DIRECT PREDICTION OF INTRINSICALLY DISORDERED PROTEIN CONFORMATIONAL PROPERTIES FROM SEQUENCE

Version 2 [2023-05-11]

Jeffrey M. Lotthammer[1,2,*], Garrett M. Ginell[1,2,*], Daniel Griffith[1,2,*], Ryan J. Emenecker[1,2], Alex S. Holehouse[1,2]

1 - Department of Biochemistry and Molecular Biophysics Washington University School of Medicine, St. Louis, MO, USA
2 – Center for Biomolecular Condensates, Washington University in St. Louis, St. Louis, MO, USA

* These authors contributed equally
Correspondence: alex.holehouse@wustl.edu

ABSTRACT

Intrinsically disordered proteins and protein regions (IDPs/IDRs - hereafter IDRs) are ubiquitous across all domains of life. Unlike their folded domain counterparts, IDRs sample a diverse ensemble of conformations - rapidly interconverting between heterogeneous states. Analogous to how folded proteins adhere to a sequence-structure-function relationship, IDRs follow a sequence-ensemble-function paradigm. While experimental methods to study the conformational properties of IDRs exist, they can be challenging, time-consuming, and often require specialized equipment and expertise. Recent methodological advances in biophysical modeling offer a unique opportunity to explore sequence ensemble relationships; however, these methods are often limited in throughput and require both software and technical expertise. In this work, we integrated rational sequence design, large-scale molecular simulations, and deep learning to develop ALBATROSS, a deep learning model for predicting IDR ensemble dimensions from sequence. ALBATROSS is lightweight, easy to use, and readily accessible as both a locally-installable software package, as well as a point-and-click style interface in the cloud. We first demonstrate the applicability of our predictors by examining the generalizability of sequence-ensemble relationships in IDRs. Then, we leverage the high-throughput nature of our networks to characterize emergent biophysical behavior of both local and global IDR ensemble features across the human proteome.

ASAPBio statement

This preprint is following the ASAPBio philosophy of communicating and sharing new results at the speed at which they emerge. The results and tools presented in this preprint are robust, but we are continuing to iterate on improving the accuracy and stability of the methods presented and expanding the scope of the analyses our new tools enable. As a result, future versions of this manuscript will report differences in training data, predictor accuracy, and new analyses.

INTRODUCTION

Intrinsically disordered proteins and protein regions (IDRs) make up an estimated 30% of most eukaryotic proteomes and play a variety of roles in molecular and cellular function[1–4]. Although folded domains are often well-described by a single (or small number of) three-dimensional (3D) structures, IDRs are defined by extensive conformational heterogeneity. This means they exist in a conformational ensemble - a collection of rapidly interconverting states that prohibits structural classification by any single reference structure[5,6]. This heterogeneity challenges many experimental, computational, and conceptual approaches developed for folded domains, necessitating the application of polymer physics to describe, classify, and interpret IDRs in a variety of contexts[7–18].

Although IDRs are defined by the absence of a defined folded state, they are not "unstructured"[19]. The same chemical moieties that drive protein folding and enable molecular recognition in folded domains are also found within IDRs. As such, while folded domains subscribe to a sequence-structure relationship, IDRs have an analogous sequence-ensemble relationship[19]. Over the last fifteen years, there has been a substantial effort to decode the mapping between IDR sequence and conformational properties, the so-called 'sequence-ensemble relationship' [5,12,19–35].

IDR conformational properties can be local or global. Local conformational properties typically involve transient secondary structure, especially transient helicity[36]. Global conformational properties report on ensemble-average dimensions - that is, the overall size and shape that the ensemble occupies[5,19,36,37]. Two common properties measured by both experiment and simulation are the radius of gyration (R_g) and end-to-end distance (R_e). The R_g reports on the volume an ensemble occupies, while the R_e reports on the average distance between the first and the last residue. Ensemble shape can be quantified in terms of asphericity, a parameter that lies between 0 (sphere) and 1 (prolate ellipsoid), and reports on how spherical an ensemble is. While R_e, R_g, and asphericity are relatively coarse-grain, they can offer insight into the molecular conformations accessible to an IDR, as well as provide hints at the types of intramolecular interactions which may also be relevant for intermolecular interactions (especially in the context of low-complexity sequences)[23,38,39].

An *in vitro* assessment of sequence-ensemble relationships involves expression, purification, and measurement of ensemble properties using various biophysical techniques[16,40,41]. The experimental methods commonly used to study conformational properties include single-molecule fluorescence spectroscopy, nuclear magnetic resonance (NMR) spectroscopy, and small angle X-ray scattering (SAXS) [16,40–42]. While powerful, all three of these approaches can be technically demanding, necessitate access to specific instrumentation, and in the case of NMR and SAXS, require relatively high concentrations of protein. Beyond *in vitro* assessment, integrating all-atom simulations with biophysical measurements has proven invaluable in obtaining a holistic description of sequence-ensemble relationships, yet these integrative studies can also be challenging [22,23,25,28,43–47]. As such, obtaining insight into sequence-specific conformational biases for disordered proteins is often challenging for groups with a limited background in molecular biophysics.

Recent efforts have led to a marked improvement in the accuracy of coarse-grained force fields for disordered protein simulations [48–53]. In particular, simulations performed with the CALVADOS and Mpipi force fields offer robust predictions of global conformational properties for disordered proteins. However, setting up, running, and analyzing molecular simulations necessitate a level of expertise and resources beyond many (arguably most) research groups. As such, the democratization of exploring sequence-to-ensemble relationships in disordered proteins demands easy-to-use tools that are readily accessible (i.e., available in a web browser without any hardware constraints).

Here, we address this gap by developing a rapid and accurate predictor for disordered protein global dimensions from sequences. We do this through a combination of rational sequence design, large-scale coarse-grained simulations, and deep learning (**Fig. 1A**). The resulting predictor (ALBATROSS; A deep-Learning Based Approach for predicTing pRoperties Of diSordered proteinS) not only pushes the boundaries of acronym development but provides a means to predict IDR global dimensions (R_g, R_e, asphericity) directly from sequence.

ALBATROSS was developed with ease of use and portability in mind; no specific hardware is required, and predictions can be performed on either CPUs or GPUs. We provide both a locally-installable implementation of ALBATROSS as well as point-and-click Google Colab notebooks that enable predictions to be performed on 30-60 sequences per second on a CPU and thousands of sequences per second on a GPU (**Fig. 1B**). In this work, we use ALBATROSS to demonstrate the generality of core sequence-ensemble relationships identified by foundational prior work, as well as assess general conformational biases observed at proteome-wide scales. Lastly, we propose that local conformational behavior offers a route to discretize IDRs into conformationally-distinct subdomains.

As a final note, while we rigorously validate the accuracy of ALBATROSS against both simulated and experimental data, we do not see it as a replacement for well designed simulation or experimental studies. Instead, our goal is for ALBATROSS to estimate emergent biophysical properties from IDR encoded sequence chemistry to aid in hypothesis generation as well as the interpretation and design of experiments.

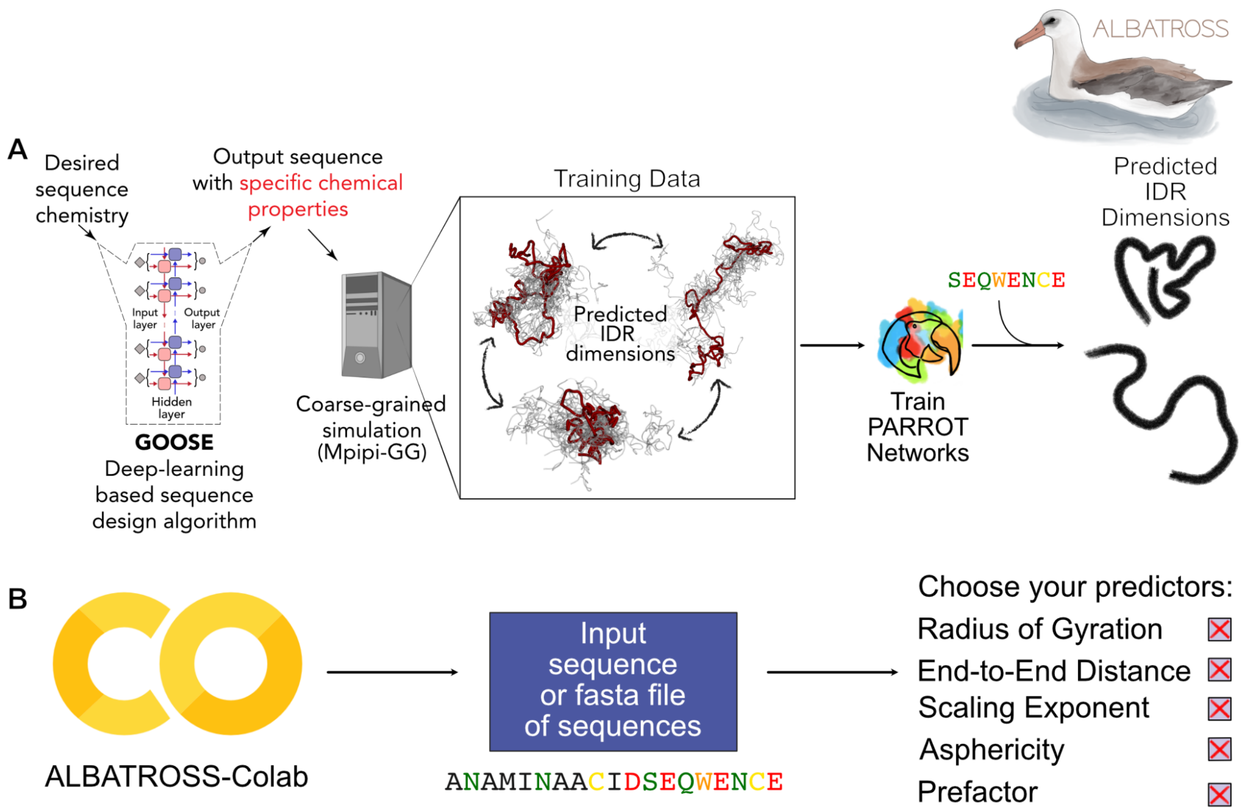

Figure 1. ALBATROSS is a deep-learning framework for predicting sequence-dependent IDR ensemble properties. A) Sequence design and simulation approach to generate training data for ALBATROSS networks. The Python package GOOSE is used to generate synthetic IDRs across a diverse area of sequence space. Coarse-grained molecular dynamics simulations are performed for each sequence to generate labeled data for downstream deep neural network training and validation. **B)** ALBATROSS is implemented as a point-and-click style interface on Google Colaboratory with support for CPU and GPU inference. The user simply specifies the amino acid sequence or a fasta file of amino acid sequences and then selects the predictions they would like to perform.

METHODS

The overall approach for developing ALBATROSS involved several steps. First, we generated a library of synthetic disordered proteins that systematically titrated across compositional space using our artificial disordered protein design package GOOSE[54]. Next, we fine-tuned the Mpipi force field, making small changes to the previously published parameters to address minor shortcomings, leading to a version we refer to as Mpipi-GG[49]. We then performed simulations of the synthetic training sequences using Mpipi-GG and calculated ensemble-average parameters[55]. Finally, we trained bidirectional recurrent neural networks with long short-term memory cells (LSTM-BRNNs) to map between amino acid sequence and simulation-derived ensemble-average parameters[56]. Network weights, along with software to perform sequence-ensemble predictions, were then packaged into our sequence analysis package SPARROW and via an easy-to-use Google Colab notebook[57].

Sequence Library Design

Using the IDR design package GOOSE, we assembled a library of chemically diverse synthetic disordered proteins (https://github.com/idptools/goose)[54]. Sequences varied charge, hydropathy, and charge patterning, as well as titrated across the amino acid composition. All sequences generated were between 10 and 750 residues in length. In total, we generated 16,885 disordered protein sequences across a diverse sequence space (see *Supplementary Information*).

Biological Sequences for Validation

In addition to the synthetic sequence library, we curated a set of 19,075 naturally occurring IDRs by randomly sampling disordered proteins ranging in length from 10-750 residues from one of each of the following proteomes: *Homo sapien, Mus musculus, Dictyostelium discoideum, Escherichia coli, Drosophilia melanogaster, Saccharomyces cerevisiae, Neurospora crassa, Schizosaccharomyces pombe, Xenopus laevis, Caenorhabditis elegans, Arabidopsis thaliana, and Danio rerio.* All annotated IDRs from the aforementioned proteomes are available at https://github.com/holehouse-lab/shephard-data/tree/main/data/proteomes.

Coarse-Grained Simulations

All reported simulations were performed with the LAMMPS simulation engine and either the newly parameterized Mpipi-GG or Mpipi (for comparison to Mpipi-GG) force fields[49,58]. Initial disordered protein starting configurations were built by assembling beads as a random coil in the excluded volume limit. Each simulation was minimized for a maximum of 1000 iterations or until the force tolerance was below 1 x 10^{-8} (kcal/mol)/Å. All simulations were performed with 150 mM implicit salt concentration in the canonical (NVT) ensemble at a target temperature of 300 K. The simulation temperature was maintained with a weakly-coupled Langevin thermostat that is adjusted every 100 picoseconds, and an integration timestep of 20 femtoseconds for all production runs. Simulations were performed with periodic boundary conditions in a 500 $Å^3$

cubic box. Output coordinates for each trajectory were saved every 2 nanoseconds. All simulations were initially equilibrated for 10 ns, and structures from this equilibration period were discarded. Production simulations of disordered sequences with less than 250 residues were performed for 6 μs, whereas sequences greater than 250 residues were simulated for 10 μs. In terms of LAMMPS simulation parameters, these settings reflect saving IDR conformations every 1 x 10^5 simulation steps, discarding the first 5 x 10^5 simulation steps as equilibration, and performing simulations for 3 x 10^8 steps for short sequences and 1 x 10^9 steps for long sequences. Simulation analysis was performed using SOURSOP and MDTraj [55,59].

Deep Learning

We leveraged Bidirectional Recurrent Neural Networks with Long Short-Term Memory cells (BRNN-LSTM) for all sequence-to-ensemble property prediction tasks with the flexible recurrent neural network framework PARROT[56]. We generated training, validation, and test data from coarse-grained simulations performed with the Mpipi-GG force field.

Specifically, we developed predictors for the radius of gyration (R_g), end-to-end distance (R_e), and asphericity, along with the polymer scaling law prefactors and scaling exponents [60,61]. For each of these ensemble and polymeric property prediction tasks, we split the synthetic IDR data randomly into three sets: a training, validation, and a held-out test set via a 70%, 15%, and 15% random split, respectively. Following previous PARROT network protocols, we employed a one-hot encoding scheme to translate the protein sequence data into numerical vectors amenable for deep neural network training. We used a training objective that sought to minimize an L1 loss function between the predictions and labeled data for each of the sequence-to-ensemble property predictors. For each network, we chose a default learning rate of 0.001, and we performed a hyperparameters grid search over the following parameters: batch size (8 to 32 incrementing by powers of 2), number of hidden layers (1 to 4), and a hidden dimension size (10 to 70). We selected the optimal hyperparameters for each network by monitoring the predictive performance of the held-out synthetic data. To evaluate the generalization error of our models on sequences relevant to biological function, we evaluated the most accurate networks for each predictor using the entire set of naturally occurring biological sequences.

Bioinformatics

Proteome-wide bioinformatics was performed using SPARROW (https://github.com/idptools/sparrow) and SHEPHARD[62]. SPARROW is an in-development Python package for calculating IDR sequence properties, while SHEPHARD is a hierarchical analysis framework for annotating and analyzing large sets of protein sequences. IDRs and proteome data are available at https://github.com/holehouse-lab/shephard-data. Disordered regions were predicted using metapredict (V2), and proteomes were obtained from UniProt[63,64]. For Fig. 5 and 6, predictions were performed using the standard R_g prediction network (`predictor.radius_of_gyration()`). For Fig. 7, local subregions were predicted using the end-to-end distance predictor from the scaled network (`predictor.end_to_end_distance(use_scaled=True)`), a decision we made as the scaled networks show slightly better performance for shorter IDRs. Normalized chain dimensions (normalized R_e and normalized R_g) were calculated as the ALBATROSS-predicted R_e or R_g divided by the Analytical Flory Random Coil (AFRC)-derived R_e or R_g. The AFRC is a model that reports on the sequence-specific chain dimensions expected if an IDR behaved as a Gaussian chain (i.e., a Flory scaling exponent of 0.5)[65].

ALBATROSS implementation and distribution

ALBATROSS is implemented within the SPARROW sequence analysis package (https://github.com/idptools/sparrow). In addition, a point-and-click style interface to ALBATROSS is provided via a stand-alone Google Colab notebook for both single-sequence and large-scale predictions of hundreds of sequences. If a FASTA file is uploaded and GPUs are selected, this notebook enables predictions for thousands of IDRs per second, facilitating in-browser proteome-wide analysis.

The notebook is available at:
https://colab.research.google.com/github/holehouse-lab/ALBATROSS-colab/blob/main/example_notebooks/polymer_property_predictors.ipynb

Finally, for IDRs predicted from protein sequences at https://metapredict.net/, the predicted R_g and R_e are also returned instantaneously.

Data and code availability

All code used for sequence analysis, training weights, bioinformatic data, the SPARROW implementation, and the Google Colab notebook are linked from this manuscript's main GitHub directory: https://github.com/holehouse-lab/supportingdata/tree/master/2023/ALBATROSS_2023

RESULTS

Our approach in developing ALBATROSS was to perform coarse-grained simulations of a set of training sequences that would enable an LSTM-BRNN model to learn the mapping between IDR sequence and global conformational behavior. To this end, four distinct phases in this process were required: (1) Selecting an appropriate force field, (2) Designing a library of synthetic sequences, (3) Performing simulations of those sequences, and 4) Optimizing deep learning models for sequence-to-ensemble mapping.

Force field selection and fine-tuning

Coarse-grained simulations were performed using Mpipi-GG, a fine-tuned version of the previously published Mpipi forcefield.. Mpipi is a one-bead-per-residue coarse-grained force field that was parameterized via a bottom-up, data-driven approach using statistics obtained from the PDB coupled with quantum mechanical calculations and all-atom simulations[49]. We (and others) have had great success in using Mpipi to provide molecular insight into a range of systems[66,67]. While Mpipi generally shows very good accuracy when compared with experiments, in performing initial calibration simulations, we noticed a few minor discrepancies between known experimental trends and Mpipi behavior (see *Supplementary Information* and **Fig. S1-S4**). Focusing on specific sets of interactions where the chemical basis for those discrepancies was interpretable, we made several small modifications to the underlying parameters, yielding a version of Mpipi we refer to as Mpipi-GG. The changes made to the interaction matrix can be visualized in **Fig. 2A**, which reports on the change in the overall interaction parameter between Mpipi-GG and Mpipi. The overall interaction parameter reports the net integral of both the short-range (Wang-Frenkel) and long-range (Coulombic) interaction potentials. We emphasize that these changes were made explicitly with single-chain behavior in mind and have not been tested in terms of their impact on phase behavior. As such, while the original Mpipi model may be preferable for studying two-phase systems, we proceeded to use Mpipi-GG for single-chain sequence-ensemble predictions.

To assess how Mpipi-GG differs in terms of overall accuracy compared to the original Mpipi parameters, we curated a set of 137 trusted radii of gyration obtained from the literature for disordered proteins that are diverse in sequence chemistry and complexity (**Fig. 2B**). Comparing the predictive power of Mpipi-GG to the original Mpipi force field for these sequences reveals comparable accuracy, with Mpipi-GG performing modestly better with an R^2 of 0.921 vs. 0.896 for Mpipi, although both models are highly accurate (**Fig. 2C-D**).

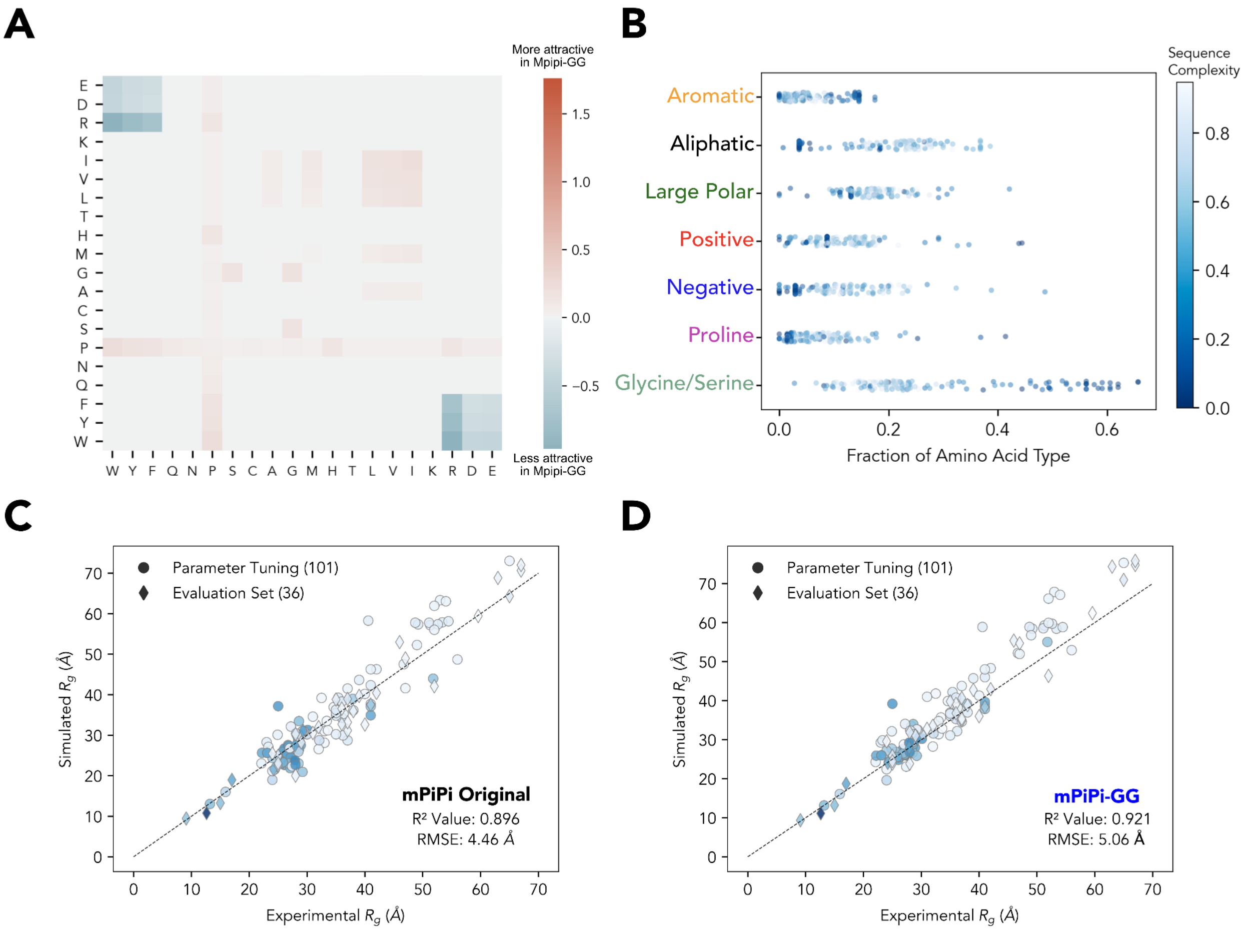

Figure 2. Reparameterization and accuracy of the Mpipi-GG force field. A) Pairwise interaction matrix for the reparameterized Mpipi-GG force field. Pairwise interactions are colored by the relative change in interaction energies between the Mpipi force field and the new Mpipi-GG force field. Interaction energies are the sum of the net contributions from both the pairwise Coulombic interactions as well as the pairwise WF interactions. **B)** Composition of the curated experimental SAXS sequence dataset by amino acid type. The blue color gradient signifies the Wootton-Federhen complexity of the sequence. **C-D)** Correlations and RMSEs between the original Mpipi and Mpipi-GG force fields and a curated set of 137 experimental radii of gyration. 101 sequences (circles) were used for validating the Mpipi-GG force field, and 36 were new sequences held-out during parameter fitting (diamonds). The same color gradient signifying sequence complexity used in B is used in panel C.

Design of a library of synthetic IDRs

Next, we created a library of artificially disordered regions that more extensively titrate across the relevant chemical space accessible to disordered proteins. To do this, we used GOOSE, our recently developed computational package for synthetic IDR design, to construct a library of 16,885 sequences that titrate across a range of sequence features known to impact IDR conformational behavior (see *Methods*). We reasoned that systematically exploring IDR sequence space would provide a good representation of the sequence chemistries relevant for disordered protein conformational behavior as opposed to training on biological sequences where intrinsic biases may limit the number of sequences with certain sequence compositions. Moreover, we opted to take advantage of GOOSE's ability to limit compositional exploration to sequences predicted to be disordered, such that our initial library is centered on sequences predicted with high confidence to be IDRs.

We first began by titrating sequences with different compositions of hydropathy and net charge per residue, two parameters known to alter average chain dimensions in IDRs (**Fig. 3A**). Next, we ensured that our sequence library had broad coverage of another important IDR sequence parameter kappa (κ), which describes the patterning of oppositely charged residues in a sequence, by systematically generating sequences that had varied fractions of charged residues each at different residue positions (**Fig. 3B**). Our synthetic disordered sequence library also covered broad chemical space in terms of the fraction of aliphatic and polar residues (**Fig. 3C**) as well as the fraction of positively charged residues and aromatic residues (**Fig. 3D**). An overview of the amino acid compositions and overall sequence complexity for the synthetic library is summarized in **Fig. 3E**. Moreover, we also ensured our sequence library had broad coverage of the sequence charge decoration (SCD) parameter defined by Sawle and Ghosh as well as the sequence hydropathy decoration parameters (SHD) (**Fig. S9**)[11,68].

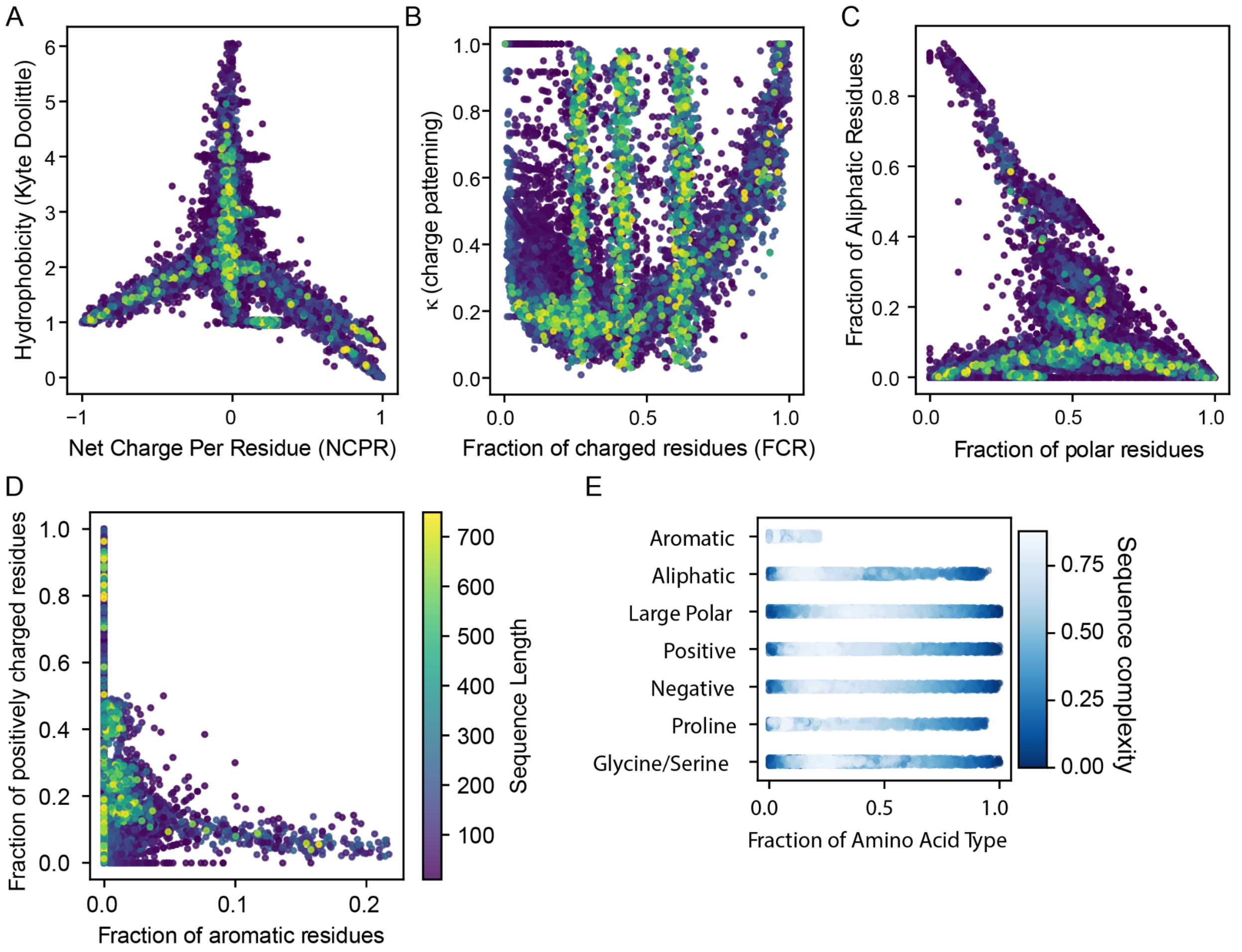

Figure 3. Composition of the synthetic IDR sequence library used for training. Two-dimensional scatter plots showing the chemical space explored by our synthetic IDR library. Each point in all panels is colored by the length of that particular sequence. **A)** Net charge per residue versus the Kyte Doolittle hydrophobicity of the sequence. **B)** Fraction of charged residues versus the charge patterning parameter kappa (κ). **C)** Fraction of polar residues versus the fraction of aliphatic residues in a given sequence. **D)** Fraction of aromatic residues and the fraction of positively charged residues (RK). **E)** Composition of the synthetic sequence library by amino acid type. The blue color gradient signifies the Wootton-Federhen complexity of the sequence.

Training an IDR sequence-to-ensemble deep learning model

After designing synthetic IDR sequences and both selecting and tuning our force field, we performed molecular dynamics simulations of all 16,885 sequences to examine their sequence-dependent ensemble features. Specifically, we focused on the radius of gyration, end-to-end distance, asphericity, and the scaling exponent and prefactor for the polymer scaling law fit the internal scaling data[23,69,70]. These data served as the foundation for training bidirectional recurrent neural networks with LSTM cells with PARROT for sequence-dependent property prediction tasks (see *Methods*). The collective group of these networks we term ALBATROSS.

We first began training the ALBATROSS R_g network. We leveraged the PARROT framework to train LSTM-based deep learning models on the simulated radius of gyration data. Promisingly, we saw a strong correlation on a synthetic sequence test set held out during the training (R^2 = 0.997, **Fig. 4A)**. We note that these synthetic sequences, which explore a more extreme and diverse region of disordered sequence space, deviate strongly from the expected radius of gyration obtained from the Analytical Flory Random Coil (AFRC), a Gaussian-chain-like model for disordered proteins (R^2 = 0.676, **Fig. S5A**). We next turned to evaluate the accuracy of our networks on the R_e prediction task. Similarly to the ALBATROSS R_g network, we observed a strong correlation between the ALBATROSS R_e and the Mpipi-GG R_e (R^2 = 0.994, **Fig. 4B**). The impact of sequence chemistry is more pronounced on the end-to-end distance than the radius of gyration, as illustrated by the weaker correlation between the ALBATROSS-predicted R_e and the AFRC-derived R_e (R^2 = 0.470, **Fig. S5B**).

In addition to these R_g and R_e networks, we also trained networks for the mean asphericity, which displayed strong quantitative agreement on the synthetic sequence prediction test set (R^2 = 0.956, **Fig. 4C**). Finally, we trained predictors based on the two parameters obtained by fitting the internal scaling of the beads to a polymer scaling model; the scaling exponent and prefactor. The accuracy of the predictions from these networks was comparably strong on the synthetic sequences with correlation coefficients of 0.978 and 0.924, respectively (**Fig. S6**).

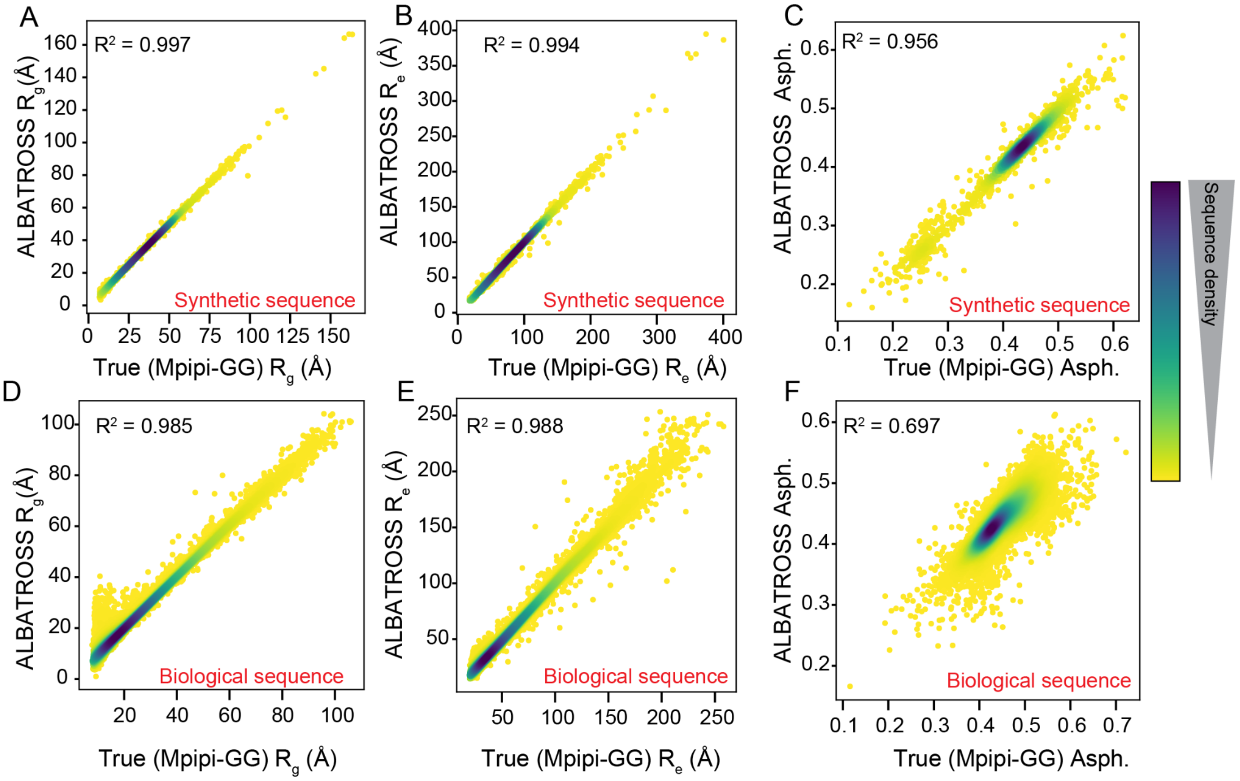

Figure 4. ALBATROSS network accuracy on two independent test sets of both synthetic sequences and naturally occurring biological sequences. A, B, C) Accuracy of ALBATROSS in predicting R_g, R_e, and asphericity for previously unseen synthetic sequences. **D, E, F**) Accuracy of ALBATROSS in predicting R_g, R_e, and asphericity for previously unseen biological sequences. For each correlation plot, a Gaussian kernel density estimation is used, where darker colors indicate regions where there are many sequences sharing a particular prediction value.

Evaluating our networks on naturally occurring biological sequences

While our networks were trained solely on synthetic disordered proteins, we hypothesized the sequence diversity explored during training would permit these networks to generalize to naturally occurring biological sequences. We first began by confirming the ALBATROSS radii of gyration match the Mpipi-GG radii of gyration for the same experimental data presented in **Fig. 2.** Indeed, we see strong quantitative agreement between ALBATROSS-derived radii of gyration and the experimental radii of gyration, despite the fact none of these sequences were in our training data (R^2 = 0.92, **Fig. S7**). Inspired by this result, we next sought to assess how accurately ALBATROSS was able to predict the simulated Mpipi-GG R_g values. To quantify the predictive power for this task, we randomly selected 19,075 predicted biological IDRs from several different model organism proteomes and performed coarse-grained molecular dynamics simulations for these sequences (see *Methods*). These IDRs served as an independent test set to evaluate the true generalization error of our tuned models in the sequence space relevant to biological function. Quantitative comparison of the Mpipi-GG radii of gyration to the

ALBATROSS R_g predictions on the biological sequences reveals excellent predictive power (R^2 = 0.985, **Fig. 4D**). We also test the accuracy of the ALBATROSS R_e network and observe a comparatively strongly predictive model (R^2 = 0.988, **Fig. 4E**).

The ALBATROSS asphericity networks performed moderately well on the biological test set (R^2 = 0.697, **Fig. 4F**). However, the correlation between the ALBATROSS ν network and the Mpipi-GG scaling exponent and scaling law prefactor were comparatively poor (R^2 = 0.358 and 0.347, **Fig. S6**). As such, while we are continuing to improve our ability to predict scaling exponents, we will focus on the three bulk polymeric properties (R_g, R_e, asphericity) from here on out.

ALBATROSS performance

While our three main networks are highly accurate, the primary benefit they provide is throughput for the systematic exploration of sequence-ensemble relationships. While coarse-grained simulations can take minutes, hours, or even days, ALBATROSS enables thousands of predictions per minute. A summary of our performance benchmarks on modest commodity CPU hardware is provided in **Fig. S8**, a criterion we focussed on, given many researchers do not have access to high-end GPUs. However, we note that one can compute R_g predictions for the entire human proteome in ~8 seconds via our Google Colab notebook running on GPUs. As such, ALBATROSS offers an accurate and high-performance route to map sequence-ensemble relationships for R_e, $R_{g,}$ and asphericity.

Systematic assessment of sequence-to-ensemble properties

Having developed and assessed ALBATROSS' ability to predict global dimensions from Mpipi-GG simulations, we used it to systematically interrogate sequence ensemble relationships. A common critique of machine learning approaches is that while they offer remarkable predictive power, they generally do not provide mechanistic insight. Here, we addressed this limitation here by returning to our sequence design approach to assess how systematic variation in different sequence parameters dictates global ensemble dimensions. Specifically, by generating thousands of synthetic disordered sequences that are well-controlled in terms of length and composition, we can systematically interrogate how individual sequence features influence global IDR dimensions.

Early work established that an IDR's absolute net charge strongly influences global dimensions[12,21,22,24]. In support of this, a systematic titration across the diagram of states developed by Das & Pappu revealed a strong dependence of IDR dimensions on the net charge per residue (**Fig. 5A, B**)[12]. For net neutral sequences (Net Charge Per Residue, NCPR ≈ 0), global dimensions were strongly influenced by the combination of the fraction of charged residues and the patterning of oppositely charged residues, quantified here by the parameter kappa (κ) (**Fig. 5C**)[12]. Sequences with evenly-distributed charged residues (low κ) become more expanded as the fraction of charged residues increases, whereas sequences with segregated charged residues (high κ) become more compact as the fraction of charged residues increases. While in line with prior computational and experimental observations, this analysis confirms the broad generality of these findings and provides calibration for expected dimensions given a sequence's composition [11,12,71].

We next systematically investigated the impact of different types of amino acids using libraries of sequences constructed using GOOSE, in which the overall sequence fraction of a specific residue was systematically titrated from 0% to 40%. While several caveats should be considered (see *Discussion*), this analysis provides a first-order approximation to the relative roles of different residues in an otherwise 'neutral' IDR background.

Increasing the fraction of aromatic amino acids lead to systematic chain compaction, with the rank order of W > Y > F in terms of the strength of interactions, in agreement with prior work establishing the relative strength of aromatic residues in the context of phase separation [23,72–74]. Increasing the fraction of aliphatic residues, at least to 40%, has a seemingly minimal impact on global dimensions, largely in agreement with work to date (**Fig. 5D**)[75–77]. Increasing the fraction of polar amino acids leads to modest compaction for Q and N, yet for T, S, and H, very little change in global dimensions is observed, in line with expectations from prior work on low-complexity polar-rich sequences (**Fig. 5D**)[20,35,76,78,79]. Finally, while glycine and cysteine have only a minor impact on global dimensions over the range explored, proline drives chain expansion, in agreement with previous studies (**Fig. 5D**)[25,28,32,80].

In addition to titrating the aromatic fraction, we designed synthetic repeat proteins consisting of glycine-serine-repeat "spacers" and poly-tyrosine "stickers" [72,81,82]. These synthetic IDRs allow us to assess how spacer length and sticker strength (tuned by the number of tyrosine residues in a sticker) influence chain dimensions. Our results demonstrate that both spacer length and sticker strength can synergistically influence IDR global dimensions (**Fig. 5E**). The dependence of the individual chain R_g on spacer length (y-axis) and sticker strength (x-axis) mirrors conclusions drawn from sticker-spacer architecture polymers from simulations and experiment [23,83–85].

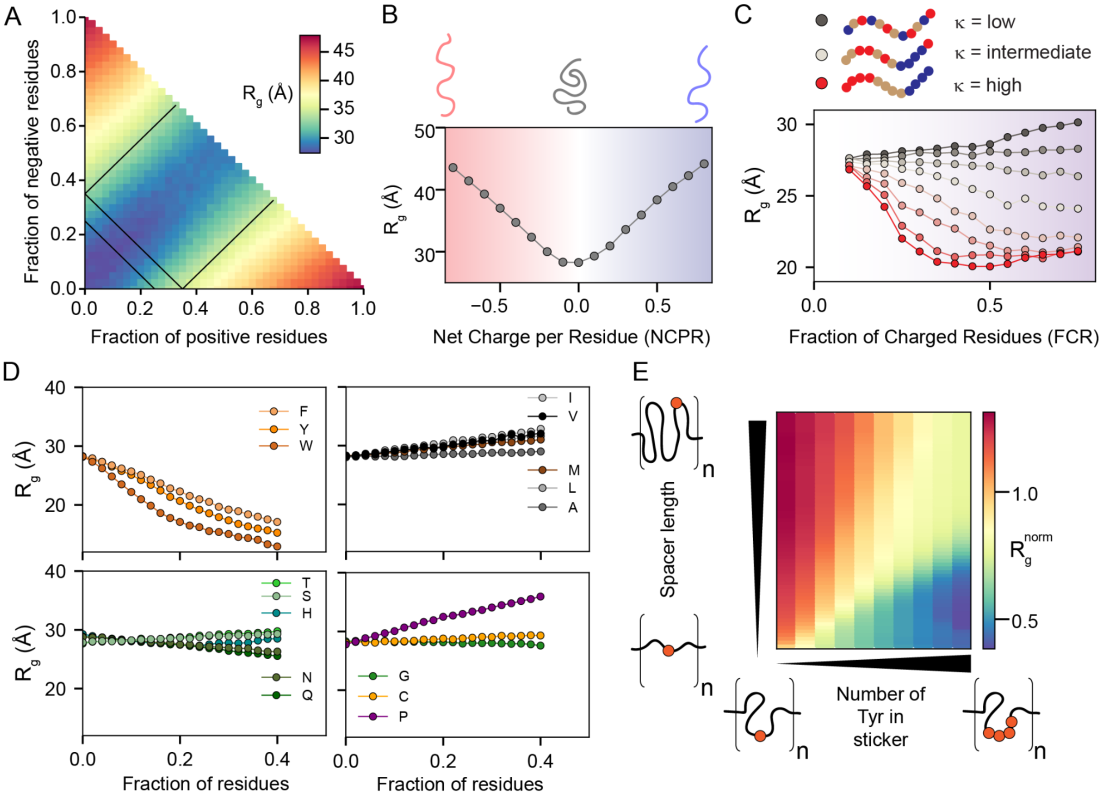

Figure 5. Sequence composition modulates the conformational preferences in disordered proteins. For panels A-D, each data point reports the average of many 100-residue synthetic disordered sequences with the specified composition. **A)** Diagram of states for weak to strong polyampholytes. Sequences are colored by a blue-to-yellow-to-red gradient based on their ALBATROSS radii of gyration. B) ALBATROSS radii of gyration as a function of net charge per residue. Both net negative (red) and net positive (blue) charged polyampholytes can drive chain expansion. **C)** The patterning of positively or negatively charged residues dictates the radius of gyration for highly-charged sequences but not those with a low fraction of charged residues. **D)** ALBATROSS radii of gyration as a function of the fraction of amino acid content for sixteen of the different amino acids. Aromatic residues drive compaction, while proline drives expansion. In each case, the fraction of other residues was held approximately fixed while one specific residue was systematically varied. **E)** Dependence of the normalized radius of gyration for sticker-spacer IDRs in which spacers are glycine serine repeats, and stickers are one or more tyrosine residues. The normalized radius of gyration is calculated as the ALBATROSS R_g divided by the R_g expected for a sequence-matched version of the protein behaving as a Gaussian chain (the AFRC model)[65]. Each sequence here contains 8 sticker-spacer repeats. Each repeat contains spacer regions (glycine-serine dipeptide repeats) that vary in length from 2 to 120 residues and sticker regions (poly-tyrosine repeats) that vary in length from 0 tyrosines to 8 tyrosines.

Predicting emergent biophysical properties throughout the human proteome

Given ALBATROSS' accuracy and throughput, we next performed large-scale bioinformatic characterization of the biophysical properties of disordered regions across the human proteome (**Fig. 6A, B**). Focusing initially on IDRs between 35 and 750 residues in length, we calculated normalized radii of gyration (**Fig. 6C**), normalized end-to-end distance (**Fig. 6D**), and asphericity (**Fig. 6E**). Normalization here was essential to account for the variability in absolute radii of gyration with sequence length, and was achieved by dividing the ALBATROSS R_g with the sequence-specific R_g expected if the IDR behaved as a Gaussian chain[65]. These analyses suggest that most IDRs behave as relatively expanded chains, although we recognize there are likely several important caveats to this interpretation (see *Discussion*). Assessing the absolute radius of gyration vs. IDR length, the majority of more compact IDRs are enriched for aromatic residues (**Fig. 6F**). Indeed, plotting the asphericity (a measure of IDR ensemble shape) vs. the normalized radius of gyration and coloring by either the fraction of aromatic residues (**Fig. 6G**) or the absolute net charge and the fraction of proline residues (**Fig. 6H**) suggest that IDRs with an ensemble that is expanded and elongated have a net charge and/or are enriched for proline, whereas IDRs with an ensemble that is compact and more spherical are enriched for aromatic residues. Segregating IDRs into the 1000 most compact and 1000 most expanded sequences reveals that compact IDRs tend to be depleted in proline residues and have a low NCPR, whereas those that are expanded are enriched in proline and/or have an absolute NCPR, although we found many examples of proline-rich charge depleted IDRs that were relatively expanded. Taken together, our analysis of the human IDR-ome mirrors insights gleaned from the analysis of synthetic sequences in **Fig. 5**.

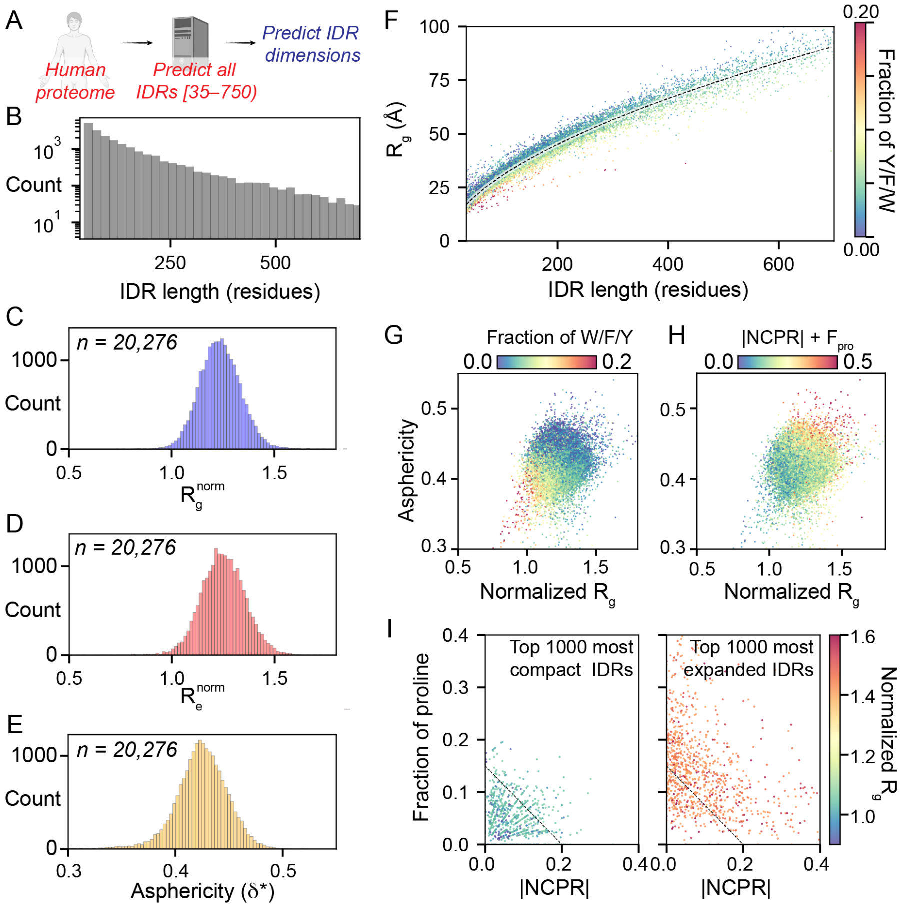

Figure 6. Human proteome-wide biophysical characterization of predicted IDRs.
A) ALBATROSS was used to perform sequence-dependent ensemble predictions for all IDRs in the human proteome. **B)** Histogram of all human IDRs ranging from 35 to 750 residues. **C)** Normalized mean ALBATROSS R_g distribution for all human IDRs. **D)** Normalized mean ALBATROSS R_e distribution for all human IDRs. **E)** Mean ALBATROSS asphericity distribution for all IDRs in the human proteome. **F)** Mean ALBATROSS radius of gyration as a function sequence length. Individual data points are colored by the fraction of aromatic residues in the sequence. The dashed line represents the fitted scaling law, which reports an apparent scaling exponent of 0.56. Deviations above and below this line suggest sequence-specific expansion or compaction, respectively. **G)** Full distribution of human IDRs plotted in terms of the normalized radius of gyration and asphericity, colored by the fraction of aromatic residues. **H)** Full distribution of human IDRs plotted in terms of the normalized radius of gyration and asphericity, colored by the absolute net charge per residue plus the fraction of proline residues. **I)** Top 1000 most compact (left) and top 1000 most expanded (right) IDRs plotted in terms of the fraction of proline residues and absolute net charge per residue.

Characterizing local dimensions of IDR subsequences

Our proteome-wide analysis in **Fig. 6** focused on ensemble-average properties calculated for entire IDRs. While convenient for revealing gross properties, we reasoned that for large (200+ residue) IDRs, it may be more informative to assess local conformational behavior with a sliding-window analysis. To this end, using a window size of 51 residues, we calculated the local end-to-end distance across every 51-mer fragment in the human proteome, enabling us to extract the 2,146,400 51-mer fragments that lay entirely within every IDR (**Fig. 7A**). The distribution of normalized end-to-end distances is tighter than the corresponding distribution for full-length IDRs, with 212,565 (i.e., around 10%) of subregions behaving as polymers more compact than a corresponding Gaussian chain (normalized $R_g < 1$) (**Fig. 7B**).

The linear assessment of local dimensions enables the demarcation of conformationally-distinct subdomains within an IDR. As a proof-of-concept, we plotted the normalized local end-to-end distance for two large IDRs, revealing distinct subregions within each (**Fig. 7C, D**). First, we analyzed the 2227 residue IDR from the nuclear speckle protein Son, identifying distinct subregions with specific conformational properties that map to previously analyzed subregions within the sequence (**Fig. 7C**)[86]. Second, we analyzed the N-terminal IDR of GIGYF1, a highly disordered protein with a potential role in Type II diabetes [87–89]. The N-terminal IDR in GIGYF1 contains three subregions, an expanded N-terminal region that may fold upon binding (residues 1-90), a comparatively compact central region (residues 91-280), and an expanded C-terminal acidic region (residues 281-469). The ability to – from sequence alone – demark potential subdomains within an IDR paves the way for more sophisticated mutagenesis studies, as well as the ability to predict if and how mutations might influence local conformational behavior.

Finally, we used the set of ~2 million IDR subregions to assess which residues were enriched in expanded or compact IDRs (**Fig. 7E, D**). Enrichment was assessed based on the fraction of the twenty amino acids in subregions taken from the top/bottom 2.5% of all subregions with respect to normalized end-to-end distance, compared to the overall fraction for all subregions. Aromatic residues, histidine, arginine, glycine, and glutamine were all found to be enriched in compact subregions. In contrast, proline and glutamic acid were found to be enriched for expanded subregions. Intriguingly, the residues most strongly enriched for compact IDRs match those residues known to engage in RNA binding[66,90–93]. Moreover, a gene ontology analysis for proteins with 10 or more compact subfragments found strong enrichment for RNA binding (**Table S1**). In contrast, we saw no obvious patterns in proteins that possessed expanded subregions (**Table S2**). Taken together, our analysis suggests IDRs that favor intramolecular interaction may share a common molecular function in RNA binding, whereas those that are highly expanded likely play a variety of context-specific roles.

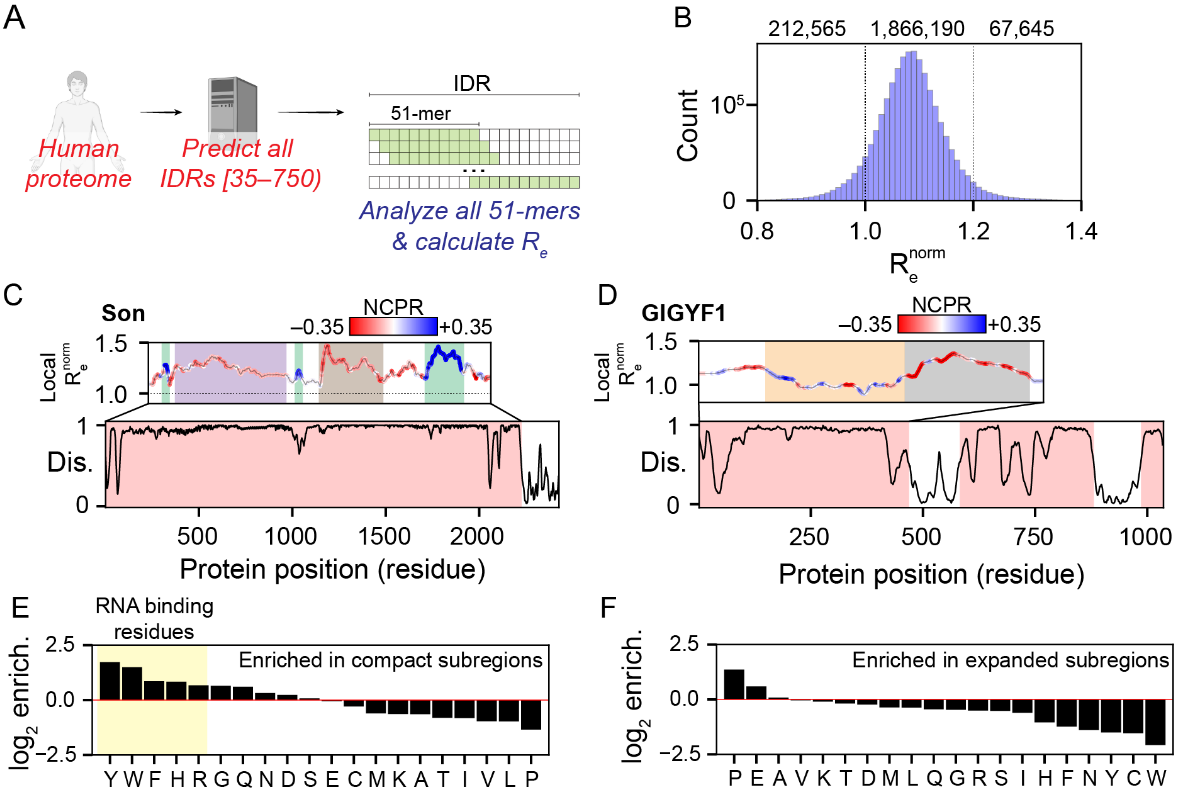

Figure 7. Local analysis of disordered protein subregions reveals sequence-dependent expansion and compaction. A) Graphical summary illustrating the sliding window subregion analysis presented in this figure. **B)** Distribution of the normalized end-to-end distance obtained from all 51-residue subfragments within IDRs in the human proteome. **C)** Linear analysis of local subregions in the 2227 residue IDR from the nuclear speckle protein Son, with conformationally-distinct subregions highlighted (UniProt: P18583). **D)** Linear analysis of local subregions in the 469-residue IDR from the cytosolic GIGYF1, with conformationally-distinct subregions highlighted (UniProt: O75420). **E)** Log_2-fold enrichment for amino acids found in compact subregions. Residues implicated in RNA binding are highlighted. **D)** Log_2-fold enrichment for amino acids found in expanded subregions.

DISCUSSION

Intrinsically disordered proteins and protein regions (IDRs) are ubiquitous, yet the absence of a fixed 3D structure coupled with limited sequence conservation has challenged conventional routes for mapping between protein sequence and molecular function. Given IDR function can be influenced or even dictated by the sequence-encoded conformational biases, a robust understanding of sequence-ensemble relationships remains an important feature for interpreting how IDRs conduct their cellular roles [19,94,95].

Here, we present ALBATROSS, a deep learning approach trained on coarse-grained simulations that allow for direct prediction of ensemble-average global dimensions from protein sequences. While there are several caveats that should be considered (discussed below), ALBATROSS enables us to assess sequence-to-ensemble relationships for both synthetic and natural IDRs. By providing ALBATROSS as both a locally-installable Python package and an easy-to-use Google Colab notebook, we aim to lower the barrier for sequence-to-ensemble predictions for single IDRs or for entire proteomes.

Our proteome-wide analysis suggests that IDR expansion can be driven by net charge, proline residues, or a combination of the two (**Fig. 6I**). In contrast, the subset of amino acids (Y/W/F/H/R/G/Q) enriched in compact IDR subregions overlap strongly with those residues previously reported to engage in RNA binding (**Fig. 7E**). Previous work has shown that disordered regions can chaperone RNA, both in isolation and in the context of biomolecular condensates[66,96–99]. Interestingly, these same RNA binding residues are also over-represented in IDR subregions that can drive phase separation *in vitro* and form condensates *in vivo* [23,72,74,100–104]. One interpretation of these observations is that compact IDRs have evolved to self-assemble and recruit RNA into condensates. Another interpretation is that these RNA-binding IDRs are constitutively bound to RNA in cells where they exchange compaction-driving intramolecular protein:protein interactions for expansion-driving intermolecular protein:RNA interactions. Under this interpretation, compact IDRs are only compact in an unphysiological RNA-free context, such that they expand to envelop and chaperone RNA molecules while themselves being reciprocally chaperoned by RNA. These interpretations are not mutually exclusive, nor do they prohibit a model in which RNA chaperoning requires many copies of RNA-binding proteins forming dynamic condensates.

Recent work from several groups touches on ideas or results that dovetail well with our own. As a proof-of-principle, Janson et al. trained a generative adversarial network using a transformer architecture with self-attention (idpGAN) to predict ensemble properties for coarse-grained simulations[105]. This study demonstrates the potential for multi-resolution models that interpolate between coarse-grained and atomistic simulations, with implications for enhanced sampling, whereby latent configurational space could be explored via a learned network to seed distinct but thermodynamically reasonable starting configurations. In parallel, Chao *et al.* presented a novel approach to represent IDR ensembles and train several different machine learning architectures to predict global dimensions from sequence[106]. This work suggests alternative representation schemes may be useful to capture sequence-specific effects and that representing IDRs in a length-free way using the Bag of Amino Acids (BAA) representation offers some advantages in terms of input for model training. Finally, Tesei & Trolle et al. recently performed an analogous proteome-wide assessment of the IDR-ome using the CALVADOS2 force field[48,53,107]. Despite utilizing an entirely distinct coarse-grained simulation forcefield, this work reached similar conclusions to us with respect to the human proteome-wide analysis. Indeed, the correlation between CALVADOS2 simulations of the human proteome and ALBATROSS predictions is high, with root mean squared errors within the range of experimental error (**Fig. S11**, R^2 = 0.98, RMSE = 4.17 Å, n=29,998, prediction time ~200 seconds on a CPU).

Moreover, we arrive at similar conclusions for the propensity for relatively expanded IDRs, the importance of net charge, charge patterning, and aromatic residues in tuning overall dimensions, and the association between RNA binding proteins and compact IDRs. Overall, the distribution of IDR dimensions from CALVADOS2 is slightly more compact than from Mpipi-GG, a difference we suspect reflects an underestimation of aliphatic residue interactions in the Mpipi-GG force field. Nevertheless, the general trends between the two studies show good agreement, a compelling result given the differences in approaches, force fields, and assumptions.

While our benchmarks demonstrate the predictive power of simulations performed in the Mpipi-GG force field and subsequently our ALBATROSS networks, there are a few important limitations to consider. Mpipi-GG is a one-bead-per-residue, coarse-grained force field. With this in mind, all of the caveats associated with one-bead-per-residue coarse-grained simulations should be considered when interpreting ALBATROSS predictions or Mpipi-GG simulations. These include the inability to acquire secondary or tertiary structure, the assumption of an isotropic interaction potential, and the steric approximation of non-spherical amino acids as spherical beads. Despite these simplifying assumptions, both this work and recent work have demonstrated that coarse-grained simulations can achieve good accuracy, at least in terms of global ensemble properties[49,53]. Furthermore, coarse-grained simulations have computational advantages over all-atom simulations. Namely, performing tens of thousands of coarse-grained simulations for analysis or for generating deep learning datasets is feasible, whereas performing tens of thousands of all-atom simulations is out of reach for nearly all academic groups[107]. Nevertheless, we suggest a few specific caveats that should be considered when evaluating Mpipi-GG simulations or ALBATROSS predictions. Firstly, we likely underestimate the impact of solvation effects on charged amino acids, such that highly charged net-neutral IDRs are likely less expanded than they should be. Secondly, our coarse-grained model and predictors do not account for transient secondary structure elements, a pervasive source of local conformational heterogeneity in many IDRs. Finally, we likely underestimate the hydrophobic effect for aliphatic residues, an intrinsically challenging phenomenon to capture in coarse-grained force fields for IDR simulations. These two final points mean we likely overestimate the predicted dimensions of IDRs that possess hydrophobicity-driven secondary structure, a caveat that should be carefully considered for IDRs enriched for helicity-promoting and/or aliphatic residues.

Beyond the limitations associated with coarse-grained simulations, a second batch of limitations stems from the training dataset. We opted to train on a large set of synthetic sequences that titrate across IDR sequence features known to impact ensemble dimensions (composition, charge, charge patterning, *etc.*). Although our predictive power for unseen synthetic sequences is excellent, for some of the networks (e.g., the scaling exponent network), our ability to accurately predict properties for biological sequences is limited. Encouragingly, for the radius of gyration and end-to-end distance predictors, the accuracy with biological sequences after training on synthetic sequences is extremely high, suggesting we are already in a relatively robust regime. Nevertheless, further work is being carried out to refine and improve the available training data and better capture sequence features that were missing in our original training set.

Our decision to focus on an LSTM-BRNN architecture for training was motivated by the desire to develop trained networks that were performant (10-50 sequences/second) on CPU commodity hardware. While more complex architectures (e.g., transformer-based networks) may offer more accurate predictors, we see two central limitations here. First, transformer-based architectures are memory intensive, and although some low-memory transformer-based architectures exist, most pre-trained biological transformers have memory requirements that scale quadratically with sequence length[108–113]. These large memory requirements can be prohibitive on commodity hardware, and we wanted to focus on developing and distributing portable tools for the community. Second, our LSTM-based architecture generates predictions that are already quite accurate. The error associated with our predictions is on the order of the experimental error (0 - 4 Å), so treating model architecture as a tunable hyperparameter for the performance of these prediction tasks, while an interesting question, did not merit further experimentation.

CONCLUSION

In this work, we present ALBATROSS, an accessible and accurate route to predict IDR global dimensions from sequence. All of the data associated with the proteome-wide analysis presented in **Fig. 6** and **Fig. 7** are shared as SHEPHARD-compliant datafiles, and we encourage other groups to explore these predictions in the context of other protein annotations using SHEPHARD and the set of precomputed annotations provided therein[62]. Our results are in good agreement with prior experimental and recent analogous computational work, suggesting that ALBATROSS offers a convenient route to obtain biophysical insight into IDR sequence-ensemble relationships.

ACKNOWLEDGEMENTS

We thank members of the Holehouse lab for extensive discussions on many aspects of this work. We thank Shubh Minhas for the ALBATROSS logo. We are also indebted to Kresten Lindorff-Larsen for his willingness to delay preprinting of their manuscript such that both preprints could reference one another. We also thank past and present members of the Pappu lab, Sukenik lab, Soranno lab, and Mittag lab for many illuminating discussions on sequence-ensemble relationships over the years. Funding for this work was provided by the Human Frontiers in Science Program (HFSP RGP0015/2022) to A.S.H. by the Longer Life Foundation, an RGA/Washington University in St. Louis Collaboration to A.S.H., and by the National Science Foundation with award 2128068 to A.S.H. JML was supported by the National Science Foundation via grant number DGE-2139839. DG was supported by the NSF via grant number DGE-2139839. We also thank members of the Water and Life Interface Institute (WALII), supported by NSF DBI grant #2213983, for helpful discussions.

LITERATURE CITED

1. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E. & Babu, M. M. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114,** 6589–6631 (2014).
2. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293,** 321–331 (1999).
3. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. R., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C. H., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, M., Garner, E. C. & Obradovic, Z. Intrinsically disordered protein. *J. Mol. Graph. Model.* **19,** 26–59 (2001).
4. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27,** 527–533 (2002).
5. Mittag, T. & Forman-Kay, J. D. Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* **17,** 3–14 (2007).
6. Dyson, H. J. & Wright, P. E. Equilibrium NMR studies of unfolded and partially folded proteins. *Nat. Struct. Biol.* **5 Suppl,** 499–503 (1998).
7. Tran, H. T., Wang, X. & Pappu, R. V. Reconciling observations of sequence-specific conformational propensities with the generic polymeric behavior of denatured proteins. *Biochemistry* **44,** 11369–11380 (2005).
8. Pappu, R. V., Wang, X., Vitalis, A. & Crick, S. L. A polymer physics perspective on driving forces and mechanisms for protein aggregation - Highlight Issue: Protein Folding. *Arch. Biochem. Biophys.* **469,** 132–141 (2008).
9. Brangwynne, C. P., Tompa, P. & Pappu, R. V. Polymer physics of intracellular phase transitions. *Nat. Phys.* **11,** 899–904 (2015).

10. Borg, M., Mittag, T., Pawson, T., Tyers, M., Forman-Kay, J. D. & Chan, H. S. Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 9650–9655 (2007).

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

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

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

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

15. Cubuk, J. & Soranno, A. Macromolecular crowding and intrinsically disordered proteins: A polymer physics perspective. *ChemSystemsChem* (2022). doi:10.1002/syst.202100051

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

17. Vancraenenbroeck, R., Harel, Y. S., Zheng, W. & Hofmann, H. Polymer effects modulate binding affinities in disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 19506–19512 (2019).

18. Dzuricky, M., Roberts, S. & Chilkoti, A. Convergence of Artificial Protein Polymers and Intrinsically Disordered Proteins. *Biochemistry* **57,** 2405–2414 (2018).

19. Das, R. K., Ruff, K. M. & Pappu, R. V. Relating sequence encoded information to form and

function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **32,** 102–112 (2015).

20. Crick, S. L., Jayaraman, M., Frieden, C., Wetzel, R. & Pappu, R. V. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 16764–16769 (2006).
21. Müller-Späth, S., Soranno, A., Hirschfeld, V., Hofmann, H., Rüegger, S., Reymond, L., Nettels, D. & Schuler, B. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 14609–14614 (2010).
22. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. & Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 8183–8188 (2010).
23. Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V. & Mittag, T. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367,** 694–699 (2020).
24. Marsh, J. A. & Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophys. J.* **98,** 2383–2390 (2010).
25. Martin, E. W., Holehouse, A. S., Grace, C. R., Hughes, A., Pappu, R. V. & Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **138,** 15323–15335 (2016).
26. Bowman, M. A., Riback, J. A., Rodriguez, A., Guo, H., Li, J., Sosnick, T. R. & Clark, P. L. Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 23356–23364 (2020).
27. Song, J., Li, J. & Chan, H. S. Small-Angle X-ray Scattering Signatures of Conformational Heterogeneity and Homogeneity of Disordered Protein Ensembles. *J. Phys. Chem. B* **125,** 6451–6478 (2021).
28. Gomes, G.-N. W., Krzeminski, M., Namini, A., Martin, E. W., Mittag, T., Head-Gordon, T.,

Forman-Kay, J. D. & Gradinaru, C. C. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. *J. Am. Chem. Soc.* **142,** 15697–15710 (2020).

29. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **115,** E4758–E4766 (2018).

30. Lazar, T., Martínez-Pérez, E., Quaglia, F., Hatos, A., Chemes, L. B., Iserte, J. A., Méndez, N. A., Garrone, N. A., Saldaño, T. E., Marchetti, J., Rueda, A. J. V., Bernadó, P., Blackledge, M., Cordeiro, T. N., Fagerberg, E., Forman-Kay, J. D., Fornasari, M. S., Gibson, T. J., Gomes, G.-N. W., Gradinaru, C. C., Head-Gordon, T., Jensen, M. R., Lemke, E. A., Longhi, S., Marino-Buslje, C., Minervini, G., Mittag, T., Monzon, A. M., Pappu, R. V., Parisi, G., Ricard-Blum, S., Ruff, K. M., Salladini, E., Skepö, M., Svergun, D., Vallet, S. D., Varadi, M., Tompa, P., Tosatto, S. C. E. & Piovesan, D. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.* **49,** D404–D411 (2021).

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

32. Gibbs, E. B., Lu, F., Portz, B., Fisher, M. J., Medellin, B. P., Laremore, T. N., Zhang, Y. J., Gilmour, D. S. & Showalter, S. A. Phosphorylation induces sequence-specific conformational switches in the RNA polymerase II C-terminal domain. *Nat. Commun.* **8,** 15233 (2017).

33. Gibbs, E. B. & Showalter, S. A. Quantification of Compactness and Local Order in the Ensemble of the Intrinsically Disordered Protein FCP1. *J. Phys. Chem. B* **120,** 8960–8969 (2016).

34. Moses, D., Guadalupe, K., Yu, F., Flores, E., Perez, A., McAnelly, R., Shamoon, N. M.,

Cuevas-Zepeda, E., Merg, A. D., Martin, E. W., Holehouse, A. S. & Sukenik, S. Structural biases in disordered proteins are prevalent in the cell. *bioRxiv* 2021.11.24.469609 (2022). doi:10.1101/2021.11.24.469609

35. Moses, D., Yu, F., Ginell, G. M., Shamoon, N. M., Koenig, P. S., Holehouse, A. S. & Sukenik, S. Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment. *J. Phys. Chem. Lett.* **11,** 10131–10136 (2020).
36. Daughdrill, G. W. Disorder for Dummies: Functional Mutagenesis of Transient Helical Segments in Disordered Proteins. *Methods Mol. Biol.* **2141,** 3–20 (2020).
37. Thomasen, F. E. & Lindorff-Larsen, K. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. *Biochem. Soc. Trans.* **50,** 541–554 (2022).
38. Dignon, G. L., Zheng, W., Best, R. B., Kim, Y. C. & Mittal, J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 9929–9934 (2018).
39. Lin, Y.-H. & Chan, H. S. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophys. J.* **112,** 2043–2046 (2017).
40. Martin, E. W., Hopkins, J. B. & Mittag, T. Small-angle X-ray scattering experiments of monodisperse intrinsically disordered protein samples close to the solubility limit. *Methods Enzymol.* **646,** 185–222 (2021).
41. Gibbs, E. B., Cook, E. C. & Showalter, S. A. Application of NMR to studies of intrinsically disordered proteins. *Arch. Biochem. Biophys.* **628,** 57–70 (2017).
42. Bernadó, P. & Svergun, D. I. Analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Methods Mol. Biol.* **896,** 107–122 (2012).
43. Borgia, A., Borgia, M. B., Bugge, K., Kissling, V. M., Heidarsson, P. O., Fernandes, C. B., Sottini, A., Soranno, A., Buholzer, K. J., Nettels, D., Kragelund, B. B., Best, R. B. & Schuler, B. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555,** 61–66 (2018).
44. Kassem, N., Araya-Secchi, R., Bugge, K., Barclay, A., Steinocher, H., Khondker, A., Wang,

Y., Lenard, A. J., Bürck, J., Sahin, C., Ulrich, A. S., Landreh, M., Pedersen, M. C., Rheinstädter, M. C., Pedersen, P. A., Lindorff-Larsen, K., Arleth, L. & Kragelund, B. B. Order and disorder-An integrative structure of the full-length human growth hormone receptor. *Sci Adv* **7,** eabh3805 (2021).

45. Cragnell, C., Durand, D., Cabane, B. & Skepö, M. Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS. *Proteins: Struct. Funct. Bioinf.* **84,** 777–791 (2016).

46. Borgia, A., Zheng, W., Buholzer, K., Borgia, M. B., Schüler, A., Hofmann, H., Soranno, A., Nettels, D., Gast, K., Grishaev, A., Best, R. B. & Schuler, B. Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J. Am. Chem. Soc.* **138,** 11714–11726 (2016).

47. Zheng, W., Borgia, A., Buholzer, K., Grishaev, A., Schuler, B. & Best, R. B. Probing the action of chemical denaturant on an intrinsically disordered protein by simulation and experiment. *J. Am. Chem. Soc.* **138,** 11702–11713 (2016).

48. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. U. S. A.* **118,** (2021).

49. Joseph, J. A., Reinhardt, A., Aguirre, A., Chew, P. Y., Russell, K. O., Espinosa, J. R., Garaizar, A. & Collepardo-Guevara, R. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat Comput Sci* **1,** 732–743 (2021).

50. Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B. & Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **14,** e1005941 (2018).

51. Regy, R. M., Thompson, J., Kim, Y. C. & Mittal, J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* **30,**

1371–1379 (2021).

52. Wu, H., Wolynes, P. G. & Papoian, G. A. AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. *J. Phys. Chem. B* **122,** 11115–11125 (2018).
53. Tesei, G. & Lindorff-Larsen, K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res. Eur.* **2,** 94 (2023).
54. Emenecker, R. & Holehouse, A. *GOOSE - a tool for the rational design of intrinsically disordered regions*. (2022). doi:10.5281/zenodo.6878703
55. Lalmansingh, J. M., Keeley, A. T., Ruff, K. M., Pappu, R. V. & Holehouse, A. S. SOURSOP: A Python package for the analysis of simulations of intrinsically disordered proteins. *bioRxiv* (2023). doi:10.1101/2023.02.16.528879
56. Griffith, D. & Holehouse, A. S. PARROT is a flexible recurrent neural network framework for analysis of large protein datasets. *Elife* **10,** (2021).
57. Holehouse, A. S. *sparrow: a tool for integrative analysis and prediction from protein sequence data*. (2022). doi:10.5281/zenodo.6891920
58. Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in 't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C. & Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271,** 108171 (2022).
59. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J. & Pande, V. S. MDTraj: a modern, open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109,** 1528–1532 (2015).
60. Rubinstein, M. & Colby, R. H. *Polymer Physics*. (Oxford University Press, 2003).
61. Holehouse, A. S., Garai, K., Lyle, N., Vitalis, A. & Pappu, R. V. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain

expansion via chemical denaturation. *J. Am. Chem. Soc.* **137,** 2984–2995 (2015).

62. Ginell, G. M., Flynn, A. J. & Holehouse, A. S. SHEPHARD: a modular and extensible software architecture for analyzing and annotating large protein datasets. *bioRxiv* 2022.09.18.508433 (2022). doi:10.1101/2022.09.18.508433

63. Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* **120,** 4312–4319 (2021).

64. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47,** D506–D515 (2019).

65. Alston, J. J., Ginell, G. M., Soranno, A. & Holehouse, A. S. The analytical Flory random coil is a simple-to-use reference model for unfolded and disordered proteins. *bioRxiv* 2023.03.12.531990 (2023).

66. Cubuk, J., Alston, J. J., Jeremías Incicco, J., Holehouse, A. S., Hall, K. B., Stuchell-Brereton, M. D. & Soranno, A. The disordered N-terminal tail of SARS CoV-2 Nucleocapsid protein forms a dynamic complex with RNA. *bioRxiv* 2023.02.10.527914 (2023). doi:10.1101/2023.02.10.527914

67. Sanchez-Burgos, I., Espinosa, J. R., Joseph, J. A. & Collepardo-Guevara, R. RNA length has a non-trivial effect in the stability of biomolecular condensates formed by RNA-binding proteins. *PLoS Comput. Biol.* **18,** e1009810 (2022).

68. Zheng, W., Dignon, G. L., Brown, M., Kim, Y. C. & Mittal, J. Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *J. Phys. Chem. Lett.* (2020). doi:10.1021/acs.jpclett.0c00288

69. Peran, I., Holehouse, A. S., Carrico, I. S., Pappu, R. V., Bilsel, O. & Raleigh, D. P. Unfolded states under folding conditions accommodate sequence-specific conformational preferences with random coil-like dimensions. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 12301–12310 (2019).

70. Mao, A. H., Lyle, N. & Pappu, R. V. Describing sequence–ensemble relationships for intrinsically disordered proteins. *Biochem. J* **449,** 307–318 (2013).
71. Sherry, K. P., Das, R. K., Pappu, R. V. & Barrick, D. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E9243–E9252 (2017).
72. Wang, J., Choi, J.-M., Holehouse, A. S., Lee, H. O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., Poser, I., Pappu, R. V., Alberti, S. & Hyman, A. A. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **174,** 688–699.e16 (2018).
73. Alshareedah, I., Borcherds, W. M., Cohen, S. R., Farag, M., Singh, A., Bremer, A., Pappu, R. V., Mittag, T. & Banerjee, P. R. Sequence-encoded grammars determine material properties and physical aging of protein condensates. *bioRxiv* (2023). doi:10.1101/2023.04.06.535902
74. Dzuricky, M., Rogers, B. A., Shahid, A., Cremer, P. S. & Chilkoti, A. De novo engineering of intracellular condensates using artificial disordered proteins. *Nat. Chem.* **12,** 814–825 (2020).
75. Riback, J. A., Katanski, C. D., Kear-Scott, J. L., Pilipenko, E. V., Rojek, A. E., Sosnick, T. R. & Drummond, D. A. Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168,** 1028–1040.e19 (2017).
76. Sørensen, C. S. & Kjaergaard, M. Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 23124–23131 (2019).
77. Riback, J. A., Bowman, M. A., Zmyslowski, A. M., Knoverek, C. R., Jumper, J. M., Hinshaw, J. R., Kaye, E. B., Freed, K. F., Clark, P. L. & Sosnick, T. R. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358,** 238–241 (2017).

78. Mukhopadhyay, S., Krishnan, R., Lemke, E. A., Lindquist, S. & Deniz, A. A. A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures. *Proceedings of the National Academy of Sciences* **104,** 2649–2654 (2007).

79. Lu, X. & Murphy, R. M. Asparagine Repeat Peptides: Aggregation Kinetics and Comparison with Glutamine Repeats. *Biochemistry* **54,** 4784–4794 (2015).

80. Boze, H., Marlin, T., Durand, D., Pérez, J., Vernhet, A., Canon, F., Sarni-Manchado, P., Cheynier, V. & Cabane, B. Proline-rich salivary proteins have extended conformations. *Biophys. J.* **99,** 656–665 (2010).

81. Ginell, G. M. & Holehouse, A. S. An Introduction to the Stickers-and-Spacers Framework as Applied to Biomolecular Condensates. *Methods Mol. Biol.* **2563,** 95–116 (2023).

82. Choi, J.-M., Holehouse, A. S. & Pappu, R. V. Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. *Annu. Rev. Biophys.* **49,** 107–133 (2020).

83. Harmon, T. S., Holehouse, A. S., Rosen, M. K. & Pappu, R. V. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *Elife* **6,** (2017).

84. Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V. & Mittag, T. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14,** 196–207 (2022).

85. Choi, J.-M., Hyman, A. A. & Pappu, R. V. Generalized models for bond percolation transitions of associative polymers. *Phys Rev E* **102,** 042403 (2020).

86. Fei, J., Jadaliha, M., Harmon, T. S., Li, I. T. S., Hua, B., Hao, Q., Holehouse, A. S., Reyer, M., Sun, Q., Freier, S. M., Pappu, R. V., Prasanth, K. V. & Ha, T. Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *J. Cell Sci.* **130,** 4180–4192 (2017).

87. Zhao, Y., Stankovic, S., Koprulu, M., Wheeler, E., Day, F. R., Lango Allen, H., Kerrison, N. D., Pietzner, M., Loh, P.-R., Wareham, N. J., Langenberg, C., Ong, K. K. & Perry, J. R. B.

GIGYF1 loss of function is associated with clonal mosaicism and adverse metabolic health. *Nat. Commun.* **12,** 4178 (2021).

88. Deaton, A. M., Parker, M. M., Ward, L. D., Flynn-Carroll, A. O., BonDurant, L., Hinkle, G., Akbari, P., Lotta, L. A., Regeneron Genetics Center, DiscovEHR Collaboration, Baras, A. & Nioi, P. Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *Sci. Rep.* **11,** 21565 (2021).
89. Chen, G., Yu, B., Tan, S., Tan, J., Jia, X., Zhang, Q., Zhang, X., Jiang, Q., Hua, Y., Han, Y., Luo, S., Hoekzema, K., Bernier, R. A., Earl, R. K., Kurtz-Nelson, E. C., Idleburg, M. J., Madan-Khetarpal, S., Clark, R., Sebastian, J., Fernandez-Jaen, A., Alvarez, S., King, S. D., Ramos, L. L., Santos, M. L. S., Martin, D. M., Brooks, D., Symonds, J. D., Cutcutache, I., Pan, Q., Hu, Z., Yuan, L., Eichler, E. E., Xia, K. & Guo, H. GIGYF1 disruption associates with autism and impaired IGF-1R signaling. *J. Clin. Invest.* **132,** (2022).
90. Calnan, B. J., Tidor, B., Biancalana, S., Hudson, D. & Frankel, A. D. Arginine-mediated RNA recognition: the arginine fork. *Science* **252,** 1167–1171 (1991).
91. Cléry, A., Blatter, M. & Allain, F. H.-T. RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.* **18,** 290–298 (2008).
92. Hall, K. B. RNA–protein interactions. *Curr. Opin. Struct. Biol.* **12,** 283–288 (2002).
93. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Mol. Cell* **78,** 9–29 (2020).
94. González-Foutel, N. S., Glavina, J., Borcherds, W. M., Safranchik, M., Barrera-Vilarmau, S., Sagar, A., Estaña, A., Barozet, A., Garrone, N. A., Fernandez-Ballester, G., Blanes-Mira, C., Sánchez, I. E., de Prat-Gay, G., Cortés, J., Bernadó, P., Pappu, R. V., Holehouse, A. S., Daughdrill, G. W. & Chemes, L. B. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat. Struct. Mol. Biol.* **29,** 781–790 (2022).
95. Zeno, W. F., Baul, U., Snead, W. T., DeGroot, A. C. M., Wang, L., Lafer, E. M., Thirumalai, D. & Stachowiak, J. C. Synergy between intrinsically disordered domains and structured

proteins amplifies membrane curvature sensing. *Nat. Commun.* **9,** 4152 (2018).

96. Holmstrom, E. D., Liu, Z., Nettels, D., Best, R. B. & Schuler, B. Disordered RNA chaperones can enhance nucleic acid folding via local charge screening. *Nat. Commun.* **10,** 2453 (2019).

97. Nott, T. J., Craggs, T. D. & Baldwin, A. J. Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters. *Nat. Chem.* **8,** 569–575 (2016).

98. Sarni, S. H., Roca, J., Du, C., Jia, M., Li, H., Damjanovic, A., Małecka, E. M., Wysocki, V. H. & Woodson, S. A. Intrinsically disordered interaction network in an RNA chaperone revealed by native mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **119,** e2208780119 (2022).

99. Zúñiga, S., Sola, I., Moreno, J. L., Sabella, P., Plana-Durán, J. & Enjuanes, L. Coronavirus nucleocapsid protein is an RNA chaperone. *Virology* **357,** 215–227 (2007).

100. Martin, E. W. & Holehouse, A. S. Intrinsically disordered protein regions and phase separation: sequence determinants of assembly or lack thereof. *Emerg Top Life Sci* **4,** 307–329 (2020).

101. Nott, T. J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T. D., Bazett-Jones, D. P., Pawson, T., Forman-Kay, J. D. & Baldwin, A. J. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57,** 936–947 (2015).

102. Han, T. W., Kato, M., Xie, S., Wu, L. C., Mirzaei, H., Pei, J., Chen, M., Xie, Y., Allen, J., Xiao, G. & McKnight, S. L. Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell* **149,** 768–779 (2012).

103. Kato, M., Han, T. W., Xie, S., Shi, K., Du, X., Wu, L. C., Mirzaei, H., Goldsmith, E. J., Longgood, J., Pei, J., Grishin, N. V., Frantz, D. E., Schneider, J. W., Chen, S., Li, L., Sawaya, M. R., Eisenberg, D., Tycko, R. & McKnight, S. L. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149,**

753–767 (2012).

104. Lin, Y., Currie, S. L. & Rosen, M. K. Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. *J. Biol. Chem.* **292,** 19110–19120 (2017).

105. Janson, G., Valdes-Garcia, G., Heo, L. & Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **14,** 774 (2023).

106. Chao, T.-H., Rekhi, S., Mittal, J. & Tabor, D. P. Data-Driven Models for Predicting Intrinsically Disordered Protein Polymer Physics Directly from Composition or Sequence. *ChemRxiv* (2023). doi:10.26434/chemrxiv-2023-wrnq1

107. Tesei, G., Trolle, A. I., Jonsson, N., Betz, J., Pesce, F., Johansson, K. E. & Lindorff-Larsen, K. Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation. *bioRxiv* 2023.05.08.539815 (2023). doi:10.1101/2023.05.08.539815

108. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018). at <http://arxiv.org/abs/1810.04805>

109. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. & Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (2023).

110. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. & Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* 2022.07.20.500902 (2022). doi:10.1101/2022.07.20.500902

111. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J. & Fergus, R. Biological structure and function emerge from scaling unsupervised learning

to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118,** e2016239118 (2021).

112. Elnaggar, A., Essam, H., Salah-Eldin, W., Moustafa, W., Elkerdawy, M., Rochereau, C. & Rost, B. Ankh ☥ : Optimized Protein Language Model Unlocks General-Purpose Modelling. *bioRxiv* 2023–2001 (2023).

113. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: The Long-Document Transformer. *arXiv [cs.CL]* (2020). at <http://arxiv.org/abs/2004.05150>