predictors does not translate to poor structure predictor performance (Fig. 5(d)(i-ii) res. 130–145). Notably, some regions that dip suddenly in IUPred predicted disorder experience small periods of poor 2D-CNN predictor performance and short molecular longevity at the edges of such regions (e.g. Figure 5(a)(i-ii) res. 90–110, Fig. 5(b)(i-ii) res. 70–85).

SSPRO8 predictions contain far less error than the other three predictors, but around major consensus peaks such as the ones mentioned before, there are regions of poor SSPRO8 predictor performance (Fig. 5(a–e)(i-ii)). However, some key regions already highlighted do not demonstrate any dips in SSPRO8 predictor performance when the other three predictors did demonstrate dips (e.g. Figure 5(a)(i-ii) res. 110–120, Fig. 5(c)(i-ii) res. 53–70). As for the previous three structure predictors, SSPRO8 performance also dipped during bands of short molecular longevity.

Comparing prediction accuracy or confidence to per-residue structure longevity (Fig. 5(a–e)(ii) in molecular dynamics simulation shows that poorly predicted regions align with regions that display shorter average structural longevity (Fig. 6(a–b)(i–v)), which indicates more motion and flexibility in the region during molecular simulation. All predictors demonstrate this trend, either with predictors predicting incorrectly or with low confidence at these regions of highly dynamic motion. Combined with the correlation between disorder and structure predictions, we find many consensus peaks in disorder align with dynamic regions as determined by molecular simulation across all sampled protein models.

However, some key regions display a high degree of flexibility without a corresponding peak in disorder predictors. Areas with average structural longevity reaching zero suggest consistent fluctuation in the protein structure in simulation and should suggest a high degree of disorder. Despite this, some regions display low SS predictor performance and structural longevity, but no significant signal in any disorder predictor (e.g. Figure 5(c)(i-ii) res. 100–150). Other regions display varying SS predictor performance and low structural longevity with conflicting signals in disorder predictors (Fig. 5(a)(i-ii) res. 90–100).

Regions of higher longevity per-residue generally correspond to helix and beta structures while regions of lower longevity align with coiled regions, as labeled by DSSP (Fig. 6(b)(i–v)). We also observe that in increasing structural order, overall structure longevity also increases (Fig. 6(a–b)(i–v)).

Additionally, we confirm our longevity measure with RMSF per-residue plots (Fig. 7). Peaks and higher-value regions in RMSF generally align with low longevity regions, such as in Fig. 7(a), res. 135–150, Fig. 7(c), res. 50–70, and Fig. 7(e), res. 105–115.

Discussion

Labeled disordered regions within the structures considered, as determined by disorder predictors, all align with regions along the sequence that undergo more structural fluctuation, quantified as low molecular longevity, and poor secondary structure predictor performance. Yet, some regions with similar trends in longevity and secondary structure predictor performance do not have a corresponding significant peak in disorder. This disparity suggests that current disorder predictors may fail to capture some disordered regions with high molecular motion, especially in the case of globular proteins like the ones tested in this study. Furthermore, several regions did not reach a consensus on disorder levels among disorder predictors, which aligns well with variability results and

under-prediction of disorder found in Nielsen 2019[56]. However, as secondary structure predictors tend to predict incorrectly or predict with low confidence at these key regions, secondary structure predictor performance may be used as an additional marker of disorder that current disorder predictors fail to capture.

In our analysis, we find that trends in longevity and SS predictor performance align most closely with the disorder prediction from DISOclust. The DISOclust method is based off variation in residue positions in multiple fold recognition models, given the assumption that on a per-residue basis, residues that are more structurally aligned are more ordered. This characterization compares most closely with our molecular longevity definition, where high longevity corresponds to conserved secondary structure on a per-residue basis. Since the three disorder predictors root from three inherently different definitions of disorder, occasionally there is not a consensus on whether a region is predicted to be disordered.

We find consensus among secondary structure predictors: while the individual algorithms or machine learning method used are different, much of the input data and input format are similar, which results in similar outputs across predictors. Most structure predictors utilize a subset of the PDB or UniProt databases to train models, and they almost always use a PSSM format to better represent an "input sequence" for prediction. However, disorder predictors do not form a consensus in many regions considered, due to the number of different algorithm formats (deterministic vs learned) and input data sources (fold recognition data vs DisProt and PDB data). As a result, structure predictors can help identify and clarify disordered regions in which different predictors may not reach a consensus.

These trends exist for structures with varying structural content, as proteins were sampled from bins separated by helix/beta/coil content. Upon inspection of the protein chains (Fig. 6(a–b)(i–v)), many regions of low molecular longevity are coiled regions connecting more structured regions, possibly serving as flexible linkers. Compared to results shown in (Fig. 5(a–e)(i-ii)), these regions are also potentially disordered, suggesting a partial order-disorder continuum within even well-characterized globular proteins. Such regions may include residues 48 to 50 in 3UMH and residues 119 to 121 in 3HZ8, which connect helix regions (Fig. 6(b)(iv-v)) and display weak structural predictor performance and disorder peaks (Fig. 5(d,e)(i-ii)). For instance, molecular dynamics simulation can provide insight into short disordered regions within globular proteins that are difficult to identify but may have biological implications. In addition, molecular dynamics may help to extrapolate incorrectly assigned structures within disordered regions in globular proteins and clarify the propensity of structure or lack thereof within these regions. Extending this approach to existing IDPs and IDRs, this method can identify disordered regions not previously highlighted through other predictors.

Our longevity measure taken from molecular dynamics simulations matches closely with RMSF per-residue plots also derived from simulations (Fig. 7), which further supports the usage of molecular dynamics as an additional determinant of disordered regions in proteins. Both approaches characterize protein flexibility, notably in regions where existing predictors miss disordered regions, as mentioned previously. RMSF per-residue values can also help detail different types of high longevity regions. For instance, high longevity ordered regions would experience low fluctuation, but high longevity disordered regions (e.g. some coiled regions) would experience higher fluctuation. In the specific cases in this study, most coiled regions experienced low longevity – because our longevity measure accounted for Q8 labels, many of these regions alternated between coil and turn state, which are generally both considered coiled in other contexts.

Because secondary structure predictors are still largely based on static structural databases, incorrectly predicted regions may suggest higher degrees of flexibility and disorder. While current databases grow to include dynamic ensemble information, static structural information can still be leveraged to make conclusions about disorder, especially at disordered regions that current disorder predictors fail to identify. With this insight, future studies may consider utilizing such data to form more complete disorder predictions or database entries. As current databases take advantage of multiple sources for disorder prediction, the addition of molecular dynamics information would contribute to a more thorough analysis of a structure's disorder. As experimental techniques for disorder classification grow, using molecular dynamics and static structure predictor performance as an indicator of disorder can contribute to disorder determination to characterize the role of IDPs and IDRs in disease, cell signaling, and drug design.

Conclusion

We provide an overview of current experimental methods for the determination of IDPs and IDRs as well as the current state and shortcomings of disorder prediction. To contribute to the more accurate identification of disorder, we have presented secondary structure predictions and molecular longevity measurements as additional markers of disorder, especially in cases where existing disorder predictors fail to reach a consensus. Regions that are marked disordered by multiple predictors also experience poor secondary structure predictor performance and low per-residue structural longevity, but some regions that are marked disordered by only one or two predictors can be further clarified through molecular longevity data. This method can contribute to identification of disordered regions in proteins where disorder may be more subtle or under-predicted, as shown in the five sampled globular proteins, which can contribute to the identification of additional disordered regions key to biological functions.

Data availability

Structures used for molecular simulations are available in the Protein Data Bank, online at https://www.rcsb.org/. All predictors are available online as described in the corresponding references. The culled datasets used to conduct experiments are available at http://dunbrack.fccc.edu/Guoli/pisces_download.php.

Received: 24 July 2019; Accepted: 16 October 2019;
Published online: 07 February 2020

References

1. Meng, F., Uversky, V. N. & Kurgan, L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell. Mol. Life Sci.* **74**, 3069–3090, https://doi.org/10.1007/s00018-017-2555-4 (2017).
2. Xue, B., Dunker, A. K. & Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **30**, 137–149, https://doi.org/10.1080/07391102.2012.675145 (2012).
3. Davey, N. E. *et al.* Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281, https://doi.org/10.1039/c1mb05231d (2012).
4. Diella, F. *et al.* Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.* **13**, 6580–6603 (2008).
5. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29, https://doi.org/10.1038/nrm3920 (2015).
6. van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114**, 6589–6631, https://doi.org/10.1021/cr400525m (2014).
7. Mohan, A. *et al.* Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* **362**, 1043–1059, https://doi.org/10.1016/j.jmb.2006.07.087 (2006).
8. Wright, P. E. & Dyson, H. J. Linking folding and binding. *Curr. Opin. Struct. Biol.* **19**, 31–38, https://doi.org/10.1016/j.sbi.2008.12.003 (2009).
9. He, B. *et al.* Predicting intrinsic disorder in proteins: an overview. *Cell Res.* **19**, 929 (2009).
10. Uversky, V. N. Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. *Front. Aging Neurosci.* **7**, 18, https://doi.org/10.3389/fnagi.2015.00018 (2015).
11. Wu, K.-P., Weinstock, D. S., Narayanan, C., Levy, R. M. & Baum, J. Structural Reorganization of α-Synuclein at Low pH Observed by NMR and REMD Simulations. *J. Mol. Biol.* **391**, 784–796, https://doi.org/10.1016/j.jmb.2009.06.063 (2009).
12. Santamaria, N., Alhothali, M., Alfonso, M. H., Breydo, L. & Uversky, V. N. Intrinsic disorder in proteins involved in amyotrophic lateral sclerosis. *Cell. Mol. Life Sci.* **74**, 1297–1318, https://doi.org/10.1007/s00018-016-2416-6 (2017).
13. Kim, H. J. *et al.* Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nat.* **495**, 467–473, https://doi.org/10.1038/nature11922 (2013).
14. Uversky, V. N. Targeting intrinsically disordered proteins in neurodegenerative and protein dysfunction diseases: another illustration of the D(2) concept. *Expert. Rev. Proteom.* **7**, 543–564, https://doi.org/10.1586/epr.10.36 (2010).
15. Maity, B. K. *et al.* Spontaneous Fluctuations Can Guide Drug Design Strategies for Structurally Disordered Proteins. *Biochem.* **57**, 4206–4213, https://doi.org/10.1021/acs.biochem.8b00504 (2018).
16. Palombo, M. *et al.* The relationship between folding and activity in UreG, an intrinsically disordered enzyme. *Sci. Rep.* **7**, 5977, https://doi.org/10.1038/s41598-017-06330-9 (2017).
17. Bellay, J. *et al.* Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.* **12**, R14, https://doi.org/10.1186/gb-2011-12-2-r14 (2011).
18. Brown, C. J. *et al.* Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55**, 104–110, https://doi.org/10.1007/s00239-001-2309-6 (2002).
19. Chen, J. W., Romero, P., Uversky, V. N. & Dunker, A. K. Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *J. Proteome Res.* **5**, 888–898, https://doi.org/10.1021/pr060049p (2006).
20. Brown, C. J., Johnson, A. K., Dunker, A. K. & Daughdrill, G. W. Evolution and disorder. *Curr. Opin. Struct. Biol.* **21**, 441–446, https://doi.org/10.1016/j.sbi.2011.02.005 (2011).
21. Sormanni, P. *et al.* Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.* **13**, 339 (2017).
22. Balasubramaniam, D. & Komives, E. A. Hydrogen-exchange mass spectrometry for the study of intrinsic disorder in proteins. *Biochim. Biophys. Acta* **1834**, 1202–1209, https://doi.org/10.1016/j.bbapap.2012.10.009 (2013).
23. LeBlanc, S. J., Kulkarni, P. & Weninger, K. R. Single Molecule FRET: A Powerful Tool to Study Intrinsically Disordered Proteins. *Biomolecules* **8**, https://doi.org/10.3390/biom8040140 (2018).
24. Receveur-Brechot, V. & Durand, D. How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr. Protein Pept. Sci.* **13**, 55–75 (2012).
25. Gilliland, G. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242, https://doi.org/10.1093/nar/28.1.235 (2000).
26. Konrat, R. NMR contributions to structural dynamics studies of intrinsically disordered proteins. *J. Magn. Reson.* **241**, 74–85, https://doi.org/10.1016/j.jmr.2013.11.011 (2014).
27. Kosol, S., Contreras-Martos, S., Cedeño, C. & Tompa, P. Structural characterization of intrinsically disordered proteins by NMR spectroscopy. *Molecules* **18**, 10802–10828, https://doi.org/10.3390/molecules180910802 (2013).
28. Sormanni, P., Camilloni, C., Fariselli, P. & Vendruscolo, M. The s2D Method: Simultaneous Sequence-Based Prediction of the Statistical Populations of Ordered and Disordered Regions in Proteins. *J. Mol. Biol.* **427**, 982–996, https://doi.org/10.1016/j.jmb.2014.12.007 (2015).
29. Fenwick, R. B., van den Bedem, H., Fraser, J. S. & Wright, P. E. Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc. Natl Acad. Sci. USA* **111**, E445–454, https://doi.org/10.1073/pnas.1323440111 (2014).
30. van den Bedem, H. & Fraser, J. S. Integrative, dynamic structural biology at atomic resolution–it's about time. *Nat. Methods* **12**, 307–318, https://doi.org/10.1038/nmeth.3324 (2015).
31. Lindorff-Larsen, K., Best, R. B., Depristo, M. A., Dobson, C. M. & Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nat.* **433**, 128–132, https://doi.org/10.1038/nature03199 (2005).
32. Cavalli, A., Salvatella, X., Dobson, C. M. & Vendruscolo, M. Protein structure determination from NMR chemical shifts. *Proc. Natl Acad. Sci. USA* **104**, 9615–9620, https://doi.org/10.1073/pnas.0610313104 (2007).
33. Shen, Y. *et al.* Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl Acad. Sci. USA* **105**, 4685–4690, https://doi.org/10.1073/pnas.0800256105 (2008).
34. Stuchfield, D. *et al.* The Use of Mass Spectrometry to Examine IDPs: Unique Insights and Caveats. *Methods Enzymol.* **611**, 459–502, https://doi.org/10.1016/bs.mie.2018.09.038 (2018).
35. Galea, C. A. *et al.* Large-scale analysis of thermostable, mammalian proteins provides insights into the intrinsically disordered proteome. *J. Proteome Res.* **8**, 211–226, https://doi.org/10.1021/pr800308v (2009).
36. Beveridge, R., Chappuis, Q., Macphee, C. & Barran, P. Mass spectrometry methods for intrinsically disordered proteins. *Analyst* **138**, 32–42, https://doi.org/10.1039/c2an35665a (2013).
37. Zhou, J. *et al.* Conformational dynamics of 1-deoxy-d-xylulose 5-phosphate synthase on ligand binding revealed by H/D exchange MS. *Proc. Natl Acad. Sci. USA* **114**, 9355–9360, https://doi.org/10.1073/pnas.1619981114 (2017).
38. Zhu, S. *et al.* Hyperphosphorylation of intrinsically disordered tau protein induces an amyloidogenic shift in its conformational ensemble. *PLoS One* **10**, e0120416, https://doi.org/10.1371/journal.pone.0120416 (2015).
39. Oganesyan, I., Lento, C. & Wilson, D. J. Contemporary hydrogen deuterium exchange mass spectrometry. *Methods* **144**, 27–42, https://doi.org/10.1016/j.ymeth.2018.04.023 (2018).
40. Goswami, D. *et al.* Time window expansion for HDX analysis of an intrinsically disordered protein. *J. Am. Soc. Mass. Spectrom.* **24**, 1584–1592, https://doi.org/10.1007/s13361-013-0669-y (2013).
41. Cheng, Y. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **161**, 450–457, https://doi.org/10.1016/j.cell.2015.03.049 (2015).

42. Righetto, R. D., Biyani, N., Kowal, J., Chami, M. & Stahlberg, H. Retrieving high-resolution information from disordered 2D crystals by single-particle cryo-EM. *Nat. Commun.* **10**, 1722, https://doi.org/10.1038/s41467-019-09661-5 (2019).
43. Ketterer, P. *et al*. DNA origami scaffold for studying intrinsically disordered proteins of the nuclear pore complex. *Nat. Commun.* **9**, 902, https://doi.org/10.1038/s41467-018-03313-w (2018).
44. Riback, J. A. *et al*. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Sci.* **358**, 238–241, https://doi.org/10.1126/science.aan5774 (2017).
45. Kikhney, A. G. & Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* **589**, 2570–2577, https://doi.org/10.1016/j.febslet.2015.08.027 (2015).
46. Choi, U. B., Weninger, K. R. & Bowen, M. E. Immobilization of proteins for single-molecule fluorescence resonance energy transfer measurements of conformation and dynamics. *Methods Mol. Biol.* **896**, 3–20, https://doi.org/10.1007/978-1-4614-3704-8_1 (2012).
47. Ward, A. B., Sali, A. & Wilson, I. A. Biochemistry. Integrative structural biology. *Sci.* **339**, 913–915, https://doi.org/10.1126/science.1228565 (2013).
48. Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756, https://doi.org/10.1110/ps.4210102 (2002).
49. Sickmeier, M. *et al*. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* **35**, D786–793, https://doi.org/10.1093/nar/gkl893 (2007).
50. Wang, S., Weng, S., Ma, J. & Tang, Q. DeepCNF-D: Predicting Protein Order/Disorder Regions by Weighted Deep Convolutional Neural Fields. *Int. J. Mol. Sci.* **16**, 17315–17330, https://doi.org/10.3390/ijms160817315 (2015).
51. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinforma.* **31**, 857–863, https://doi.org/10.1093/bioinformatics/btu744 (2015).
52. Dunker, A. K. *et al*. D2P2: database of disordered protein predictions. *Nucleic Acids Res.* **41**, D508–D516, https://doi.org/10.1093/nar/gks1226 (2012).
53. Piovesan, D. *et al*. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.* **46**, D471–D476, https://doi.org/10.1093/nar/gkx1071 (2018).
54. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinforma.* **7**, 208, https://doi.org/10.1186/1471-2105-7-208 (2006).
55. Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A. & Fidelis, K. Assessment of protein disorder region predictions in CASP10. *Proteins* **82**(Suppl 2), 127–137, https://doi.org/10.1002/prot.24391 (2014).
56. Nielsen, J. T. & Mulder, F. A. A. Quality and bias of protein disorder predictors. *Sci. Rep.* **9**, 5137, https://doi.org/10.1038/s41598-019-41644-w (2019).
57. Varadi, M. *et al*. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.* **42**, D326–335, https://doi.org/10.1093/nar/gkt960 (2014).
58. Piovesan, D. *et al*. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227, https://doi.org/10.1093/nar/gkw1056 (2017).
59. Holley, L. H. & Karplus, M. Protein secondary structure prediction with a neural network. *Proc. Natl Acad. Sci. USA* **86**, 152–156 (1989).
60. Kneller, D. G., Cohen, F. E. & Langridge, R. Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214**, 171–182, https://doi.org/10.1016/0022-2836(90)90154-E (1990).
61. Muskal, S. M. & Kim, S. H. Predicting protein secondary structure content. A tandem neural network approach. *J. Mol. Biol.* **225**, 713–727 (1992).
62. Qian, N. & Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865–884 (1988).
63. Hua, S. & Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinforma.* **17**, 721–728, https://doi.org/10.1093/bioinformatics/17.8.721 (2001).
64. Rost, B. & Sander, C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl Acad. Sci.* **90**, 7558–7562, https://doi.org/10.1073/pnas.90.16.7558 (1993).
65. Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961, https://doi.org/10.1016/0022-2836(87)90501-8 (1987).
66. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices11Edited by G. Von Heijne. *J. Mol. Biol.* **292**, 195–202, https://doi.org/10.1006/jmbi.1999.3091 (1999).
67. Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics* **40**, 502-511, 10.1002/1097-0134(20000815)40:3<502::AID-PROT170>3.0.CO;2-Q (2000).
68. Yu, C. H., Qin, Z., Martin-Martinez, F. J. & Buehler, M. J. A Self-Consistent Sonification Method to Translate Amino Acid Sequences into Musical Compositions and Application in Protein Design Using Artificial Intelligence. *ACS Nano*, https://doi.org/10.1021/acsnano.9b02180 (2019).
69. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324, https://doi.org/10.1109/5.726791 (1998).
70. Liu, Y., Chen, Y. & Cheng, J. *in 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 1771-1775.*
71. Li, Z. & Yu, Y. Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks. (2016).
72. Sønderby, S. K. & Winther, O. Protein Secondary Structure Prediction with Long Short Term Memory Networks. *arXiv*:1412.*7828 [cs, q-bio]* (2014).
73. Heffernan, R., Yang, Y., Paliwal, K. & Zhou, Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinforma.* **33**, 2842–2849, https://doi.org/10.1093/bioinformatics/btx218 (2017).
74. Magnan, C. N. & Baldi, P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinforma.* **30**, 2592–2597, https://doi.org/10.1093/bioinformatics/btu352 (2014).
75. Wang, Z., Zhao, F., Peng, J. & Xu, J. Protein 8-class secondary structure prediction using conditional neural fields. *Proteom.* **11**, 3786–3792, https://doi.org/10.1002/pmic.201100196 (2011).
76. Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci. Rep.* **6**, 18962 (2016).
77. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins* **86**(Suppl 1), 7–15, https://doi.org/10.1002/prot.25415 (2018).
78. Rost, B. Review: Protein Secondary Structure Prediction Continues to Rise. *J. Struct. Biol.* **134**, 204–218, https://doi.org/10.1006/jsbi.2001.4336 (2001).
79. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolym.* **22**, 2577–2637, https://doi.org/10.1002/bip.360221211 (1983).
80. Cheng, J., Tegge, A. N. & Baldi, P. Machine Learning Methods for Protein Structure Prediction. *IEEE Rev. Biomed. Eng.* **1**, 41–49, https://doi.org/10.1109/RBME.2008.2008239 (2008).

81. Dosztányi, Z. Prediction of protein disorder based on IUPred. *Protein Science: A Publ. Protein Soc.* **27**, 331–340, https://doi.org/10.1002/pro.3334 (2018).
82. Linding, R. *et al*. Protein Disorder Prediction: Implications for Structural Proteomics. *Structure* **11**, 1453–1459, https://doi.org/10.1016/j.str.2003.10.002 (2003).
83. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. & Uversky, V. N. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochimica et. Biophysica Acta - Proteins Proteom.* **1804**, 996–1010, https://doi.org/10.1016/j.bbapap.2010.01.011 (2010).
84. Kozlowski, L. P. & Bujnicki, J. M. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinforma.* **13**, 111, https://doi.org/10.1186/1471-2105-13-111 (2012).
85. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E. & Dunker, A. K. in Proceedings of International Conference on Neural Networks (ICNN'97). 90-95 vol.91.
86. Atkins, J. D., Boateng, S. Y., Sorensen, T. & McGuffin, L. J. Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies. *Int. J. Mol. Sci.* **16**, 19040–19054, https://doi.org/10.3390/ijms160819040 (2015).
87. Walsh, I. *et al*. Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinforma.* **31**, 201–208, https://doi.org/10.1093/bioinformatics/btu625 (2015).
88. Wang, G. & Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinforma.* **19**, 1589–1591 (2003).
89. McGuffin, L. J. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinforma.* **24**, 1798–1804, https://doi.org/10.1093/bioinformatics/btn326 (2008).
90. Gruenig, M. C. *et al*. Creating directed double-strand breaks with the Ref protein: a novel RecA-dependent nuclease from bacteriophage P1. *J. Biol. Chem.* **286**, 8240–8251, https://doi.org/10.1074/jbc.M110.205088 (2011).
91. Genomics, J. C. f. S. Crystal Structure of FMN-Binding Protein (NP_142786.1) from Pyrococcus Horikoshii at 1.35 A Resolution, www.rcsb.org/structure/2R6V (2007).
92. Todone, F., Weinzierl, R. O., Brick, P. & Onesti, S. Crystal structure of RPB5, a universal eukaryotic RNA polymerase subunit and transcription factor interaction target. *Proc. Natl Acad. Sci. USA* **97**, 6306–6310, https://doi.org/10.1073/pnas.97.12.6306 (2000).
93. Lafaye, C. *et al*. Biochemical and structural study of the homologues of the thiol-disulfide oxidoreductase DsbA in Neisseria meningitidis. *J. Mol. Biol.* **392**, 952–966, https://doi.org/10.1016/j.jmb.2009.07.056 (2009).
94. Dahms, S. O. *et al*. Metal binding dictates conformation and function of the amyloid precursor protein (APP) E2 domain. *J. Mol. Biol.* **416**, 438–452, https://doi.org/10.1016/j.jmb.2011.12.057 (2012).
95. Van Der Spoel, D. *et al*. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718, https://doi.org/10.1002/jcc.20291 (2005).
96. MacKerell, A. D. *et al*. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616, https://doi.org/10.1021/jp973084f (1998).
97. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935, https://doi.org/10.1063/1.445869 (1983).
98. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690, https://doi.org/10.1063/1.448118 (1984).
99. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **4**, 116–122, https://doi.org/10.1021/ct700200b (2008).
100. Essmann, U. *et al*. A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593, https://doi.org/10.1063/1.470117 (1995).
101. Becker, J., Maes, F. & Wehenkel, L. On the encoding of proteins for disordered regions prediction. *PLoS One* **8**, e82252, 10.1371/journal.pone.0082252 (2013).
102. Towns, J. *et al*. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **16**, 62-74, 10.1109/MCSE.2014.80. (2014).
103. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33-38, 27-38 (1996).

Acknowledgements

CCH acknowledges support from the MIT Undergraduate Research Opportunities Program. MJB acknowledges support from ONR (grant # N00014–16–1–651 2333) and NIH U01 EB014976. This work utilized the Extreme Science and Engineering Discovery Environment (XSEDE)[102], which is supported by National Science Foundation grant number ACI-1053575. XSEDE resources Stampede 2 and Ranch at the Texas Advanced Computing Center and Comet at the San Diego Supercomputing Center through allocation TG-MCB180008 were used.

Author contributions

C.C.H. and A.T. designed and performed research. A.T. performed molecular simulations and C.C.H. and A.T. analyzed data and prepared all figures. C.C.H., A.T. and M.J.B. wrote, reviewed, and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58868-w.

Correspondence and requests for materials should be addressed to A.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.