Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation

Giulio Tesei[1,◕,*], Anna Ida Trolle[1,◕], Nicolas Jonsson[1], Johannes Betz[1], Francesco Pesce[1], Kristoffer E. Johansson[1], Kresten Lindorff-Larsen[1,*]

[1] Structural Biology and NMR Laboratory & the Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark

◕These authors contributed equally to this work. * giulio.tesei@bio.ku.dk, lindorff@bio.ku.dk

Abstract

Intrinsically disordered proteins and regions (collectively IDRs) are pervasive across proteomes in all kingdoms of life, help shape biological functions, and are involved in numerous diseases. IDRs populate a diverse set of transiently formed structures, yet defy commonly held sequence-structure-function relationships. Recent developments in protein structure prediction have led to the ability to predict the three-dimensional structures of folded proteins at the proteome scale, and have enabled large-scale studies of structure-function relationships. In contrast, knowledge of the conformational properties of IDRs is scarce, in part because the sequences of disordered proteins are poorly conserved and because only few have been characterized experimentally. We have developed an efficient model to generate conformational ensembles of IDRs, and thereby to predict their conformational properties from sequence only. Here, we applied this model to simulate all IDRs of the human proteome. Examining conformational ensembles of 29,998 IDRs, we show how chain compaction is correlated with cellular function and localization, including in different types of biomolecular condensates. We train a model to predict compaction from sequence and use this to show conservation of structural properties across orthologs. Our results recapitulate observations from previous studies of individual protein systems, and enable us to study the relationship between sequence, conservation, conformational ensembles, biological function and disease variants at the proteome scale.

Introduction

Intrinsically disordered proteins (IDPs) and regions (IDRs) within and between folded domains play important and highly diverse roles in determining and regulating biological processes. IDPs and IDRs—from hereon collectively IDRs—are highly dynamic molecules whose structures must be represented by conformational ensembles. This dynamics in turn complicates the interpretation of experimental data, and limits our ability to understand the relationship between protein sequences, conformational ensembles and biological functions of IDRs.

Recent developments in protein structure prediction have led to a transformative ability to predict structures of folded proteins, and to perform structure-based analyses of protein function at the proteome level [1]. When methods such as AlphaFold are applied to IDRs they, however, result in a single or few configurations of low confidence that do not generally represent the conformational heterogeneity and functional states of IDRs [2]. Indeed, the AlphaFold pLDDT confidence score is an accurate predictor of disorder, because the structures of disordered proteins are generally predicted with low confidence [1]. AlphaFold has, however, been shown to provide information about local regions of proteins that may fold conditionally [3, 4], and used to provide starting points for refinement of IDR ensembles against experimental data [5].

Structural studies of disordered proteins generally rely on combining molecular simulations with experimental data to correct for intrinsic biases in simulation methods [6]. Such studies have provided a wealth of insights for individual IDRs [7] and revealed sequence-ensemble relationships for individual types of proteins [8, 9], but do not provide information about the structural properties across the proteome. Our inability to predict structural properties of IDRs thus limits our understanding of the functional roles of IDRs and how evolution shapes them. As a supplement to these one-by-one studies of IDRs, we have recently taken a different approach where we used experimental data on more than 50 different proteins to learn the parameters in a molecular energy function to predict conformational properties of IDRs. By globally optimizing a transferable model, called CALVADOS, we can study the conformational ensemble of any IDR in the absence of experimental data, and have validated the model extensively [10, 11]. In developing CALVADOS we focused on predicting conformational properties such as long-range interactions and overall compaction, since these properties are known to be important for function [12] and the ability of IDRs to self-associate to form biomolecular condensates [9]; other methods already exist to predict local structuring or binding regions within disordered proteins from sequence alone [13].

We have here taken advantage of the speed and accuracy of CALVADOS to perform molecular simulations and generate conformational ensembles of all 29,998 IDRs from the human proteome. We analysed the conformational properties of the ensembles and related them to the biological function of the proteins. For example, we find that some functions such as the ability to bind RNA or the localization within nuclear speckles or spliceosomes are enriched in proteins that have compact IDRs, whereas components of chromatin and nucleosomes are enriched in expanded IDRs. Our large sequence-ensemble dataset reveals the sequence properties that determine conformational ensembles, including the interplay between overall hydrophobicity and charge patterning. We take advantage of this knowledge and our ensembles to train a model to predict protein compaction from sequence only, and use it to show that IDRs in human orthologs have conserved conformational properties. In a cellular context, many IDRs form interactions with other molecules, and chain compaction and charge properties of the monomer quantify key aspects of a protein's ability to form both intra- and intermolecular interactions. Together, our work provides a new paradigm for how to link protein sequence, ensembles and function of disordered proteins, and sheds new light on the many biological roles of IDRs. Our database with conformational ensembles is freely available and will enable researchers to study the important and diverse roles of disordered proteins.

Results and Discussion

IDR selection and generation of conformational ensembles

We analysed AlphaFold2 models for 20,588 full-length human proteins and selected IDRs using low pLDDT confidence scores as predictor of disorder [1, 3, 4] (Fig. 1*A*), and obtained 29,998 sequences from 15,463 distinct proteins. The selected IDRs range in length between 30 and 500 residues (longer IDRs were divided into segments with a maximum of 500 residues), with a median value of 87 residues. Although pLDDT scores outperform many other disorder predictors, recent studies have highlighted that AlphaFold2 assigns high pLDDT scores to a fraction of known IDRs, possibly reflecting conditional folding upon binding or upon post-translational modifications [3, 4]. We therefore analysed sequence annotations in UniProt to calculate the fraction of residues within domains and structural motifs (f_{domain}), and find that only a small fraction of the IDRs overlap substantially with annotated domains (Fig. S1).

We then performed MD simulations of all 29,998 sequences using the experimentally-parameterized and broadly validated residue-level CALVADOS model [10, 11] to generate conformational ensembles for each disordered sequence, and calculated distributions of conformational properties from these ensembles (Fig. 1*B*). The model was trained on experimental data reporting on single-chain conformational properties and was shown to predict radii of gyration of

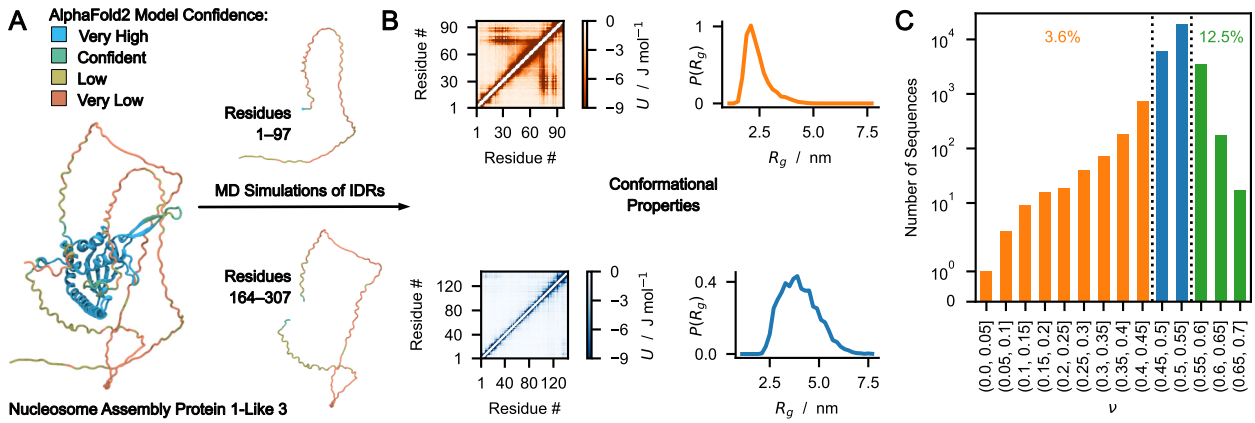

Figure 1: Schematic illustration of the approach used to obtain conformational properties for all the IDRs in the human proteome. (*A*) Selection of IDRs based on confidence of AlphaFold2 structural prediction, showing the selection of two IDRs from NP1L3 (UniProt ID: Q99457) as an example. (*B*) MD simulation of IDRs and calculation of conformational properties, showing interaction energy maps and distribution of the radius of gyration from the simulations of the two IDRs in NP1L3. (*C*) Distribution of the Flory scaling exponent, ν for 29,998 IDRs in the human proteome; note the logarithmic scale. A few IDRs have very small values of ν (0.4% have $\nu < 0.33$); these IDRs display conformational properties that are not well described by a simple scaling law. We note that we do not find evidence for a significant enrichment of misclassified structured sequences in compact IDRs (Fig. S2).

IDRs with mean relative error of 4% [11]. As measures of chain compaction that are independent of sequence length, we calculated the apparent Flory scaling exponent, ν, and the ratio of the mean-squared end-to-end distance and the mean-squared radius of gyration, $\langle R_{ee}^2 \rangle / \langle R_g^2 \rangle$; the accuracy of our model corresponds to an accuracy of predicting ν with an average error of ca. 0.01 for proteins of the lengths that we studied. For a so-called 'ideal-chain' polymer, the residue–residue, residue–solvent, and solvent–solvent interactions balance out and the polymer conformations are described by a Gaussian chain with $\nu = 0.5$ and $\langle R_{ee}^2 \rangle / \langle R_g^2 \rangle = 6$ [14, 15]. For most of the simulated IDRs, we find conformational ensembles which approximate this ideal reference state. Only 3.6% of the IDRs are substantially more compact ($\nu \leq 0.45$) whereas 12.5% of the sequences adopt more extended conformations ($\nu > 0.55$) (Fig. 1*C*). Although the relationship between $\langle R_{ee}^2 \rangle / \langle R_g^2 \rangle$ and ν is complex and sequence-dependent [15–18] (Fig. S3*A*), we obtain a similar distribution using $\langle R_{ee}^2 \rangle / \langle R_g^2 \rangle$ as a measurement of chain compaction (Fig. S3*B*). We also calculated other properties of the 29,998 conformational ensembles including interaction energy maps, asphericity (Δ) and prolateness (S). We find that both Δ and S are correlated with ν (Fig. S3*C*,*D*) so that the most compact ensembles are also most spherical and least prolate. We show 12 representative examples (Fig. S4) and provide access to all simulation trajectories and conformational properties in an online database. Generation of additional ensembles is possible via running CALVADOS using Google Colab (see Methods).

Relating compaction with function and localization

The compaction of an IDR can play a determining role in protein function [12, 19, 20]. Chain compaction also indicates poor solvation, and may correlate with increased attractive chain-chain interactions and propensity to phase separate [9, 21, 22]. Although the coupling between conformational properties and phase behaviour may break for sequences with different net charges [23], IDRs of proteins that undergo homotypic phase separation tend to have lower ν values compared to IDRs of proteins that do not phase separate [24, 25]. For these reasons, we sought to investigate the association of the presence of compact IDRs ($\nu \leq 0.45$) with the biological function and cellular localization of the corresponding full-length proteins. Our gene ontology analysis shows a significant enrichment of compact IDRs in proteins that bind RNA (odds ratio, OR = 2.4), histones (OR = 2.7), chromatin (OR = 1.9), and DNA (OR = 1.4) (Fig. 2*A*). These data resonate with the strong association between proteins with compact IDRs and their localization to nuclear complexes and membraneless

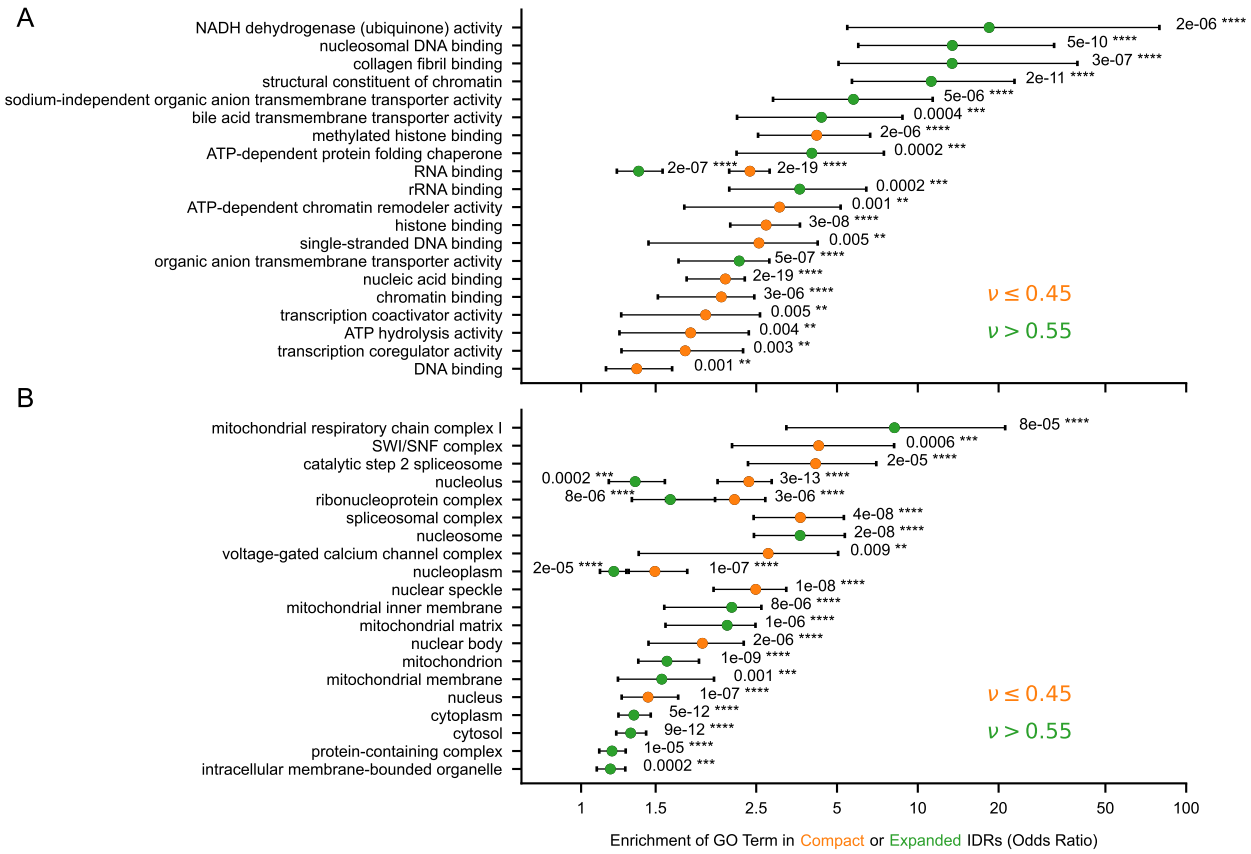

Figure 2: Gene ontology enrichment analysis of compact ($\nu \leq 0.45$, orange) and expanded ($\nu > 0.55$, green) IDRs. The strength of the association of compaction with (*A*) molecular functions and (*B*) cellular compartments is quantified by odds ratios. The statistical significance is estimated by adjusted p-values ≤ 0.001 using Fisher's exact test. Error bars show 95% confidence intervals.

organelles, namely, the SWI/SNF chromatin-remodelling complex (OR = 4.3), the spliceosomal complex (OR = 3.7), nuclear speckles (OR = 2.5), the nucleolus (OR = 2.3), and nuclear bodies (OR = 1.8) (Fig. 2*B*). Conversely, expanded IDRs are significantly enriched in histones (OR $>$ 11), in proteins of the mitochondrial respiratory chain complex I (OR $>$ 8), and in transmembrane proteins mediating the uptake of organic anions (OR $>$ 4) (Fig. 2).

We could ascribe a molecular function and cellular component annotations to 839 and 984 of the 1,081 IDRs with $\nu \leq 0.45$. Among these, we for example identified ten previously analysed [26] activation domains of transcription factors, such as pancreas transcription factor 1 subunit alpha, forkhead box proteins O4 and O6, and the zinc finger proteins 473, 561, 644, and 777. Furthermore, as examples of compact IDRs of RNA-binding proteins, we find the well-characterized low-complexity domains hnRNPA1, hnRNPA2, and FUS, whose compaction correlates with their ability to form biomolecular condensates in the cell [9, 27–29]. Multivalent interactions between IDRs are one of the driving forces for the formation of membraneless organelles [30] and our gene ontology analysis suggests a strong association between IDR compaction and protein partitioning into nuclear condensates. We identified compact IDRs for 372 proteins that localize in intracellular membraneless organelles and for 191 proteins that are annotated to localize in nucleoli, nuclear speckles, nuclear bodies, and PML bodies. For example we predicted compact IDRs for nucleophosmin, a component of the nucleolus [31]; the nuclear speckle protein snRNP subunit U1-70k, which is a splicing factor and forms insoluble aggregates in Alzheimer's disease [32, 33]; DAXX and ATRX, which are involved in chromatin remodeling and localize to PML bodies [34]; and CTR9, which is a positive regulator of RNA polymerase-II-mediated transcription [35].

Insights into sequence–compaction relationships

While the positional sequence identity of IDRs is generally poorly conserved [36, 37], amino acid composition and patterning have been proven useful for predicting conformational and phase properties as well as for identifying functionally similar IDRs [9, 26, 33, 35, 38–46].

Segregation of charged residues into oppositely charged blocks has been identified as a key driver of chain compaction [38, 39, 41] and condensate formation [44, 45, 47]. Conversely, high net charge per residue and uniform charge distribution result in expanded conformations [38, 39] and disfavor phase separation [10, 23, 48]. Recent studies have highlighted that IDRs with pronounced charge segregation are enriched in proteins involved in regulation of gene expression and mRNA processing. Specifically, for the IDRs of RNA polymerase II regulators, splicing factors, and polyadenylation factors, the extent of charge segregation has been shown to influence the selective partitioning of these proteins into transcription condensates [35], nuclear speckles [33], and nucleoli [46]. Moreover, charge blockiness has been identified as a conserved sequence feature [40] for example in homologous IDRs of the transcription elongation factors CTR9 and SPT6 [35].

Likewise, hydrophobic patterning modulates chain compaction and the propensity of IDRs to form condensates and avoid aggregation [9, 41]. Proteome-wide analyses have shown that the distributions of aromatic residues in prion-like IDRs is significantly more uniform than for random sequences [9, 49]. This finding suggests an evolutionary bias toward multivalent interactions between aromatic residues which underlie condensate formation, as opposed to stronger interactions between aromatic clusters which promote irreversible aggregation [9]. However, hydrophobic clusters have also been related to specific biological functions, as for the activation domains of transcription factors that bind coactivators. In these IDRs, clusters of aromatic and leucine residues are embedded in sequences rich in negatively charged residues [50], which may prevent chain collapse via electrostatic repulsion [26] and possibly degradation [51].

Inspired by these previous studies of individual protein families, we sought to examine the relationship between sequence and compaction across the entire set of human IDRs. For each sequence, we estimated the average residue stickiness, $\langle\lambda\rangle$ [11], as a measurement of the strength of attractive intra-chain interactions relative to solvent-protein interactions, and quantified the patterning of hydrophobic residues though the sequence hydropathy decorator (SHD) [41], using the amino-acid-specific stickiness parameters of the CALVADOS model. As metrics of charge segregation, we calculated the κ parameter [39] and the sequence charge decorator (SCD) [38]; κ and SCD capture different aspects of how uniform the charged residues are distributed along the polypeptide chain. More pronounced charge segregation is indicated by increasing κ values and decreasing SCD values. The κ parameter is normalized based on the composition of each sequence and ranges from 0 to 1 whereas SCD values correlate with the absolute net charge of the sequences. Therefore, κ and SCD are both influenced by sequence length and composition. Moreover, since κ is based on averages over windows of 5–6 residues [39], only SCD captures long-range interactions between blocks of charged residues along the linear sequence [48, 52].

We used the full set of IDR sequences to determine the sequence properties that affect chain compaction in human IDRs (Fig. 3). We find independent effects of SCD and κ (Fig. 3*A*,*B*,*G*); large negative and positive SCD values effectively distinguish chains with $\nu \approx 0.5$ from more compact and expanded IDRs, respectively (Fig. 3*A*,*E*). On the other hand, sequences with low or high values of ν are associated with larger κ values (Fig. 3*B*,*F*). With increasing compaction, sequences show increasing average stickiness, $\langle\lambda\rangle$, up to $\nu \sim 0.5$, after which $\langle\lambda\rangle$ plateaus and further decreases while the segregation of charged residues abruptly increases (Fig. 3*A*–*C*). These trends suggest that while hydrophobicity is used to drive some compaction, more extreme compaction is driven by electrostatic interactions between blocks of oppositely charged residues, in agreement with results from smaller sets of designed sequence variants [8, 39], and with the analysis of simulations of 16,885 synthetic sequences of equal chain length in an analysis performed in parallel with ours [53].

We hypothesized that Nature evolved the most compact IDRs by exploiting charge segregation rather than increased

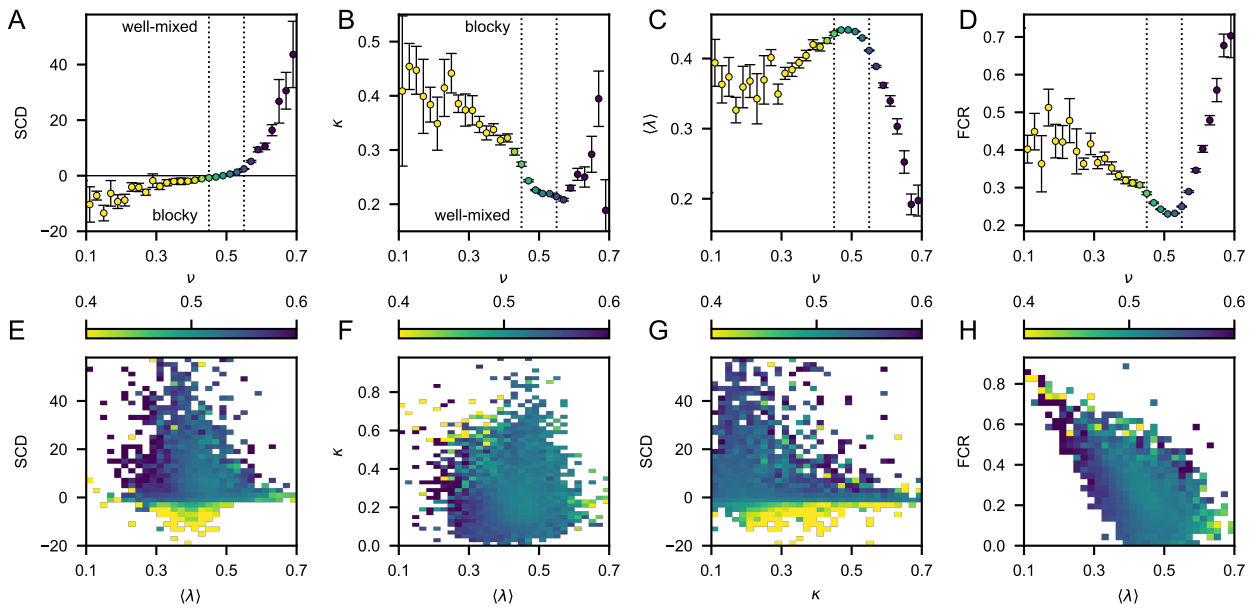

Figure 3: Sequence features that determine compaction. *A–D*: We show (*A*) sequence charge decoration, SCD, (*B*) charge segregation, κ, (*C*) average stickiness, $\langle\lambda\rangle$, and (*D*) fraction of charged residues, FCR as a function of the Flory scaling exponent, ν. *E–H*: Average ν as a function of (*E*) SCD and $\langle\lambda\rangle$, (*F*) κ and $\langle\lambda\rangle$, (*G*) SCD and κ, and (*H*) FCR and $\langle\lambda\rangle$.

hydrophobicity because sequences with large $\langle\lambda\rangle$ might lead to aggregation or degradation by the quality control apparatus [54]. To investigate this hypothesis, we used a predictor of quality control degradation signals, QCDPred, to estimate the propensities of the IDRs to be degraded [51]. The average QCDPred scores are low across the whole range of compaction, and more so for both compact and extended IDRs (Fig. S5*A*). We note that compact IDRs have on average considerably longer sequences than extended IDRs (Fig. S5*B*) and that the average QCDPred scores systematically decrease with increasing sequence length (Fig. S5*C*). This bias does therefore not explain the pronounced decrease in QCDPred score with decreasing ν below $\nu = 0.45$ and the similar values of the QCDPred score for extremely compact and expanded IDRs, which have significantly different hydrophobic content. Instead, the excess of negatively charged residues with respect to arginine and lysine residues in compact IDRs (Fig. S6) may weaken the degradation signal, as suggested by recent experiments [55].

Whereas both compact and expanded IDRs are characterized by large fractions of charged residues (FCR) (Fig. 3*D*), expanded sequences are short, acidic, and depleted in arginine residues (Fig. S5*B* and S6). Conversely, compact IDRs are considerably longer, charge segregated, and richer in positively charged residues, although they have on average negative net charge per residue (Fig. S6*C*). Notably, compact IDRs have more arginine ($\langle f_R\rangle = 7.9 \pm 0.1\%$) than lysine residues ($\langle f_K\rangle = 6.8 \pm 0.2\%$). Because of the amphiphilic character of the guanidinium side chain [56], and its ability to favorably interact with aromatic residues [23], this imbalance may contribute to stabilize compact conformations as well as protein-protein and protein-nucleotide interactions [27, 57].

Conservation of conformational and sequence features across homologs

Prompted by the observed relationships between sequence, compaction and biological function, we investigated the conservation of ν and sequence descriptors across a previously curated set of orthologs of human IDRs [58]. To estimate the scaling exponents of all the sequences, we trained a support vector regression (SVR) model against the ν values obtained from our simulations. Using permutation feature importance testing [59], we ranked the sequence descriptors based on their impact on model performance and selected the first five as input variables, namely, $\mathrm{SCD} > \mathrm{SHD} > \langle\lambda\rangle > \kappa > \mathrm{FCR}$ (Fig. S7*A*). For a test set of 2,999 IDRs, the ν values predicted by the developed model have a Pearson correlation coeffi-

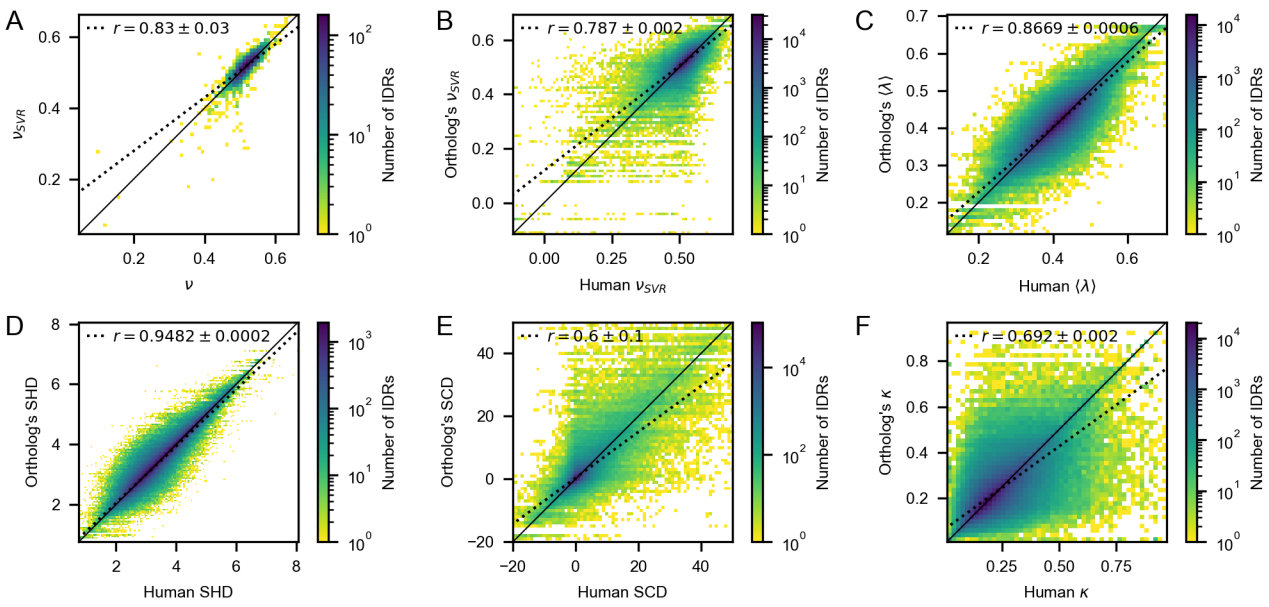

Figure 4: Conservation of conformational and sequence properties. (*A*) Correlation between ν values from simulations and ν_{SVR} values predicted using the SVR model for a held-out test set of 2,999 sequences from the set of 29,998 IDRs identified in this work. (*B*–*F*) Correlation of ν, $\langle\lambda\rangle$, SHD, SCD, and κ between human IDRs and their orthologs from a set of 513,333 homologous IDRs, ranging in sequence length from 30 to 5,601 residues [58]. Dotted lines show linear fits to the data. r is the Pearson correlation coefficient with the error estimated as the standard deviation over 1,000 bootstraps.

cient with the simulation data of 0.83 ± 0.03 and a mean absolute error of ca. 0.01 (Fig. 4*A*). Moreover, the SVR model accurately recapitulates the relationships between sequence features and ν obtained from simulations (Fig. S7*B*–*F*).

Using the SVR model, we calculated ν for the set of the 507,059 homologous IDRs (15,140 human IDRs and their 491,919 orthologs) [58] and found a high correlation between the compaction of human sequences and their orthologs (Fig. 4*B*). Average chain stickiness and hydrophobic patterning are also highly correlated between human and orthologous sequences (Fig. 4*C*,*D*). These findings are in agreement with recent bioinformatics analyses on homologous prion-like domains of FUS and FET family proteins, which showed that sequence features governing chain compaction are evolutionarily constrained [9, 23, 49, 60].

Compared to $\langle\lambda\rangle$ and SHD, we observed a weaker correlation for both SCD and κ between human and homologous IDRs (Fig. 4*E*,*F*). Despite the importance of charge segregation in governing chain compaction, this finding may suggest that this feature is less evolutionarily conserved than hydrophobic patterning. This weaker correlation observed may, however, possibly be attributed to the effect of post-translational modifications on charge patterning [61, 62], which may differ across species and was not accessible for our analysis.

Incidence of pathogenic variants and understudied genes

Given the observed evolutionary conservation of compaction, and its significant association with biological functions and cellular localization, we sought to investigate the potential relationship between IDR compaction and the incidence of pathogenic variants. We searched the ClinVar database [63] for missense variants in the human IDRs and found 2,977 pathogenic and 16,890 benign variants in 4,334 distinct proteins and 7,001 IDRs, which is less than a fourth of the IDRs in our set. Known pathogenic variants tend to be in regions with higher pLDDT scores than benign variants [64] (Fig. S8); this enrichment might be partially explained by the presence of a small fraction of folded regions within our set of IDRs (Fig. S8*B*), but may also reflect increased sensitivity in conserved regions of IDRs or those that may fold upon binding to their targets [3]. The sparsity of clinical data for IDRs could in part be ascribed to (i) the unknown functions of many IDRs and (ii) because computational methods to predict variant effects based on structure and sequence conservation are

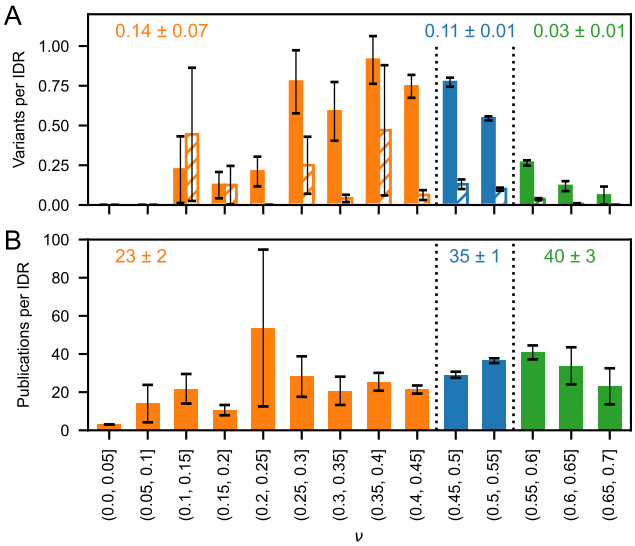

Figure 5: (*A*) Average number per IDR of benign (full bars) and pathogenic (hatched bars) ClinVar missense variants. (*B*) Average number of publications per IDR obtained from the "Find my Understudied Gene" tool [72]. Error bars show standard errors and the numbers refer to an average over the IDRs in the three groups of compaction.

less applicable [13, 65]. Nonetheless, pathogenic variants in IDRs have been associated with many diseases [66], such as autism spectrum disorder, cancer, and neurodegenerative disorders [65]. Specifically, several disease-associated variants in IDRs have been shown to alter the formation, composition, and material properties of biomolecular condensates [67–71]. Our analysis shows that expanded chains have on average a lower number of pathogenic and benign variants than more compact chains (Fig. 5*A*). However, this trend can be attributed to the sequence length of the expanded IDRs, which are on average approximately three times shorter than compact IDRs (Fig. S5*B*), and we do not find evidence for an overall enrichment of pathogenic variants in more compact IDRs. We used our SVR model to predict the change in ν for the 2,977 pathogenic and 16,890 benign missense variants in ClinVar and find very similar distributions for the two classes suggesting that changes in compaction of IDRs is not a general disease mechanism for these variants (Fig. S9*A*,*B*). Similar results were obtained for 5,676 frameshift variants that have recently been described to lead to modified disordered C-terminal regions [70] (Fig. S9*C*,*D*).

The ability to study biological functions and assign functional status to missense variants depends on our knowledge of protein function, which varies substantially across the proteome [73]. We therefore asked whether certain IDR-containing proteins are less studied than others [72]. We calculated an average of 35 publications per IDR and observed an uneven distribution with respect to chain compaction, which suggests that proteins with the most compact IDRs may have received less research attention than those with more expanded IDRs (Fig. 5*B*), perhaps because they are longer and more difficult to characterize experimentally.

Conclusions

IDRs are found in all kingdoms of life, and play essential roles in human biology and pathology. Our ability to study IDRs is, however, hampered by the fact that the relationship between sequence and structure is different from the more clearly defined patterns in folded proteins. This in turn makes it difficult to design experiments to test sequence-ensemble-function studies, and to understand the evolutionary processes that shape these. We have here performed a systematic analysis of the relationship between conformational ensembles and biological functions of human IDRs. Our conformational ensembles help rationalize previously observed relationships between sequence and function, and provides a unique dataset to continue to expand our knowledge of this understudied class of proteins, including training models to predict conforma-

tional properties [53, 74] or generate conformational ensembles [75]. In a cellular context, IDRs form interactions with other molecules, and our work provides a starting point for understanding both homo- and heterotypic interactions in IDR biology. Together with our and related methods [53] for predicting conformational properties from sequence and the mapping of biological observations onto our ensembles, we hope that our database of conformational ensembles presented in this work will encourage the investigation and the generation of hypotheses regarding the biological function, cellular localization, and variant effects of IDRs.

Methods

Selection of IDRs

Window-averaged AlphaFold2 pLDDT scores [1] were mapped to 20,215 ($> 98\%$) sequences from the human proteome as defined in UniProt release 2021_4. Residues with $\langle \mathrm{pLDDT} \rangle > 0.8$, $\langle \mathrm{pLDDT} \rangle < 0.7$, and $0.7 \leq \langle \mathrm{pLDDT} \rangle \leq 0.8$ were initially labeled as folded, disordered, and gap regions, respectively. Next, folded and disordered regions shorter than 10 residues were reclassified as gaps. Gap regions were relabeled as disordered if (i) flanked by disordered regions or (ii) terminal and preceded or followed by disordered regions. All other gap regions were instead relabeled as folded. IDRs smaller than 30 residues were discarded whereas IDRs larger than 500 residues were split into approximately equally long segments. This set of IDRs contain 37% of the residues in the UniProt proteome.

Analysis of domain annotations

AlphaFold2 may in some cases assign high pLDDT scores to IDRs, possibly reflecting conditional folding upon binding or upon post-translational modifications [3, 4]. Moreover, the absence of cofactors or ions in AlphaFold2 predictions may lead to low pLDDT scores for domains such as zinc fingers, EF-hand motifs, and C2 domains [76]. To assess the extent of folded residues in the selected IDRs, we analysed domain annotations in UniProt; for each IDR we calculated the fraction of residues within annotated domains and structural motifs, f_{domain}. Specifically, we obtained domain and zinc finger annotations for the selected IDRs from UniProt (uniprot.org/help/api) via application programming interface (API) calls (April 2023). A minimum sequence overlap of ten residues was used to evaluate the presence of the domains in the IDRs. We found that only 2.1% of the sequences have $f_{\text{domain}} > 0.5$ whereas 1.0% have $f_{\text{domain}} > 0.9$ (Fig. S1*A*). These putative folded sequences have generally short sequence length (Fig. S1*B*) and often include zinc or calcium binding sites, loops in protein kinase and myosin motor domains, and disulfide bonds in extracellular protein domains (Fig. S1*C*).

Molecular simulations

Molecular dynamics simulations were conducted in the NVT ensemble at $T = 37$ ºC using the Langevin integrator with a time step of 10 fs and friction coefficient of 0.01 ps^{-1}. The Cα-based model was implemented using the CALVADOS 2 parameters and functional forms as previously detailed [11]. Single chains of N residues are simulated using HOOMD-blue v2.9.3 [77] in a cubic box of side length $0.38 \times (N - 1) + 4$ nm under periodic boundary conditions, starting from the fully extended linear conformation. The charge number of the first or last residue of the simulated sequences is increased or decreased by one unit when it coincides with the N- or C-terminus of the full-length protein, respectively. Histidine residues are modeled in the neutral form. Conformations are saved every $\Delta t \approx 3 \times N^2$ fs if $N > 150$ and $\Delta t = 70$ ps otherwise. Each chain is simulated for a simulation time of $1,010 \times \Delta t$ and the initial 10 frames are discarded so as to obtain 1,000 weakly correlated conformations for each protein (Fig. S10*A*). The uncertainty of the averages of the radius of gyration, R_g, and of the end-to-end distance, R_{ee}, over the conformational ensembles is estimated as the standard error obtained from the blocking approach [78] implemented in the BLOCKING software

(github.com/fpesceKU/BLOCKING). Apparent Flory scaling exponents, ν, are obtained by fitting $\langle R_{ij}\rangle = R_0|i-j|^\nu$ to average residue-residue distances, $\langle R_{ij}\rangle$, calculated for separations along the linear sequence, $|i-j|$, larger than 5. The sampling error on ν and $\langle R_{ee}^2\rangle/\langle R_g^2\rangle$ was estimated to be 0.01 and 0.1, respectively, from the standard deviation over five independent replicas averaged over a subset of 33 IDRs (Fig. S10*B*,*C*). Intra-chain distances were calculated from simulations trajectories using MDTraj [79]. The charge segregation parameter κ was calculated using localCIDER [80] or set to zero for sequences without charged residues.

We also provide easy access to generate conformational ensembles using CALVADOS simulations via a Notebook that for example can be run using Google colaboratory github.com/KULL-Centre/EnsembleLab. Typical simulation times range from ca. 5 min (a 71 ns-long simulation of a protein of 70 residues), 7 min (71 ns-long simulation of 140 residues), to 34 min for a 373 ns-long simulation of a protein with 351 residues.

Gene-Ontology analysis

Gene Ontology (GO) terms for the selected IDRs were obtained directly from UniProt (uniprot.org/help/api) via API calls (January 2023). The basic version of the GO was downloaded in OBO format from the GO Consortium (purl.obolibrary.org/obo/go/go-basic.obo) (January 2023) and the Python package for complex network analysis NetworkX [81] was used to read the file as a directed acyclic graph. The graph has a hierarchical structure with three root nodes: "cellular component", "biological process", and "molecular function". In this work, we analysed the "molecular function" and "cellular component" domains and discarded all child terms of the "biological process" domain. This filtering resulted in 23,916 and 27,129 IDRs with GO annotations of the corresponding full-length proteins pertaining to "molecular function" and "cellular component", respectively. The numbers of compact IDRs ($\nu \leq 0.45$) with "molecular function" and "cellular component" annotations are 839 and 984 whereas the corresponding numbers for expanded IDRs ($\nu > 0.55$) are 3,112 and 3,478. The GO terms were mapped onto parent terms occurring at least 10 times in proteins of IDRs with $\nu \leq 0.45$. GO terms which are unrelated to any of the selected parent terms were labeled as "other". The GO terms were binned based on the ν values of the IDRs, using the threshold $\nu \leq 0.45$. For each GO term, a 2×2 contingency table was populated with the numbers of occurrences of the GO term for IDRs with ν below or above 0.45. The strength of the association between chain compaction and GO terms was quantified as the odds ratio,

$$\text{Odds Ratio} = \frac{n_{\text{GO}}(\nu \leq 0.45)}{n(\nu \leq 0.45) - n_{\text{GO}}(\nu \leq 0.45)} \Bigg/ \frac{n_{\text{GO}}(\nu > 0.45)}{n(\nu > 0.45) - n_{\text{GO}}(\nu > 0.45)}, \tag{1}$$

where n_{GO} and n denote the number of occurrences of the GO term and the total number of sequences in the ν-bin, respectively. Fisher's exact test was used to estimate the significance of the association and the p-values were adjusted for multiple testing according to the Benjamini-Hochberg method [82]. The same analysis was repeated using the threshold $\nu > 0.55$.

QCD calculations

We used QCDPred [51] to predict quality control degrons for all IDR sequences, and report the mean degron score across each IDR as representing its degron potential. Briefly, this method considers all possible 17-residues segments of an IDR and predicts a degron score based on the amino acid composition with hydrophobic residues acting to promote degradation and negatively charged residues to counteract degradation.

Support vector regression model

A support vector regression model with a radial basis function kernel [83, 84] was developed to predict ν using as input features the sequence charge decoration, SCD [38]; the sequence hydropathy decoration, SHD [41]; the mean stickiness

$\langle\lambda\rangle$; the κ parameter [39]; and the fraction of charged residues, FCR. The model was generated by randomly splitting the simulated sequences into training (30%), validation (60%), and test sets (10%). The sequence descriptors SCD, SHD, $\langle\lambda\rangle$, κ, and FCR were selected from a larger set which also included the net charge per residue, NCPR; and the fraction of aromatic, f_{aro}; arginine, f_{R}; lysine, f_{K}; glutamate, f_{E}; and aspartate residues, f_{D}. The selection was based on their impact on model performance estimated using permutation feature importance testing [59]. Optimal hyperparameters (ϵ = 0.01, C = 78) were obtained using grid search with the Pearson correlation coefficient as a performance metric, leading to a correlation of 0.83 and a mean absolute error of 0.013 on the test set. All model training and application, including permutation feature importance testing [59], was performed using scikit-learn [85]. We note that a similar model was recently described [74] using simulations with a different force field [41] that generally puts emphasis on sequence hydrophobicity and does not capture experiments as accurately as CALVADOS [10].

ClinVar analysis

ClinVar data were obtained by parsing the ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz (May 2021) as described previously [86]. Namely, we included only entries that (i) have a rating of at least one star, (ii) are single nucleotide variants, and (iii) are mapped to the Genome Reference Consortium Human Build 38 (GRCh38), using the scripts available at github.com/KULL-Centre/PRISM/tree/main/software/make_prism_files, release-tag v0.1.1. Next, we retained only ClinVar missense variants within the selected IDRs and omitted all deletion variants. All "likely benign" and "likely pathogenic" variants were reclassified as "benign" and "pathogenic", respectively. Subsequently, variants with categorical labels other than "benign" and "pathogenic" were discarded.

Data and code availability

The 29,998 conformational ensembles and precalculated conformational properties are available via sid.erda.dk/sharelink/AVZAJvJnCO, which also provides access to code and data to reproduce the results presented in this work, and to `Supporting_File_1.xlsx`, which lists the sequences and various sequence- and conformational properties of all 29,998 IDRs. Code and data are also available via github.com/KULL-Centre/_2023_Tesei_IDRome. CALVADOS is available from github.com/KULL-Centre/CALVADOS and can be run using Google Colab via github.com/KULL-Centre/EnsembleLab.

Acknowledgments

We thank Drs. Tanja Mittag and Xavier Salvatella for useful comments on our work, and Dr. Alex Holehouse for sharing work before publication. We acknowledge access to computational resources from the Biocomputing Core Facility at the Department of Biology, University of Copenhagen. This work is a contribution from the PRISM (Protein Interactions and Stability in Medicine and Genomics) centre funded by the Novo Nordisk Foundation (to K.L.-L.; NNF18OC0033950). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101025063.

Supporting Figures

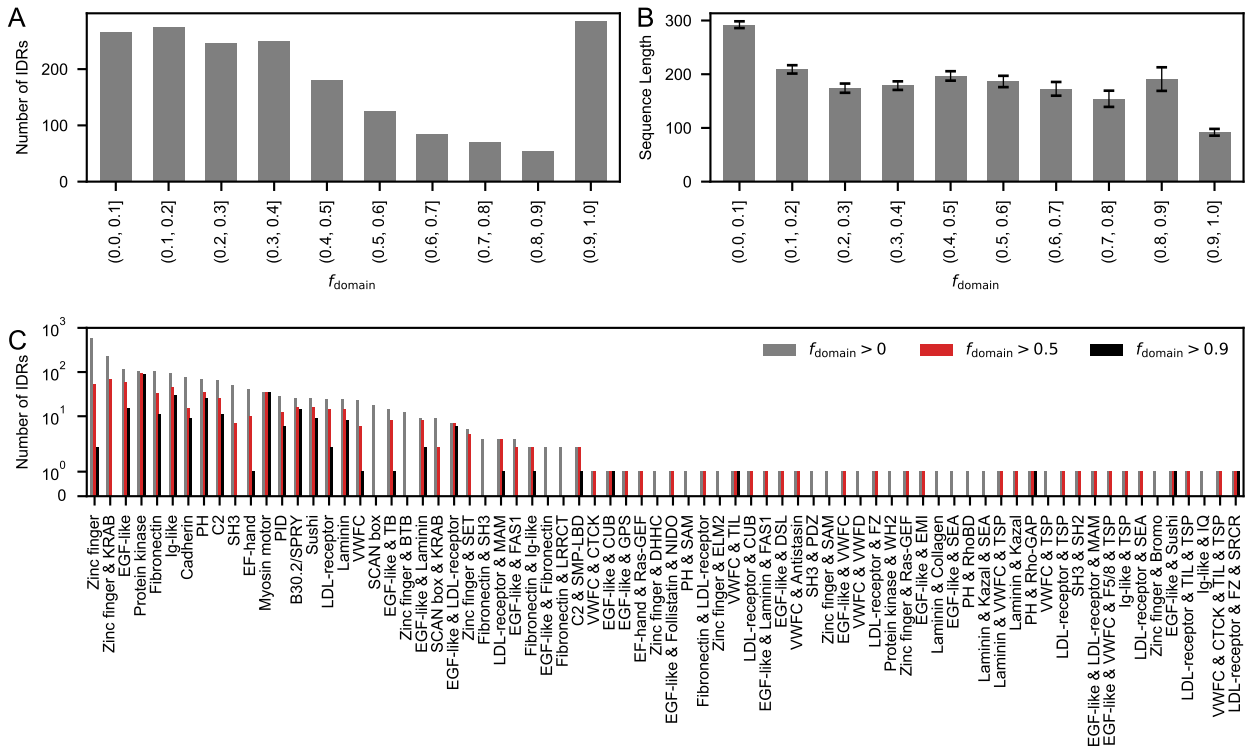

Figure S1: Annotation of the 1,837 IDRs (6%) with any residues in annotated domains. (*A*) Number of IDRs and (*B*) average sequence length of IDRs as a function of the fraction of residues in domains, f_{domain}. Error bars show standard errors. (*C*) Number of IDRs in which $f_{\text{domain}} > 0$ (gray), > 0.5 (red), and > 0.9 (black) based on domain and zinc finger annotations in UniProt.

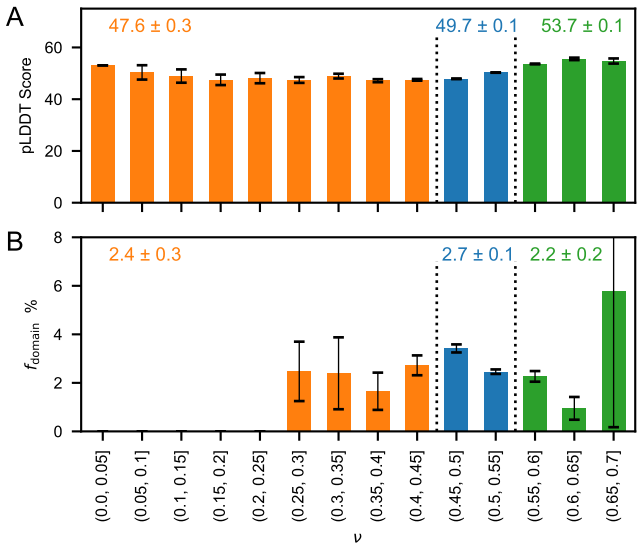

Figure S2: (*A*) Average pLDDT scores of the human IDRs selected in this study as a function of ν. (*B*) Average fraction of residues in domains annotated in UniProt for the IDRs as a function of ν. Error bars show standard errors and the numbers are averages and standard errors over IDRs with $\nu \leq 0.45$ (orange), $0.45 < \nu \leq 0.55$ (blue), and $\nu > 0.55$ (green).

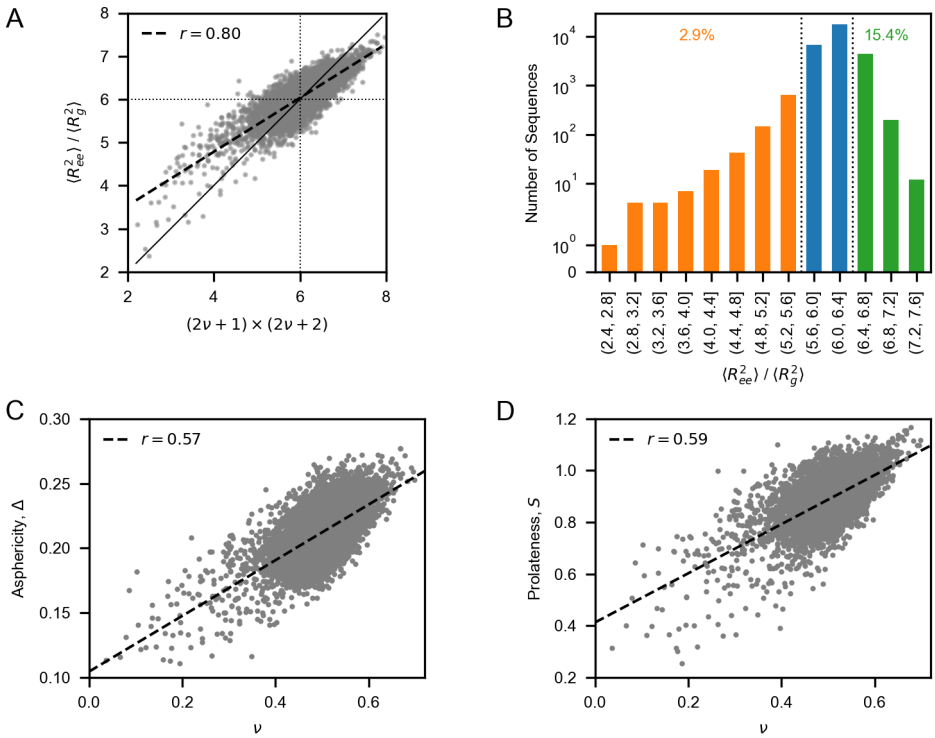

Figure S3: (*A*) Correlation between $\langle R_{ee}^2 \rangle$ / $\langle R_g^2 \rangle$ calculated from simulation trajectories and the approximate relation $\langle R_{ee}^2 \rangle$ / $\langle R_g^2 \rangle = (2\nu + 1) \times (2\nu + 2)$, where ν is the apparent Flory scaling exponent [87]. (*B*) Distribution of $\langle R_{ee}^2 \rangle$ / $\langle R_g^2 \rangle$ for IDRs in the human proteome; note the logarithmic scale. (*C*) Correlation between asphericity, Δ, and ν [88]. (*D*) Correlation between prolateness, S, and ν [88]. r is the Pearson correlation coefficient.

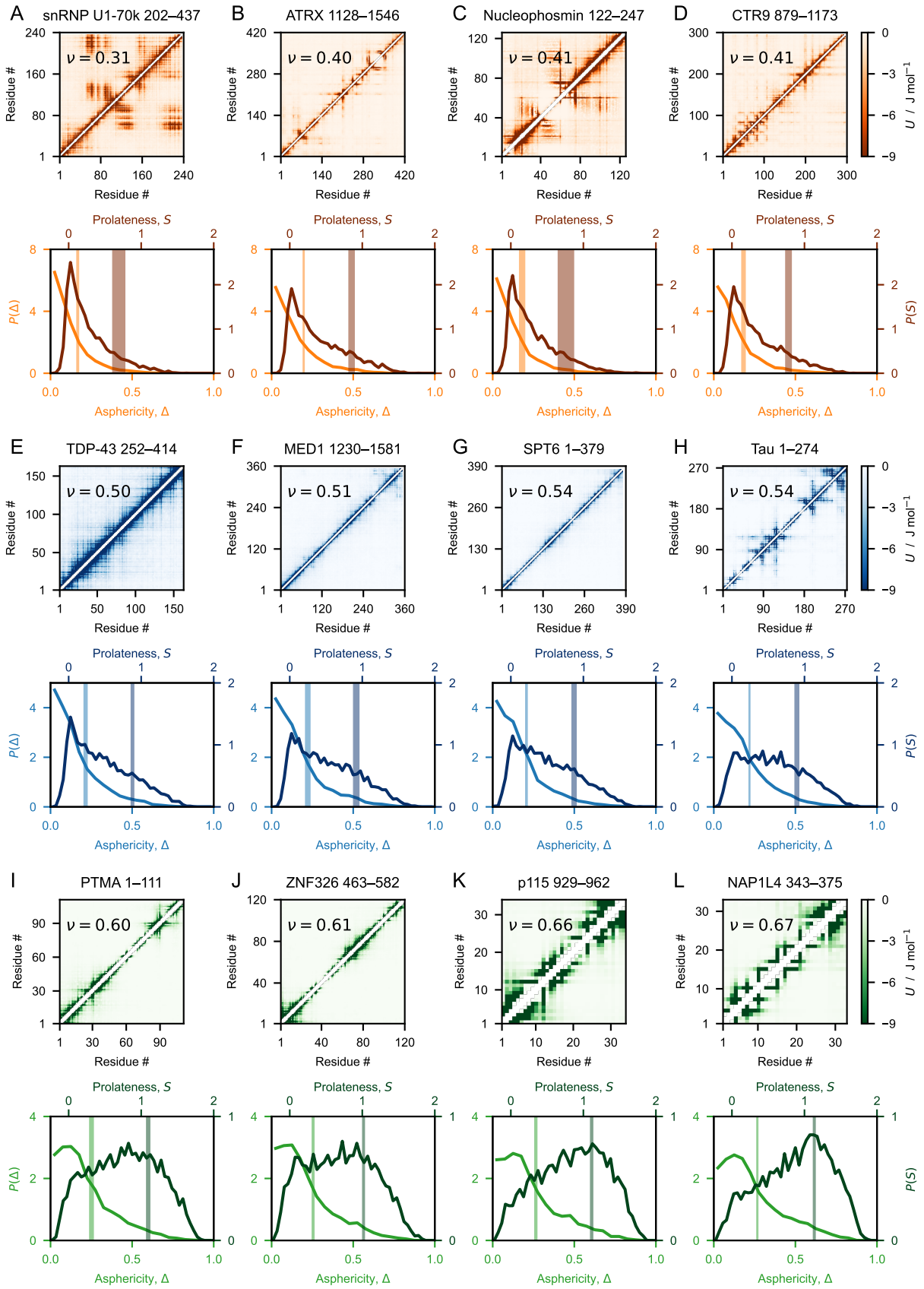

Figure S4: Representative intra-chain energy maps (*Top* panels), and distributions of the conformational parameters asphericity, Δ, and prolateness, S, (*Bottom* panels) for a selection of IDRs from our database. The UniProt IDs of the IDRs are (*A*) P08621, (*B*) P46100, (*C*) P06748, (*D*) Q6PD62, (*E*) Q13148, (*F*) Q15648, (*G*) Q7KZ85, (*H*) P10636, (*I*) P06454, (*J*) Q5BKZ1, (*K*) O60763, and (*L*) Q99733. The data is averaged over five ensembles of 1,000 weakly correlated conformations (Fig. S10), each obtained from an independent MD simulation. (*Top* panels) Energy maps are calculated using the non-ionic Ashbaugh-Hatch potential of the CALVADOS 2 model. (*Bottom* panels) Lines and vertical shaded areas show distributions and confidence intervals of the ensemble averages estimated from the standard deviation of Δ (light shade) and S (dark shade) over five independent simulations.

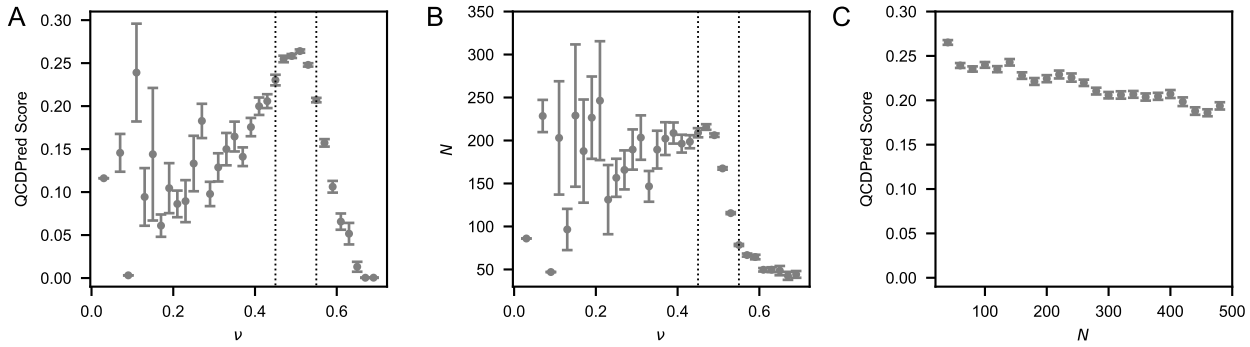

Figure S5: (*A*) QCDPred score and (*B*) sequence length, N, as a function of the apparent Flory scaling exponent, ν. (*C*) QCDPred score as a function of sequence length. Error bars show standard errors.

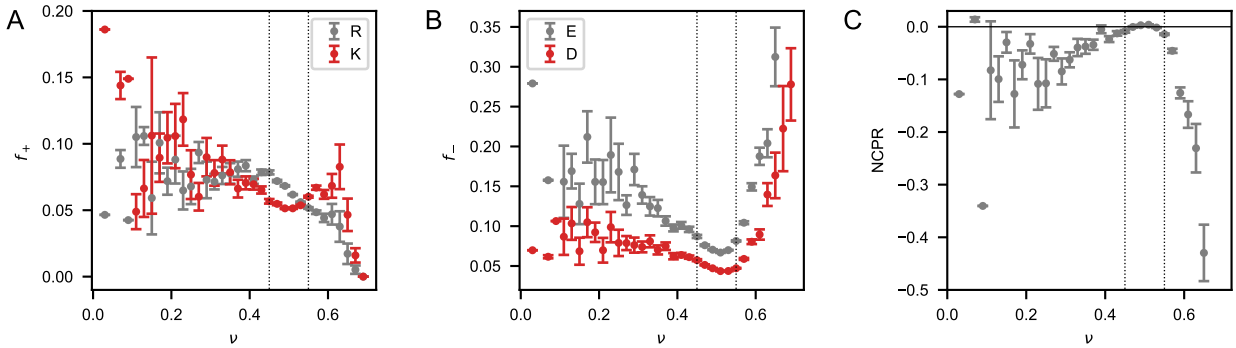

Figure S6: (*A*) Average fraction of positively charged residues Arg (gray) and Lys (red) as a function of ν. (*B*) Average fraction of negatively charged residues Glu (gray) and Asp (red) as a function of ν. Error bars show standard errors. (*C*) Average net charge per residue, NCPR, as a function of the apparent Flory scaling exponent, ν.

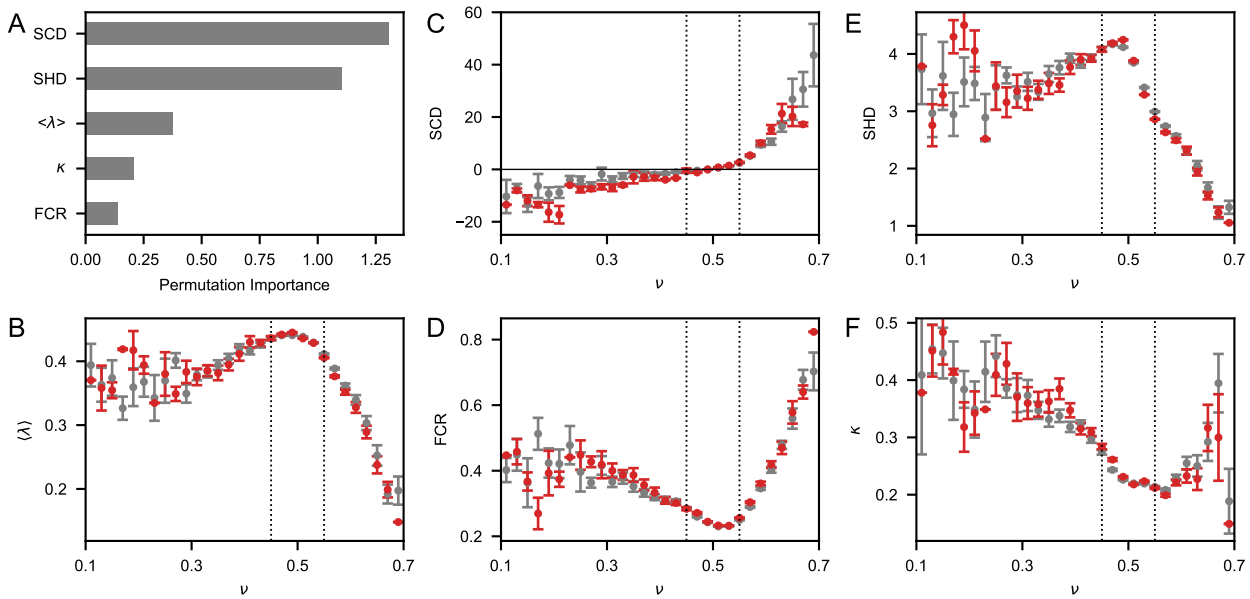

Figure S7: (*A*) Permutation importance of the sequence features used in the SVR model. (*B*–*F*) Average sequence features as a function of the apparent Flory scaling exponents, ν, calculated from simulation trajectories (gray) and using the SVR model (red). Error bars show standard errors.

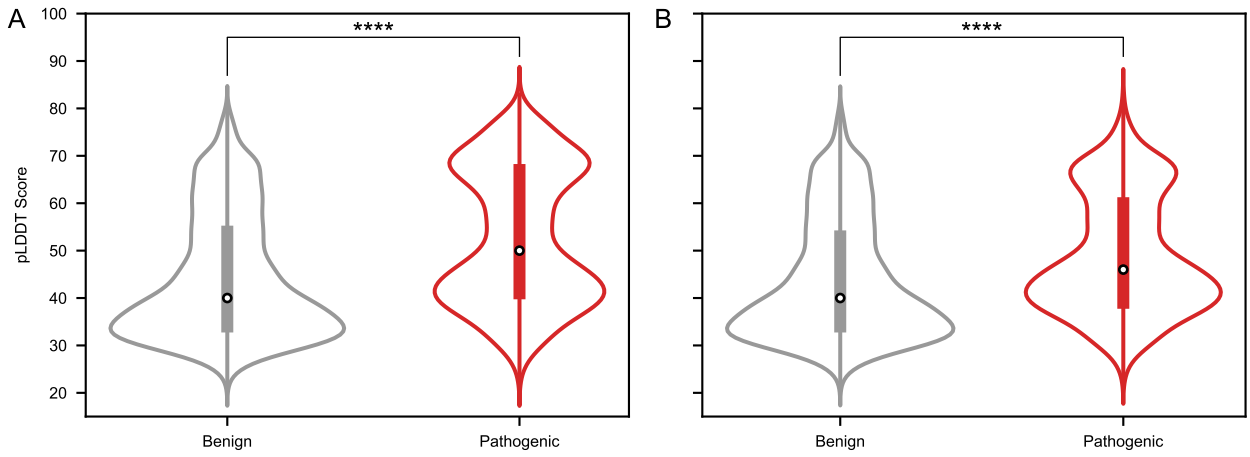

Figure S8: Violin plots of pLDDT scores for the wild-type residues substituted in benign (gray) and pathogenic (red) missense variants from ClinVar. The analysis is performed over variants within (*A*) all the IDRs in our database and (*B*) excluding variants within IDRs with $f_{\text{domain}} > 0$. **** indicates p-values $< 10^{-75}$ from a two-tailed Welch's t-test.

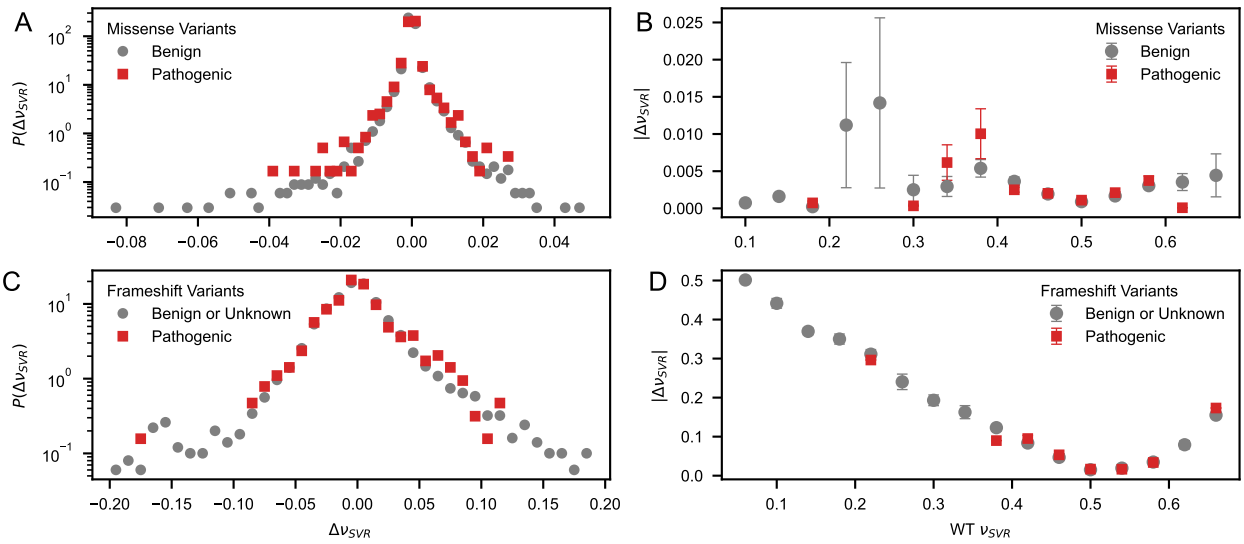

Figure S9: Distributions of the difference in ν between variant and wild-type, $\Delta\nu_{\text{SVR}}$, for (*A*) Clinvar missense variants and (*C*) frameshift variants identified by Mensah et al. [70]. Average absolute values of $\Delta\nu_{\text{SVR}}$ as a function of the wild-type ν_{SVR} value for (*B*) Clinvar missense variants and (*D*) frameshift variants identified by Mensah et al. [70]. Benign and pathogenic variants are shown in gray and red, respectively. ν_{SVR} values are estimated using the SVR model developed in this work.

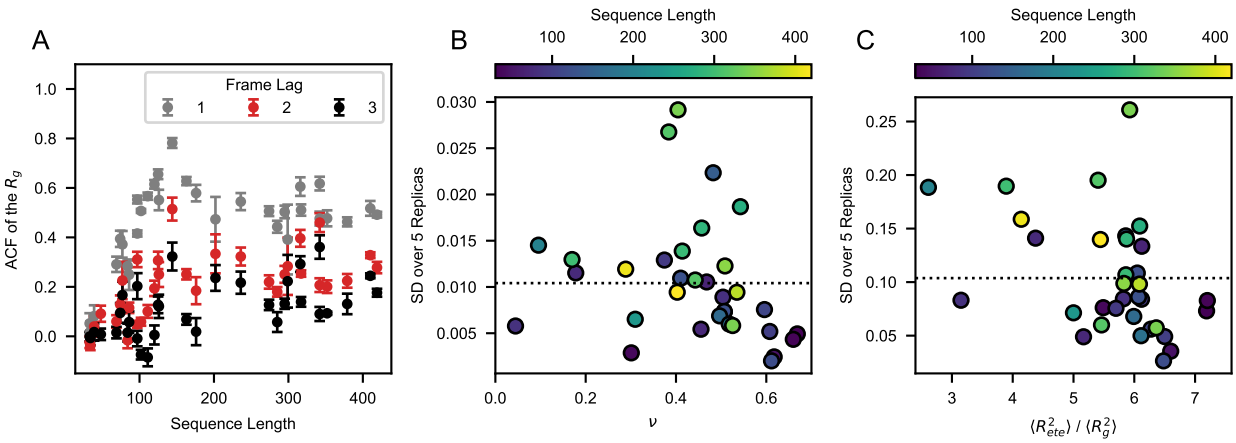

Figure S10: (*A*) Values of the autocorrelation function (ACF) of the R_g for lag times of one (gray), two (red), and three (black) frames as a function of sequence length. Error bars show standard deviation (SD) over five replicas. (*B*) Standard deviation (SD) of the apparent Flory scaling exponent, ν, as a function of ν for proteins of different sequence length. (*C*) SD of the ratio of the mean-squared end-to-end distance and the mean-squared radius of gyration, $\langle R_{ee}^2 \rangle / \langle R_g^2 \rangle$, ν, as a function of $\langle R_{ee}^2 \rangle / \langle R_g^2 \rangle$ for proteins of different sequence length. SD values are calculated over five simulation replicas. Dotted lines show the average SD.

References

[1] Mehmet Akdel, Douglas E. V. Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O. Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L. Good, Roman A. Laskowski, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Petras Kundrotas, Victoria Ruiz Serra, Carlos H. M. Rodrigues, Alistair S. Dunham, David Burke, Neera Borkakoti, Sameer Velankar, Adam Frost, Jérôme Basquin, Kresten Lindorff-Larsen, Alex Bateman, Andrey V. Kajava, Alfonso Valencia, Sergey Ovchinnikov, Janani Durairaj, David B. Ascher, Janet M. Thornton, Norman E. Davey, Amelie Stein, Arne Elofsson, Tristan I. Croll, and Pedro Beltrao. A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, November 2022. doi:10.1038/s41594-022-00849-w.

[2] Kiersten M. Ruff and Rohit V. Pappu. AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167208, October 2021. doi:10.1016/j.jmb.2021.167208.

[3] T. Reid Alderson, Iva Pritišanac, Alan M. Moses, and Julie D. Forman-Kay. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *bioRxiv*, February 2022. doi:10.1101/2022.02.18.481080.

[4] Damiano Piovesan, Alexander Miguel Monzon, and Silvio C. E. Tosatto. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Science*, 31(11), October 2022. doi:10.1002/pro.4466.

[5] Z Faidon Brotzakis, Shengyu Zhang, and Michele Vendruscolo. Alphafold prediction of structural ensembles of disordered proteins. *bioRxiv*, pages 2023–01, 2023.

[6] F Emil Thomasen and Kresten Lindorff-Larsen. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. *Biochemical Society Transactions*, 50(1):541–554, 2022.

[7] Tamas Lazar, Elizabeth Martínez-Pérez, Federica Quaglia, András Hatos, Lucía B Chemes, Javier A Iserte, Nicolás A Méndez, Nicolás A Garrone, Tadeo E Saldaño, Julia Marchetti, et al. Ped in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic acids research*, 49(D1):D404–D411, 2021.

[8] Rahul K. Das, Yongqi Huang, Aaron H. Phillips, Richard W. Kriwacki, and Rohit V. Pappu. Cryptic sequence features within the disordered protein $p27^{Kip1}$ regulate cell cycle signaling. *Proceedings of the National Academy of Sciences*, 113(20):5616–5621, May 2016. doi:10.1073/pnas.1516277113.

[9] Erik W. Martin, Alex S. Holehouse, Ivan Peran, Mina Farag, J. Jeremias Incicco, Anne Bremer, Christy R. Grace, Andrea Soranno, Rohit V. Pappu, and Tanja Mittag. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*, 367(6478):694–699, February 2020. doi:10.1126/science.aaw8653.

[10] Giulio Tesei, Thea K. Schulze, Ramon Crehuet, and Kresten Lindorff-Larsen. 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(44), 2021. doi:10.1073/pnas.2111696118.

[11] Giulio Tesei and Kresten Lindorff-Larsen. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range [version 2; peer review: 2 approved]. *Open Research Europe*, 2:94, January 2023. doi:10.12688/openreseurope.14967.2.

[12] Nicolas S Gonzalez-Foutel, Juliana Glavina, Wade M Borcherds, Matías Safranchik, Susana Barrera-Vilarmau, Amin Sagar, Alejandro Estaña, Amelie Barozet, Nicolás A Garrone, Gregorio Fernandez-Ballester, et al. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nature structural & molecular biology*, 29(8):781–790, 2022.

[13] 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. *Journal of Molecular Biology*, 433(20):167196, October 2021. doi:10.1016/j.jmb.2021.167196.

[14] H. Hofmann, A. Soranno, A. Borgia, K. Gast, D. Nettels, and B. Schuler. 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, September 2012. doi:10.1073/pnas.1207719109.

[15] Wenwei Zheng, Gül H. Zerze, Alessandro Borgia, Jeetain Mittal, Benjamin Schuler, and Robert B. Best. Inferring properties of disordered chains from FRET transfer efficiencies. *The Journal of Chemical Physics*, 148(12):123329, March 2018. doi:10.1063/1.5006954.

[16] Alessandro Borgia, Wenwei Zheng, Karin Buholzer, Madeleine B. Borgia, Anja Schüler, Hagen Hofmann, Andrea Soranno, Daniel Nettels, Klaus Gast, Alexander Grishaev, Robert B. Best, and Benjamin Schuler. Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *Journal of the American Chemical Society*, 138(36):11714–11726, September 2016. doi:10.1021/jacs.6b05917.

[17] Gustavo Fuertes, Niccolò Banterle, Kiersten M. Ruff, Aritra Chowdhury, Davide Mercadante, Christine Koehler, Michael Kachala, Gemma Estrada Girona, Sigrid Milles, Ankur Mishra, Patrick R. Onck, Frauke Gräter, Santiago Esteban-Martín, Rohit V. Pappu, Dmitri I. Svergun, and Edward A. Lemke. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proceedings of the National Academy of Sciences*, 114(31):E6342–E6351, July 2017. doi:10.1073/pnas.1704692114.

[18] Jianhui Song, Jichen Li, and Hue Sun Chan. Small-angle x-ray scattering signatures of conformational heterogeneity and homogeneity of disordered protein ensembles. *The Journal of Physical Chemistry B*, 125(24):6451–6478, June 2021. doi:10.1021/acs.jpcb.1c02453.

[19] Kathryn P Sherry, Rahul K Das, Rohit V Pappu, and Doug Barrick. Control of transcriptional activity by design of charge patterning in the intrinsically disordered ram region of the notch receptor. *Proceedings of the National Academy of Sciences*, 114(44):E9243–E9252, 2017.

[20] Mateusz Dyla and Magnus Kjaergaard. Intrinsically disordered linkers control tethered kinases via effective concentration. *Proceedings of the National Academy of Sciences*, 117(35):21413–21419, 2020.

[21] Thomas Allan Orofino and PJ Flory. Relationship of the second virial coefficient to polymer chain dimensions and interaction parameters. *The Journal of Chemical Physics*, 26(5):1067–1076, 1957.

[22] Joshua A Riback, Christopher D Katanski, Jamie L Kear-Scott, Evgeny V Pilipenko, Alexandra E Rojek, Tobin R Sosnick, and D Allan Drummond. Stress-triggered phase separation is an adaptive, evolutionarily tuned response. *Cell*, 168(6):1028–1040, 2017.

[23] Anne Bremer, Mina Farag, Wade M. Borcherds, Ivan Peran, Erik W. Martin, Rohit V. Pappu, and Tanja Mittag. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nature Chemistry*, December 2021. doi:10.1038/s41557-021-00840-w.

[24] Elisia A. Paiz, Jeffre H. Allen, John J. Correia, Nicholas C. Fitzkee, Loren E. Hough, and Steven T. Whitten. Beta turn propensity and a model polymer scaling exponent identify intrinsically disordered phase-separating proteins. *Journal of Biological Chemistry*, 297(5):101343, November 2021. doi:10.1016/j.jbc.2021.101343.

[25] Ayyam Y. Ibrahim, Nathan P. Khaodeuanepheng, Dhanush L. Amarasekara, John J. Correia, Karen A. Lewis, Nicholas C. Fitzkee, Loren E. Hough, and Steven T. Whitten. Intrinsically disordered regions that drive phase separation form a robustly distinct protein class. *Journal of Biological Chemistry*, 299(1):102801, January 2023. doi:10.1016/j.jbc.2022.102801.

[26] Max V. Staller, Eddie Ramirez, Sanjana R. Kotha, Alex S. Holehouse, Rohit V. Pappu, and Barak A. Cohen. Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Systems*, 13(4):334–345.e5, April 2022. doi:10.1016/j.cels.2022.01.002.

[27] Jie Wang, Jeong-Mo Choi, Alex S. Holehouse, Hyun O. Lee, Xiaojie Zhang, Marcus Jahnel, Shovamayee Maharana, Régis Lemaitre, Andrei Pozniakovsky, David Drechsel, Ina Poser, Rohit V. Pappu, Simon Alberti, and Anthony A. Hyman. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell*, 174(3):688–699.e16, July 2018. doi:10.1016/j.cell.2018.06.006.

[28] Veronica H. Ryan, Gregory L. Dignon, Gül H. Zerze, Charlene V. Chabata, Rute Silva, Alexander E. Conicella, Joshua Amaya, Kathleen A. Burke, Jeetain Mittal, and Nicolas L. Fawzi. Mechanistic view of hnRNPA2 low-complexity domain structure, interactions, and phase separation altered by mutation and arginine methylation. *Molecular Cell*, 69(3):465–479.e7, February 2018. doi:10.1016/j.molcel.2017.12.022.

[29] Zachary Monahan, Veronica H Ryan, Abigail M Janke, Kathleen A Burke, Shannon N Rhoads, Gül H Zerze, Robert O'Meally, Gregory L Dignon, Alexander E Conicella, Wenwei Zheng, Robert B Best, Robert N Cole, Jeetain Mittal, Frank Shewmaker, and Nicolas L Fawzi. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *The EMBO Journal*, 36(20):2951–2967, August 2017. doi:10.15252/embj.201696394.

[30] Xiangze Zeng and Rohit V Pappu. Developments in describing equilibrium phase transitions of multivalent associative macromolecules. *Current Opinion in Structural Biology*, 79:102540, 2023.

[31] Benjamin R. Sabari, Alessandra Dall'Agnese, and Richard A. Young. Biomolecular condensates in the nucleus. *Trends in Biochemical Sciences*, 45(11):961–977, November 2020. doi:10.1016/j.tibs.2020.06.007.

[32] Ian Diner, Chadwick M. Hales, Isaac Bishof, Lake Rabenold, Duc M. Duong, Hong Yi, Oskar Laur, Marla Gearing, Juan Troncoso, Madhav Thambisetty, James J. Lah, Allan I. Levey, and Nicholas T. Seyfried. Aggregation properties of the small nuclear ribonucleoprotein u1-70k in alzheimer disease. *Journal of Biological Chemistry*, 289(51):35296–35313, December 2014. doi:10.1074/jbc.m114.562959.

[33] Jamie A. Greig, Tu Anh Nguyen, Michelle Lee, Alex S. Holehouse, Ammon E. Posey, Rohit V. Pappu, and Gregory Jedd. Arginine-enriched mixed-charge domains provide cohesion for nuclear speckle condensation. *Molecular Cell*, 77(6):1237–1250.e4, March 2020. doi:10.1016/j.molcel.2020.01.025.

[34] Fiona T. M. Chang, James D. McGhie, F. Lyn Chan, Michelle C. Tang, Melissa A. Anderson, Jeffrey R. Mann, K. H. Andy Choo, and Lee H. Wong. PML bodies provide an important platform for the maintenance of telomeric chromatin integrity in embryonic stem cells. *Nucleic Acids Research*, 41(8):4447–4458, February 2013. doi:10.1093/nar/gkt114.

[35] Heankel Lyons, Reshma T. Veettil, Prashant Pradhan, Christy Fornero, Nancy De La Cruz, Keiichi Ito, Mikayla Eppert, Robert G. Roeder, and Benjamin R. Sabari. Functional partitioning of transcriptional regulators by patterned charge blocks. *Cell*, 186(2):327–345.e28, January 2023. doi:10.1016/j.cell.2022.12.013.

[36] Celeste J. Brown, Sachiko Takayama, Andrew M. Campen, Pam Vise, Thomas W. Marshall, Christopher J. Oldfield, Christopher J. Williams, and A. Keith Dunker. Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of Molecular Evolution*, 55(1):104–110, July 2002. doi:10.1007/s00239-001-2309-6.

[37] C. J. Brown, A. K. Johnson, and G. W. Daughdrill. Comparing models of evolution for ordered and disordered proteins. *Molecular Biology and Evolution*, 27(3):609–621, November 2009. doi:10.1093/molbev/msp277.

[38] Lucas Sawle and Kingshuk Ghosh. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *The Journal of Chemical Physics*, 143(8):085101, August 2015. doi:10.1063/1.4929391.

[39] Rahul K. Das and Rohit V. Pappu. 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, 2013. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1304749110, doi:10.1073/pnas.1304749110.

[40] Taraneh Zarin, Bob Strome, Alex N Nguyen Ba, Simon Alberti, Julie D Forman-Kay, and Alan M Moses. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife*, 8, July 2019. doi:10.7554/elife.46883.

[41] Wenwei Zheng, Gregory Dignon, Matthew Brown, Young C. Kim, and Jeetain Mittal. Hydropathy patterning complements charge patterning to describe conformational preferences of disordered proteins. *The Journal of Physical Chemistry Letters*, 11(9):3408–3415, March 2020. doi:10.1021/acs.jpclett.0c00288.

[42] Jonathan Huihui and Kingshuk Ghosh. Intrachain interaction topology can identify functionally similar intrinsically disordered proteins. *Biophysical Journal*, 120(10):1860–1868, May 2021. doi:10.1016/j.bpj.2020.11.2282.

[43] Jonas Wessén, Suman Das, Tanmoy Pal, and Hue Sun Chan. Analytical formulation and field-theoretic simulation of sequence-specific phase separation of protein-like heteropolymers with short- and long-spatial-range interactions. *The Journal of Physical Chemistry B*, 126(45):9222–9245, November 2022. doi:10.1021/acs.jpcb.2c06181.

[44] Timothy J. Nott, Evangelia Petsalaki, Patrick Farber, Dylan Jervis, Eden Fussner, Anne Plochowietz, Timothy D. Craggs, David P. Bazett-Jones, Tony Pawson, Julie D. Forman-Kay, and Andrew J. Baldwin. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Molecular Cell*, 57(5):936–947, March 2015. doi:10.1016/j.molcel.2015.01.013.

[45] Chi W. Pak, Martyna Kosno, Alex S. Holehouse, Shae B. Padrick, Anuradha Mittal, Rustam Ali, Ali A. Yunus, David R. Liu, Rohit V. Pappu, and Michael K. Rosen. Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. *Molecular Cell*, 63(1):72–85, July 2016. doi:10.1016/j.molcel.2016.05.042.

[46] Hiroya Yamazaki, Masatoshi Takagi, Hidetaka Kosako, Tatsuya Hirano, and Shige H. Yoshimura. Cell cycle-specific phase separation regulated by protein charge blockiness. *Nature Cell Biology*, 24(5):625–632, May 2022. doi:10.1038/s41556-022-00903-1.

[47] Benjamin S. Schuster, Gregory L. Dignon, Wai Shing Tang, Fleurie M. Kelley, Aishwarya Kanchi Ranganath, Craig N. Jahnke, Alison G. Simpkins, Roshan Mammen Regy, Daniel A. Hammer, Matthew C. Good, and Jeetain Mittal. Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proceedings of the National Academy of Sciences*, 117(21):11421–11431, May 2020. doi:10.1073/pnas.2000223117.

[48] Yi-Hsuan Lin and Hue Sun Chan. Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophysical Journal*, 112(10):2043–2046, May 2017. doi:10.1016/j.bpj.2017.04.021.

[49] Pouria Dasmeh, Roman Doronin, and Andreas Wagner. The length scale of multivalent interactions is evolutionarily conserved in fungal and vertebrate phase-separating proteins. *Genetics*, 220(1), October 2021. doi:10.1093/genetics/iyab184.

[50] Nicole DelRosso, Josh Tycko, Peter Suzuki, Cecelia Andrews, Aradhana, Adi Mukund, Ivan Liongson, Connor Ludwig, Kaitlyn Spees, Polly Fordyce, Michael C. Bassik, and Lacramioara Bintu. Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature*, April 2023. doi:10.1038/s41586-023-05906-y.

[51] Kristoffer E. Johansson, Bayan Mashahreh, Rasmus Hartmann-Petersen, Tommer Ravid, and Kresten Lindorff-Larsen. Prediction of quality-control degradation signals in yeast proteins. *Journal of Molecular Biology*, 435(2):167915, January 2023. doi:10.1016/j.jmb.2022.167915.

[52] Kiersten M. Ruff. Predicting conformational properties of intrinsically disordered proteins from sequence. In *Methods in Molecular Biology*, pages 347–389. Springer US, 2020. doi:10.1007/978-1-0716-0524-0_18.

[53] Jeffrey M. Lotthammer, Garrett M. Ginell, Daniel Griffith, Ryan J. Emenecker, and Alex S. Holehouse. Direct prediction of intrinsically disordered protein conformational properties from sequence. *bioRxiv*, 2023.

[54] Eric K Fredrickson, Joel C Rosenbaum, Melissa N Locke, Thomas I Milac, and Richard G Gardner. Exposed hydrophobicity is a key determinant of nuclear quality control degradation. *Molecular biology of the cell*, 22(13):2384–2395, 2011.

[55] Bayan Mashahreh, Shir Armony, Kristoffer Enøe Johansson, Alon Chappleboim, Nir Friedman, Richard G. Gardner, Rasmus Hartmann-Petersen, Kresten Lindorff-Larsen, and Tommer Ravid. Conserved degronome features governing quality control associated proteolysis. *Nature Communications*, 13(1), December 2022. doi:10.1038/s41467-022-35298-y.

[56] Mario Vazdar, Jan Heyda, Philip E. Mason, Giulio Tesei, Christoph Allolio, Mikael Lund, and Pavel Jungwirth. Arginine "magic": Guanidinium like-charge ion pairing from aqueous salts to cell penetrating peptides. *Accounts of Chemical Research*, 51(6):1455–1464, May 2018. doi:10.1021/acs.accounts.8b00098.

[57] Matteo Paloni, Giovanni Bussi, and Alessandro Barducci. Arginine multivalency stabilizes protein/RNA condensates. *Protein Science*, 30(7):1418–1426, May 2021. doi:10.1002/pro.4109.

[58] Alex X. Lu, Amy X. Lu, Iva Pritišanac, Taraneh Zarin, Julie D. Forman-Kay, and Alan M. Moses. Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning. *PLOS Computational Biology*, 18(6):e1010238, June 2022. doi:10.1371/journal.pcbi.1010238.

[59] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/a:1010933404324.

[60] Hermann Broder Schmidt, Ariana Barreau, and Rajat Rohatgi. Phase separation-deficient TDP43 remains functional in splicing. *Nature Communications*, 10(1), October 2019. doi:10.1038/s41467-019-12740-2.

[61] Jianhong Zhou, Suwen Zhao, and A. Keith Dunker. Intrinsically disordered proteins link alternative splicing and post-translational modifications to complex cell signaling and regulation. *Journal of Molecular Biology*, 430(16):2342–2359, 2018. Intrinsically Disordered Proteins. doi:https://doi.org/10.1016/j.jmb.2018.03.028.

[62] Estella A. Newcombe, Elise Delaforge, Rasmus Hartmann-Petersen, Karen Skriver, and Birthe B. Kragelund. How phosphorylation impacts intrinsically disordered proteins and their function. *Essays in Biochemistry*, 66(7):901–913, December 2022. doi:10.1042/ebc20220060.

[63] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, November 2017. doi:10.1093/nar/gkx1153.

[64] Axel Schmidt, Sebastian Röner, Karola Mai, Hannah Klinkhammer, Martin Kircher, and Kerstin U. Ludwig. Predicting the pathogenicity of missense variants using features derived from alphafold2. *bioRxiv*, 2022. doi:10.1101/2022.03.05.483091.

[65] Brian Tsang, Iva Pritišanac, Stephen W. Scherer, Alan M. Moses, and Julie D. Forman-Kay. Phase separation as a missing mechanism for interpretation of disease mutations. *Cell*, 183(7):1742–1756, December 2020. doi:10.1016/j.cell.2020.11.050.

[66] Shehab S Ahmed, Zaara T Rifat, Ruchi Lohia, Arthur J Campbell, A Keith Dunker, M Sohel Rahman, and Sumaiya Iqbal. Characterization of intrinsically disordered regions in proteins informed by human genetic diversity. *PLoS Computational Biology*, 18(3):e1009911, 2022.

[67] Avinash Patel, Hyun O. Lee, Louise Jawerth, Shovamayee Maharana, Marcus Jahnel, Marco Y. Hein, Stoyno Stoynov, Julia Mahamid, Shambaditya Saha, Titus M. Franzmann, Andrej Pozniakovski, Ina Poser, Nicola Maghelli, Loic A. Royer, Martin Weigert, Eugene W. Myers, Stephan Grill, David Drechsel, Anthony A. Hyman, and Simon Alberti. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell*, 162(5):1066–1077, August 2015. doi:10.1016/j.cell.2015.07.047.

[68] Amandine Molliex, Jamshid Temirov, Jihun Lee, Maura Coughlin, Anderson P. Kanagaraj, Hong Joo Kim, Tanja Mittag, and J. Paul Taylor. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*, 163(1):123–133, September 2015. doi:10.1016/j.cell.2015.09.015.

[69] Salman F. Banani, Lena K. Afeyan, Susana W. Hawken, Jonathan E. Henninger, Alessandra Dall'Agnese, Victoria E. Clark, Jesse M. Platt, Ozgur Oksuz, Nancy M. Hannett, Ido Sagi, Tong Ihn Lee, and Richard A. Young. Genetic

variation associated with condensate dysregulation in disease. *Developmental Cell*, 2022. doi:https://doi.org/10.1016/j.devcel.2022.06.010.

[70] Martin A. Mensah, Henri Niskanen, Alexandre P. Magalhaes, Shaon Basu, Martin Kircher, Henrike L. Sczakiel, Alisa M. V. Reiter, Jonas Elsner, Peter Meinecke, Saskia Biskup, Brian H. Y. Chung, Gregor Dombrowsky, Christel Eckmann-Scholz, Marc Phillip Hitz, Alexander Hoischen, Paul-Martin Holterhus, Wiebke Hülsemann, Kimia Kahrizi, Vera M. Kalscheuer, Anita Kan, Mandy Krumbiegel, Ingo Kurth, Jonas Leubner, Ann Carolin Longardt, Jörg D. Moritz, Hossein Najmabadi, Karolina Skipalova, Lot Snijders Blok, Andreas Tzschach, Eberhard Wiedersberg, Martin Zenker, Carla Garcia-Cabau, René Buschow, Xavier Salvatella, Matthew L. Kraushar, Stefan Mundlos, Almuth Caliebe, Malte Spielmann, Denise Horn, and Denes Hnisz. Aberrant phase separation and nucleolar dysfunction in rare genetic diseases. *Nature*, February 2023. doi:10.1038/s41586-022-05682-1.

[71] Carla Garcia-Cabau, Anna Bartomeu, Andreea Balaceanu, Giulio Tesei, Berta Duran-Arqué, Marcos Fernández-Alfara, Judit Martín, Cesare De Pace, Lorena Ruiz-Pérez, Jesús García, Giuseppe Battaglia, Kresten Lindorff-Larsen, Raúl Méndez, and Xavier Salvatella. Kinetic stabilization of translation-repression condensates by a neuron-specific microexon. *bioRxiv*, 2023. doi:10.1101/2023.03.19.532587.

[72] Reese AK Richardson, Heliodoro Tejedor Navarro, Luis A Nunes Amaral, and Thomas Stoeger. Data-driven design of a tool to promote the investigation of understudied genes. *bioRxiv*, 2023. URL: https://www.biorxiv.org/content/early/2023/03/01/2023.02.28.530483, doi:10.1101/2023.02.28.530483.

[73] Joao Rocha, Satish Arcot Jayaram, Tim J. Stevens, Nadine Muschalik, Rajen D. Shah, Sahar Emran, Cristina Robles, Matthew Freeman, and Sean Munro. Functional unknomics: closing the knowledge gap to accelerate biomedical research. *bioRxiv*, 2022. doi:10.1101/2022.06.28.497983.

[74] Tzu-Hsuan Chao, Shiv Rekhi, Jeetain Mittal, and Daniel P Tabor. Data-driven models for predicting intrinsically disordered protein polymer physics directly from composition or sequence. *ChemRxiv*, 2023. doi:10.26434/chemrxiv-2023-wrnq1.

[75] Giacomo Janson, Gilberto Valdes-Garcia, Lim Heo, and Michael Feig. Direct generation of protein conformational ensembles via machine learning. *Nature Communications*, 14(1):774, 2023.

[76] Maarten L. Hekkelman, Ida de Vries, Robbie P. Joosten, and Anastassis Perrakis. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature Methods*, 20(2):205–213, November 2022. doi:10.1038/s41592-022-01685-y.

[77] Joshua A. Anderson, Jens Glaser, and Sharon C. Glotzer. HOOMD-blue: A python package for high-performance molecular dynamics and hard particle monte carlo simulations. *Computational Materials Science*, 173:109363, February 2020. doi:10.1016/j.commatsci.2019.109363.

[78] H. Flyvbjerg and H. G. Petersen. Error estimates on averages of correlated data. *The Journal of Chemical Physics*, 91(1):461–466, July 1989. doi:10.1063/1.457480.

[79] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528 – 1532, 2015. doi:10.1016/j.bpj.2015.08.015.

[80] Alex S. Holehouse, Rahul K. Das, James N. Ahad, Mary O.G. Richardson, and Rohit V. Pappu. CIDER: Resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophysical Journal*, 112(1):16–21, January 2017. doi:10.1016/j.bpj.2016.11.3200.

[81] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[82] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, January 1995. doi:10.1111/j.2517-6161.1995.tb02031.x.

[83] Chang Y.-W., Hsieh C.-J., Chang K.-W., M. Ringgaard, and C.-J. Lin. Training and testing low-degree polynomial data mappings via linear svm. *Journal of Machine Learning Research*, 11:1471–1490, 2010.

[84] B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[85] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[86] Johanna Katarina Sofie Tiemann, Henrike Zschach, Kresten Lindorff-Larsen, and Amelie Stein. Interpreting the molecular mechanisms of disease variants in human transmembrane proteins. *Biophysical Journal*, January 2023. doi:10.1016/j.bpj.2022.12.031.

[87] G Jannink and J des Cloizeaux. Polymers in solution. *Journal of Physics: Condensed Matter*, 2(1):1–24, January 1990. doi:10.1088/0953-8984/2/1/001.

[88] J.A. Aronovitz and D.R. Nelson. Universal features of polymer shapes. *Journal de Physique*, 47(9):1445–1456, 1986. doi:10.1051/jphys:019860047090144500.