Direct Prediction of Intrinsically Disordered Protein Conformational Properties From Sequence

Version 3 [2023-05-28]

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 regions (IDRs) are ubiquitous across all domains of life and play a range of functional roles. While folded domains are generally well-described by a single 3D structure, IDRs exist in a collection of interconverting states known as an ensemble. This structural heterogeneity means IDRs are largely absent from the PDB, contributing to a lack of computational approaches to predict ensemble conformational properties from sequence. Here we combine 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 enables the instantaneous prediction of ensemble average properties at proteome-wide scale. ALBATROSS is lightweight, easy-to-use, and accessible as both a locally installable software package and 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 ALBATROSS to characterize emergent biophysical behavior of IDRs within and between proteomes.

Update from previous version

- This preprint reports an updated version of the ALBATROSS network weights trained on simulations of over 42,000 sequences.
- In addition, we provide new colab notebooks that enable proteome-wide IDR prediction and annotation in minutes.
- All conclusions and observations made in versions 1 and 2 of this manuscript remain true and robust.

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, and apparent polymer scaling exponent) 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.

By combining ALBATROSS with recent improvements to our state-of-the art disorder predictor (metapredict V2-FF), we also provide the ability to predict and annotate the entire set of IDRs for a given proteome (the IDR-ome) in seconds-to-minutes. This advance opens the door for large-scale unstructural bioinformatics and, importantly, makes these predictions easy for anyone with an internet connection.

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.

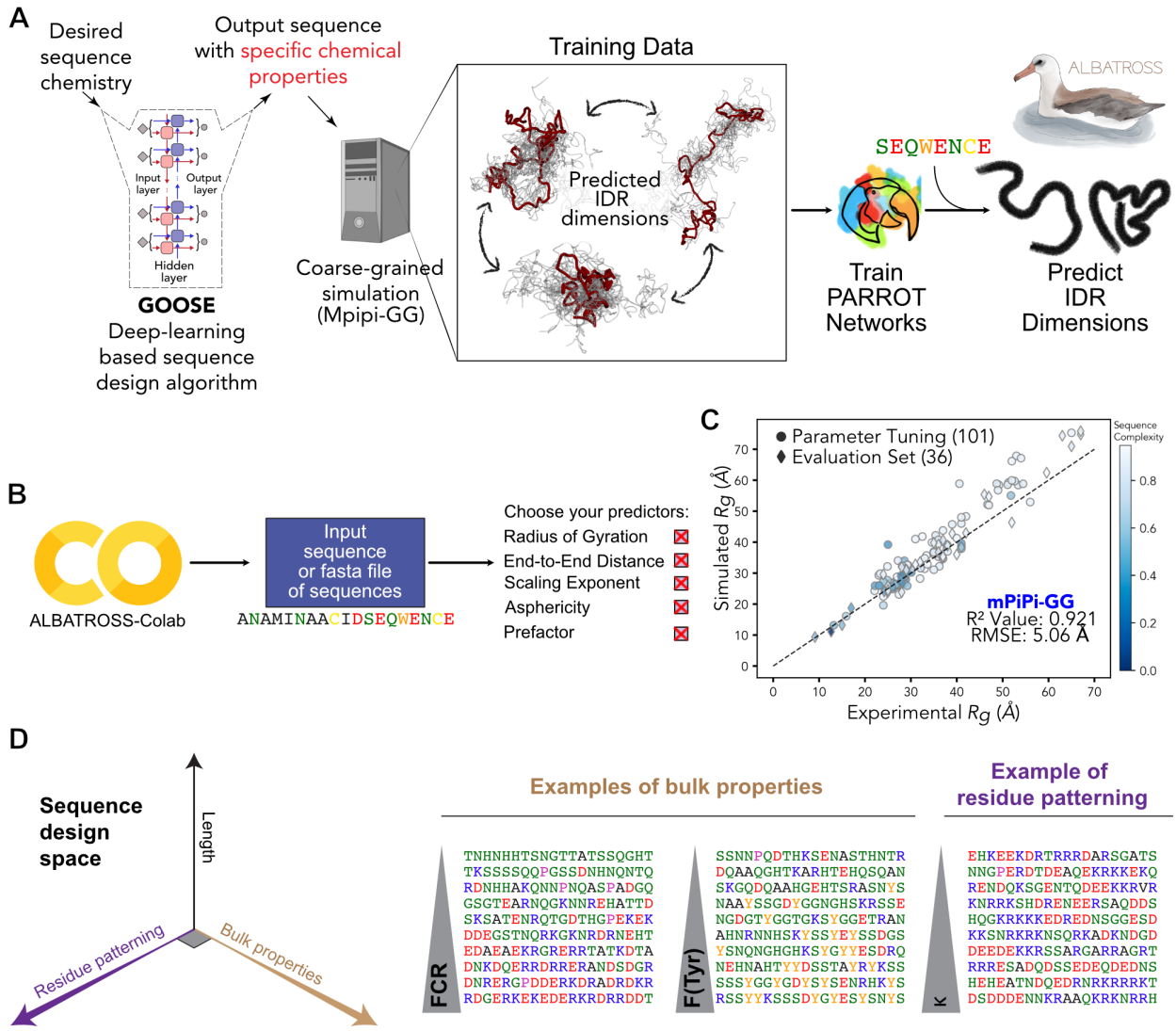

Figure 1. ALBATROSS is a deep-learning framework for predicting sequence-dependent IDR ensemble properties. A) Sequence design and simulation approach to generate training data for ALBATROSS networks. The Python package GOOSE is used to generate synthetic IDRs across a diverse area of sequence space. 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. **C)** Correlation and RMSE between the Mpipi-GG force field 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 blue color gradient signifies the Wootton-Federhen complexity of the sequence. D) Rational sequence design scheme. GOOSE was used to design sequences that titrated along different protein sequence parameter axes: length, residue patterning, and bulk amino acid properties. The fraction of charged residues and the fraction of tyrosine residues are examples of bulk properties. An example of residue patterning is the sequence parameter K which describes the asymmetry of positively and negatively charged amino acids in a sequence,

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].

ALBATROSS Training 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 23,127 disordered protein sequences across a diverse sequence space (see *Supplementary Information*). 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.

ALBATROSS Validation Sequence Library Design

To prepare a test set to accurately assess the true generalization error for each of our ALBATROSS predictors, we randomly selected an additional set of 2,501 biological IDRs from one of the aforementioned proteomes. To ensure that any newly selected biological sequences were distinct from biological sequences seen during training, we leveraged CD-HIT with default

parameters to remove sequences with >20% similarity[58]. In total, we were left with 2,306 biological IDRs in our test set. In addition, we also designed and simulated 3,731 synthetic disordered protein sequences. As before, these synthetic 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, the ALBATROSS test set we used to assess model accuracy consisted of 6,037 disordered protein sequences.

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[49,59]. 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 [55,60]. We made extensive use of GNU parallel for simulation analysis[61].

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 [62,63].

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 of these prediction tasks, we performed a hyperparameter grid search with 5-fold cross validation for 500 epochs with PARROT (64% training, 16% validation, and 20% test). The set of

hyperparameters that performed best on average across each fold were selected to train a final model with 80% of the data used as training data and 20% as validation data. Final network weights were chosen by selecting the epoch with the lowest validation loss across 750 epochs.

For each network, we chose a default learning rate of 0.001, and we performed a hyperparameters grid search over the following parameters: number of hidden layers (1 to 2), and a hidden dimension size (10 to 55). 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 curated test set of 6,051 IDR sequences which consisted of both synthetic and naturally occurring IDRs.

Disorder prediction

Disorder prediction (in this manuscript and in the associated notebooks) is provided through metapredict V2-FF [64,65]. Metapredict V2-FF is our newly implemented version of metapredict V2, which offers a 5-50x improvement in performance compared to metapredict V2 with no loss in prediction accuracy. V2-FF was developed specifically in the context of this manuscript for working with ALBATROSS, and enables proteome-wide prediction to be obtained in a reasonable (<10 min) timeframe.

Bioinformatics

Proteome-wide bioinformatics was performed using SPARROW (https://github.com/idptools/sparrow) and SHEPHARD[66]. 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. Proteomes were obtained from UniProt[65,67]. 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)[68].

Yeast homologous IDR analysis

Homologous yeast proteomes were obtained from the Yeast Genome Order Browser (YGOB)[69,70]. Syntenic genes were used to identify homologous proteins, and these sequences were then aligned using Clustal Omega[71]. For all homolog sets where there were more than 10 total proteins and a *S. cerevisiae* protein present, this protein was segmented into folded and disordered domains using metapredict (V2)[64,65]. This domain prediction was projected from the *S. cerevisiae* protein onto the multiple sequence alignment of all of the homologs, assigning any gapped regions between domains as IDRs. All sets of homologous IDRs where the *S. cerevisiae* IDR was 40 amino acids or longer were analyzed further. All IDRs belonging to one of these sets had predicted dimensions using ALBATROSS. Additionally, on each of these sets,

sequence similarity of the aligned IDRs was calculated using the pyMSA (v0.5.1) package (https://github.com/benhid/pyMSA), using the BLOSUM62 scoring matrix. Both the SumOfPairs and StarScore metrics for evaluating MSA similarity were normalized by the number of aligned sequences and length of the alignment. A similar procedure was applied to the E1A linker sequences from[72]. Sequence features for the Yak1 and Spt2 IDR homologs were computed using SPARROW.

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

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

Finally, we provide a standalone notebook for predicting and annotating all IDRs in a proteome (i.e., IDR-ome construction):
https://colab.research.google.com/github/holehouse-lab/ALBATROSS-colab/blob/main/idrome_constructor/idrome_constructor.ipynb
Specifically, IDR-ome construction combines predicting IDRs across an entire proteome with calculating IDR sequence properties and predicted IDR ensemble properties. With GPU-support on Google colab, this notebook enables the construction of the annotated human IDR-ome in ~5-8 minute. Without GPU support, the same output is achieved in ~7-10 minutes.

Data and code availability

All of the data associated with the proteome-wide analysis presented in **Fig. 4** and **Fig. 5** 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[66]. All other data and 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.

The Mpipi-GG force field accurately recapitulates IDR ensemble dimensions

Before training a deep learning model for ensemble properties, we collected training data via coarse-grained simulations performed in the Mpipi-GG force field, a fine-tuned version of the previously published Mpipi forcefield[49]. We (and others) have had great success in using Mpipi to provide molecular insight into a range of systems[73,74]. 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 (**Fig. S1-S5**). We made several small modifications to the underlying parameters, yielding a version of Mpipi we refer to as Mpipi-GG (see *Supplementary Information* for more details on force field fine-tuning).

To assess the accuracy of Mpipi-GG, we curated a set of 137 radii of gyration from previously published SAXS experiments on disordered proteins (**Fig. 1C**). 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. 1C, Fig. S1C,D**). Given this accuracy we reasoned that we could leverage Mpipi-GG simulations to generate training data for deep-learning-based models to map IDR sequence chemistry to ensemble properties.

Designing libraries of synthetic and biological IDRs for broad coverage of chemical space

Prior to performing simulations, we constructed a disordered protein sequence library with diverse sequence chemistries. We reasoned that a systematic exploration of IDR sequence space would enable our deep learning models to learn the complex underlying sequence-encoded conformational biases of disordered proteins. To build this initial synthetic library, we used GOOSE, our recently developed computational package for synthetic IDR design, to titrate across a range of sequence features known to impact IDR conformational behavior (**Fig 1D,** see *Methods*). Moreover, we opted to take advantage of GOOSE's ability to focus compositional exploration to sequences predicted to be disordered, such that our initial library is centered on sequences predicted with high confidence to be IDRs.

Using GOOSE, we designed a library of synthetic sequences that systematically explore IDR hydropathy, overall charge, net charge, charge patterning (quantified by kappa (κ)), and the overall fraction of different amino acids (**Fig. 1D**, **Fig S6**). Lastly, we also added disordered sequences by specifying random sequence fractions. Collectively, we designed a library of 23,127 synthetic sequences across a broad sequence landscape. In addition to these synthetic IDRs, we also randomly selected 19,075 IDRs from common model system proteomes. In total, we collected a library of 42,202 disordered protein sequences. This training library covered broad chemical space in terms of the fraction of aliphatic and polar residues (**Fig. S6**) as well as the fraction of positively charged residues and aromatic residues (**Fig. S6**). An overview of the amino acid compositions and overall sequence complexity for the training library is summarized in **Fig. S6**. 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. S7**)[11,75].

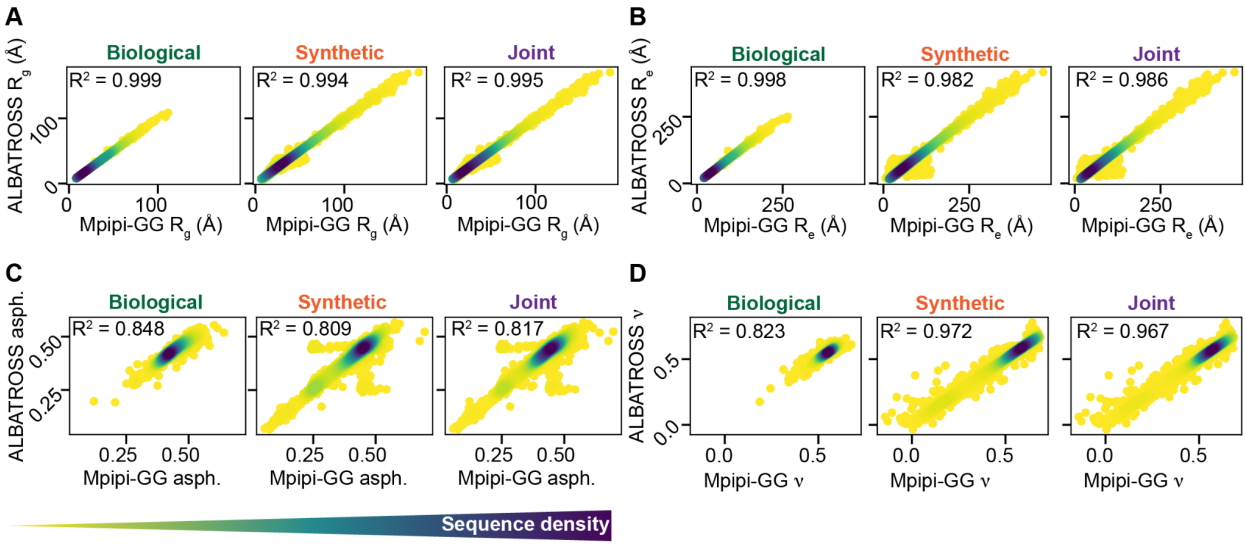

Figure 2. Assessing the ALBATROSS network accuracy on an independent test of both synthetic sequences and naturally occurring biological sequences. A, B, C, D) Accuracy of ALBATROSS in predicting R_g, R_e, Asphericity, and the polymer scaling exponent for previously unseen biological (left), synthetic (middle), and combined (right) 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.

Training and evaluating an IDR sequence-to-ensemble deep learning model

After designing our training library of IDR sequences and both selecting and tuning our force field, we performed molecular dynamics simulations of all 42,202 sequences and calculated ensemble-average parameters of interest. 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[23,76,77]. 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. After optimizing hyperparameters (see Methods), 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. 1C.** 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 = 0.998, **Fig. S8**). Inspired by this result, we next sought to assess how accurately ALBATROSS was able to predict the simulated Mpipi-GG R_g values on data unseen during training. Promisingly, when we evaluated our model on a held-out test set consisting of both synthetic and biological IDRs, we saw strong correlations in all cases (R^2 = 0.995, **Fig. 2A**). Finally, ALBATROSS was comparable or more accurate than current state-of-the-art methods for radii of gyration prediction but enables a much higher throughput (thousands of sequences/second) than comparable models (**Fig. S8**). 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 on the held out test set (R^2 = 0.986, **Fig. 2B**).

In addition to these R_g and R_e networks, we also trained networks for the mean asphericity, which displayed quantitative agreement on the test set (R^2 = 0.817, **Fig. 2C**). Lastly, 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 0.967 and 0.930, respectively, on the independent set of test sequences (**Fig. 2D, Fig. S9**). In summary, the ALBATROSS networks perform well on both synthetically designed and naturally occurring IDRs suggesting our networks have learned the role of sequence chemistry for tuning IDR ensemble dimensions (**Fig. 2, Fig. S9**).

ALBATROSS performance

While our four main networks are highly accurate, another 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. S10**, 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, asphericity, and polymer scaling exponent and prefactor.

Systematic assessment of sequence-to-ensemble properties

We next used ALBATROSS to assess how IDR sequence features influence global dimensions. Using GOOSE, we designed libraries of synthetic disordered sequences that systematically vary one sequence feature while holding others fixed. This strategy enables us to isolate and assess the average contribution of different sequence features; each data point on the panels in **Figure 3** reflects the average ensemble dimensions obtained from 100 distinct sequences with the same overall sequence features.

This analysis recapitulates and confirms a wide variety of sequence-to-ensemble relationships reported by many groups through computational and experimental studies over the last decade. In particular, our work highlights the importance of net charge in determining IDR global dimensions (**Fig. 3A, 3B**) and illustrates the fact that charge patterning becomes an increasingly important determinant of IDR dimensions as the overall fraction of charged residues increases (**Fig. 3C**)[11,12,21,22,24]. A systematic titration of individual amino acid fractions confirms that aromatic residues drive chain compaction (with tryptophan the strongest of the three), that proline residues drive chain expansion, and that glutamine (more than any other polar amino acid) drives intramolecular interactions and compaction (**Fig. 3D**) [20,23,35,78–80]. Finally, these analyses would suggest that aliphatic hydrophobes have a modest impact on IDR dimensions, a result consistent with prior work, although we caution that our predictions likely underestimate the hydrophobic effect (see Discussion) (**Fig. 3D**) [81–83]. In summary, our conclusions here are largely concordant with prior work but generalize those conclusions from individual proteins or systems to the sequence-average properties.

In addition to titrating the aromatic fraction, we designed synthetic repeat proteins consisting of glycine-serine-repeat “spacers” and poly-tyrosine “stickers” [84–86]. 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. 3E**). 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 experiments [23,87–89].

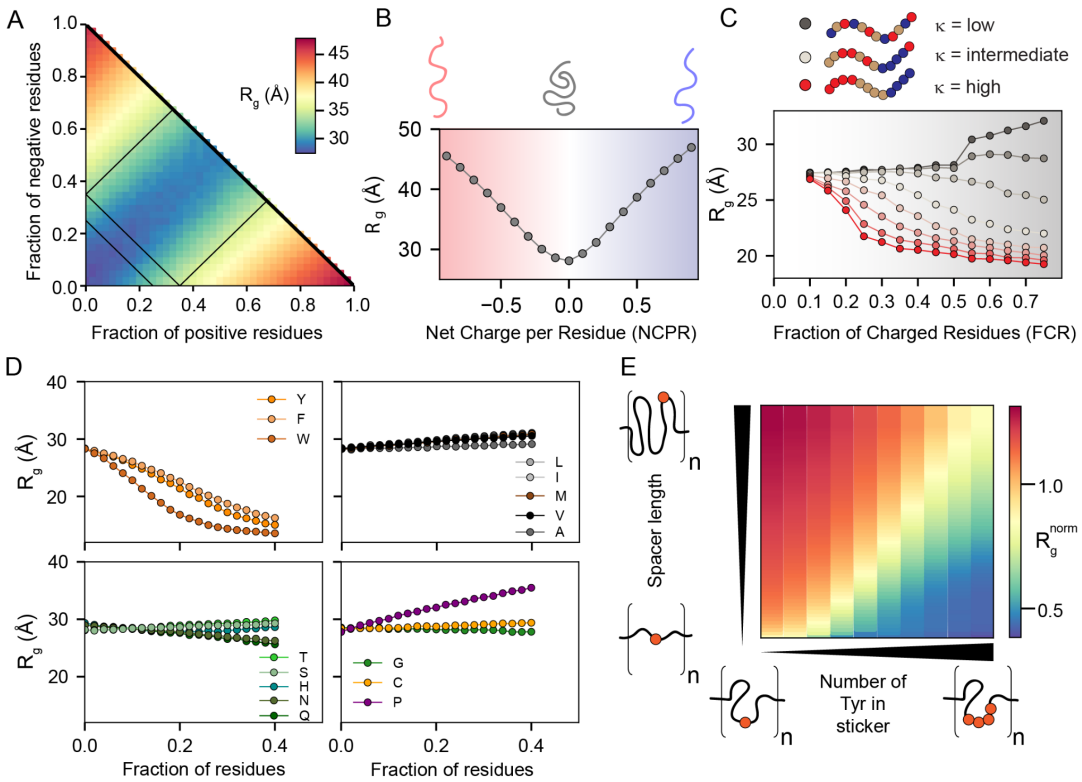

Figure 3. Sequence composition modulates the conformational preferences in disordered proteins. For panels A-D, each data point reports the average of many 100-residue synthetic disordered sequences with the specified composition. **A)** Diagram of states for weak to strong polyampholytes. Sequences 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)[68]. 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. 4A, B**). Focusing on IDRs between 35 and 750 residues in length, we calculated normalized radii of gyration (**Fig. 4C**), normalized end-to-end distance (**Fig. 4D**), and asphericity (**Fig. 4E**). 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[68]. 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. 4F**). 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. 4G**) or the absolute net charge and the fraction of proline residues (**Fig. 4H**) 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. 3**.

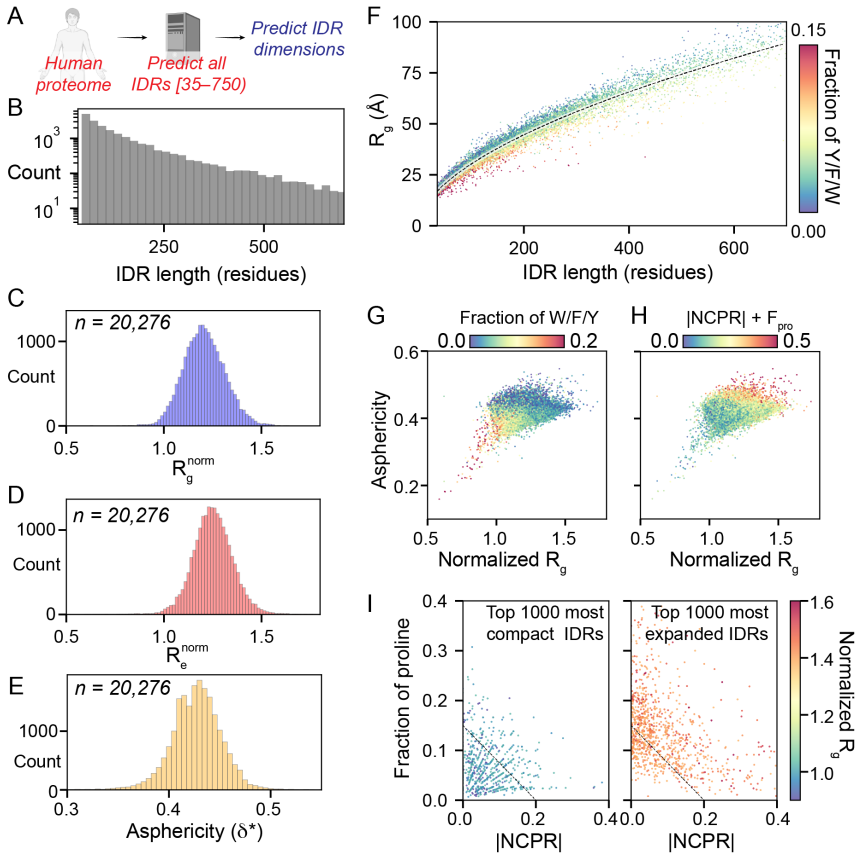

Figure 4. Human proteome-wide biophysical characterization of predicted IDRs.
A) ALBATROSS was used to perform sequence-dependent ensemble predictions for all IDRs in the human proteome. **B)** Histogram of all human IDRs ranging from 35 to 750 residues. **C)** Normalized mean ALBATROSS R_g distribution 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. 4** 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. 5A**). The distribution of normalized end-to-end distances is tighter than the corresponding distribution for full-length IDRs, with 195,555 (i.e., around 10%) of subregions behaving as polymers more compact than a corresponding Gaussian chain (normalized $R_g < 1$) (**Fig. 5B**).

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. 5C, 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. 5C**)[90]. Second, we analyzed the N-terminal IDR of GIGYF1, a highly disordered protein with a potential role in Type II diabetes [91–93]. 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, and potentially, molecular function.

Finally, we used the set of ~2 million IDR subregions to assess which residues were enriched in expanded or compact IDRs (**Fig. 5E, 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, several of the residues most strongly enriched for compact IDRs match those residues known to engage in RNA binding[73,94–97]. 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.

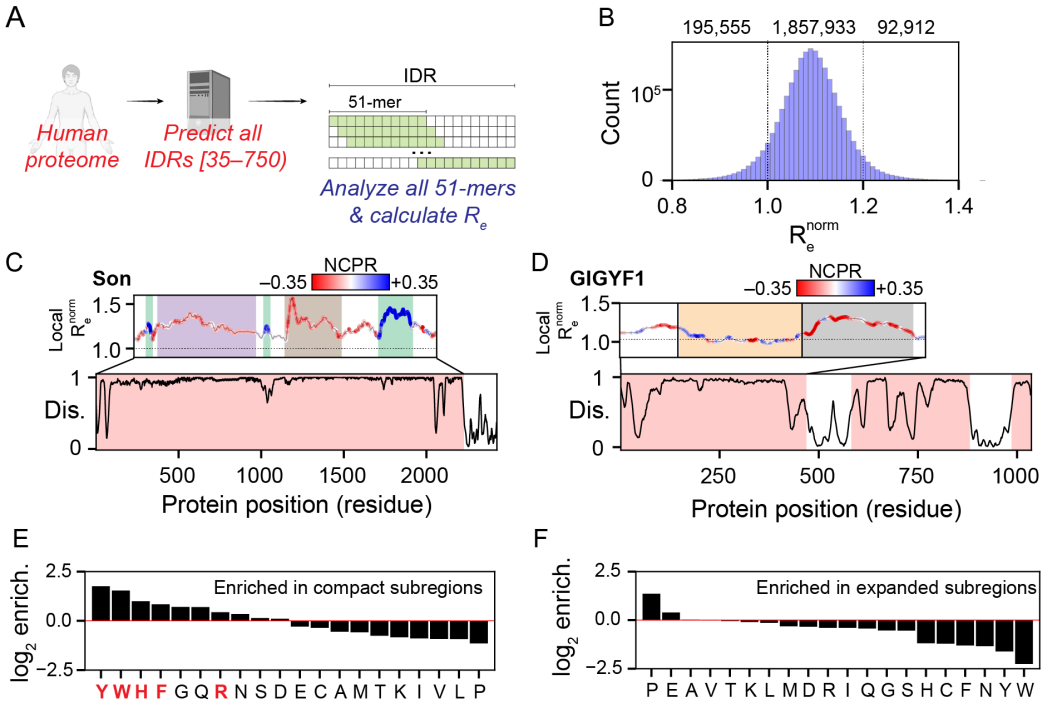

Figure 5. Local analysis of disordered protein subregions reveals sequence-dependent expansion and compaction. A) Graphical summary illustrating the sliding window subregion analysis presented in this figure. **B)** Distribution of the normalized end-to-end distance obtained from all 51-residue subfragments 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 in red. **D)** Log_2-fold enrichment for amino acids found in expanded subregions.

High-throughput IDR ensemble informatics across yeast evolution

Having demonstrated the accuracy and throughput of ALBATROSS in conducting broad, proteome-wide analyses, we next sought to showcase ALBATROSS' unique advantages for structural bioinformatics of disordered proteins. In particular, we were motivated by recent work demonstrating that IDRs can conserve global dimensions in spite of variation in amino acid sequence, as reported for a linker region in the viral protein E1A[72]. Compelled by this specific example, we suspected that there would be other instances in nature in which functionally-important IDR dimensions are evolutionarily conserved across divergent homologs.

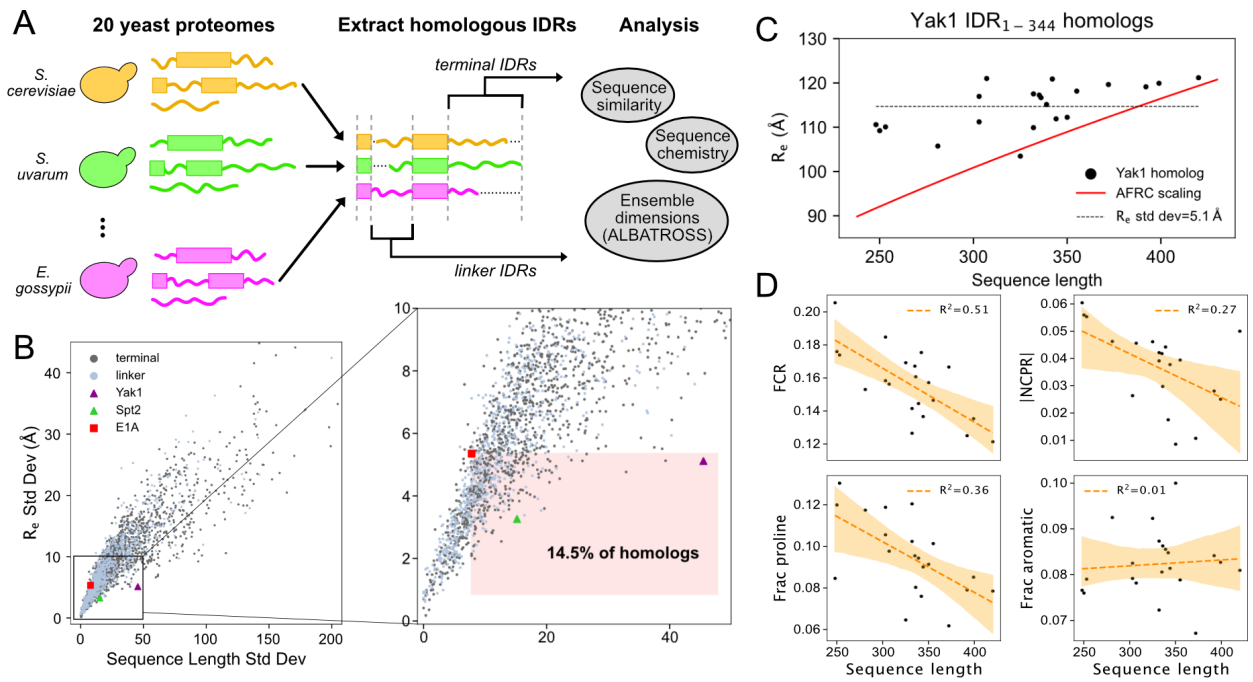

Figure 6. Applying ALBATROSS on homologous yeast IDRs identifies instances where primary sequences diverge yet IDR dimensions are maintained. A) Graphical summary illustrating how sets of homologous proteins from 20 yeast proteomes were aligned, then IDRs were extracted and analyzed using ALBATROSS and bioinformatics. **B)** For each set of homologous IDRs, standard deviation of the predicted R_e values is plotted with the standard deviation of sequence length (in number of amino acids). Linker IDR homologs and terminal IDR homologs are plotted in blue and gray, and several specific IDR homologs are marked. The right panel shows a zoomed-in inset, with a region denoting homologs with more divergent sequences and less variable dimensions than E1A are highlighted. **C)** Sequence length and ALBATROSS-predicted R_e for Yak1 and its yeast homologs. The gray dashed line denotes the mean R_e. The red line denotes R_e as a function of sequence length for an Analytical Flory Random Coil polymer scaling model. **D)** Yak1 homolog sequence length compared to various protein sequence features. Line of best fit and 95% CI denoted in orange. FCR=fraction of charged residues; |NCPR|=absolute value of net charge per residue.

To test this, we analyzed evolutionarily-related IDRs across a wide-ranging set of yeast species. Using the *S. cerevisiae* proteome as a reference, we aligned and extracted 2302 sets of homologous IDRs >40 residues from 20 yeast proteomes, totaling 49,335 IDRs (**Fig. 6A**). We predicted the R_e for all IDRs, and used the standard deviation of these predicted R_e values to quantify the conservation of IDR dimensions. We quantified homolog sequence divergence using two approaches: by computing the variation of IDR sequence lengths and by scoring the sequence similarities from the multiple sequence alignment (see *Methods*). In line with what has been reported elsewhere, both of these approaches for scoring sequence similarity show that homologous IDRs in our dataset are significantly more divergent than homologous folded domains (Mann-Whitney U test, $p < 0.001$) (**Fig. S11**) [70,98–100].

Looking at all of the sets of homologous IDRs, we see a clear relationship between sequence similarity and R_e conservation (**Fig. 6B, S12**). As expected, IDRs with more divergent sequences tend to possess larger variations in R_e, however there are a number of homologs that exhibit tightly coupled R_e relative to their sequence (dis)similarity, even compared to E1A[72] (**Fig. 6B, S12**). While characterizing all of these homologs is outside of the scope of this work, we did perform a deeper analysis into two of these candidate IDRs in order to validate this observation.

The first system we looked at was the N-terminal IDR of the DYRK-family kinase Yak1 and its homologs. Despite large variation in sequence length (248 - 420 residues), the Yak1 homologs displayed strikingly conserved dimensions, with all but two sequences having predicted R_e between 109 and 122 Å (**Fig. 6C**). In contrast, basic polymer models predict R_e differences of >27 Å across this range of sequence lengths, suggesting that there may be evolutionary constraints on the R_e of the N-terminal Yak1 IDR. Analysis of the IDR sequences reveals several trends in sequence features that may explain the buffering of chain dimensions. As sequence length increases, there is a decrease in the fraction of proline and charged residues and a modest trend towards a neutral net charge (**Fig. 6D, Table S3**). Each of these features are associated with chain compaction [72](**Fig. 3, Fig. 4, Fig. 5**). Other trends, such as a modest increase in κ, and changing sequence composition of polar and aliphatic residues may also modulate the dimensions and properties of the Yak1 homologs (**Fig. S13**).

We also examined the disordered linker (residues 57-373) from Spt2, a histone chaperone associated with chromatin remodeling during transcription. Like Yak1, the Spt2 IDR homologs have relatively constrained R_e despite spanning 186 to 239 residues in length (**Fig. 6B, S12, S14**). Through analysis of the sequence features of the Spt2 homologs, their dimension of the longer sequences appear to be modulated by an increase in proline and aliphatic content, and a more neutral net charge. Additionally, all of the homologs have high κ values which may buffer the overall chain dimensions (**Fig. S14, Table S4**).

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 [19,72,101].

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, 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. 4I**). 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. 5E**). Previous work has shown that disordered regions can chaperone RNA, both in isolation and in the context of biomolecular condensates[73,102–105]. 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* [23,78,85,106–110]. 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.

Our analysis of yeast homologs highlights two specific cases where IDR dimensions appear to be conserved across evolution despite substantial divergence in primary sequence, consistent with previous studies[72]. The homologs of the Yak1 N-terminal IDR homologs have more constrained R_e than we would expect based on polymer models (**Fig. 6C**). In the literature, Yak1 kinase activity has been shown to be regulated by its N-terminal IDR, through both intra- and intermolecular interactions[111]. We hypothesize that maintaining a narrow range of end-to-end dimensions of the N-terminal IDR across homologs could be important for preserving autophosphorylation capabilities and for facilitating specific, multivalent interactions with 14-3-3 proteins[111,112]. The IDRs of the histone chaperone Spt2 also maintain similar R_e values across divergent homolog sequences (**Fig. S14**). Spt2 has shown to interact with histone proteins and a variety of chromatin remodelers, with its IDR potentially playing an important role in these binding events[113,114]. Thus, we hypothesize that the Spt2 IDR dimensions are functionally conserved across homologs in order to preserve the protein's ability to act as a multivalent molecular scaffold. Beyond Yak1 and Spt2, our analyses suggest there are likely other IDRs with evolutionary constraints on their dimensions. Although we need further experiments to validate any of these claims we make regarding mechanism, we believe that this analysis illustrates ALBATROSS's utility in generating hypotheses and identifying candidates in large biological datasets.

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 to predict ensemble properties for coarse-grained simulations (idpGAN) [115]. This study also demonstrated the potential for multi-resolution models that interpolate between coarse-grained and atomistic simulations. Leveraging and interpolating between these multiscale models to generate thermodynamically distinct but realistic conformations may be a promising approach to enhance conformational sampling[116,117]. 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[118]. This work suggests that representing IDRs in a length-free way using the Bag of Amino Acids representation may be useful to capture sequence-specific effects and offers some advantages in model training. Finally, Tesei & Trolle et al. recently performed an analogous assessment of the human IDR-ome using the CALVADOS2 force field[48,53,119]. Despite using a different force field, 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. S15**, R^2 = 0.98, RMSE = 3.68 Å, n=29,998, ALBATROSS prediction time for all IDRs ~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 using Mpipi-GG and of ALBATROSS, there are a few important limitations to consider. Mpipi-GG is a one-bead-per-residue, coarse-grained force field that assumes an isotropic interaction potential. Despite this simplifying assumption, many independent studies have confirmed that coarse-grained models are able to capture global ensemble properties of IDRs with reasonably good accuracy[48,49,51,53,120]. 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 more compact 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.

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[121–126]. 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. Finally, combining an LSTM-based sequence-to-ensemble predictor with our high-throughout LSTM-based disorder predictor (metapredict V2-FF) ensures parity in disorder prediction and ensemble prediction. With this in mind, our IDR-ome predictor notebook enables proteome-wide predictions in minutes, democratizing high-throughput unstructural bioinformatics.

CONCLUSION

In this work, we present ALBATROSS, an accessible and accurate route to predict IDR global dimensions from sequence. 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. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. *Nat. Commun.* **8**, 15231 (2017).
32. Gibbs, E. B. *et al.* 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. *et al.* 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. *et al.* 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. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61–66 (2018).

44. Kassem, N. *et al.* 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. *et al.* Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J. Am. Chem. Soc.* **138**, 11714–11726 (2016).
47. Zheng, W. *et al.* 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. *et al.* 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. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
59. Thompson, A. P. *et al.* LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171 (2022).
60. McGibbon, R. T. *et al.* MDTraj: a modern, open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
61. Tange, O. *GNU Parallel 20210622 ('Protasevich')*. (2021). doi:10.5281/zenodo.5013933.
62. Rubinstein, M. & Colby, R. H. *Polymer Physics*. (Oxford University Press, 2003).
63. 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).
64. Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict V2: An update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *bioRxiv* 2022.06.06.494887 (2022) doi:10.1101/2022.06.06.494887.
65. 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).
66. 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.

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

68. 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).

69. Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15**, 1456–1461 (2005).

70. Langstein-Skora, I. *et al.* Sequence- and chemical specificity define the functional landscape of intrinsically disordered regions. *bioRxiv* 2022.02.10.480018 (2022) doi:10.1101/2022.02.10.480018.

71. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

72. González-Foutel, N. S. *et al.* Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat. Struct. Mol. Biol.* **29**, 781–790 (2022).

73. Cubuk, J. *et al.* 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.

74. 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).

75. 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.

76. Peran, I. *et al.* 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).

77. Mao, A. H., Lyle, N. & Pappu, R. V. Describing sequence–ensemble relationships for intrinsically disordered proteins. *Biochem. J* **449**, 307–318 (2013).
78. 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).
79. Alshareedah, I. *et al.* Sequence-encoded grammars determine material properties and physical aging of protein condensates. *bioRxiv* (2023) doi:10.1101/2023.04.06.535902.
80. 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).
81. Riback, J. A. *et al.* Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238–241 (2017).
82. 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).
83. Riback, J. A. *et al.* Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168**, 1028–1040.e19 (2017).
84. 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).
85. Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **174**, 688–699.e16 (2018).
86. 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).
87. 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).

88. Bremer, A. *et al.* Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14**, 196–207 (2022).
89. Choi, J.-M., Hyman, A. A. & Pappu, R. V. Generalized models for bond percolation transitions of associative polymers. *Phys Rev E* **102**, 042403 (2020).
90. Fei, J. *et al.* Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *J. Cell Sci.* **130**, 4180–4192 (2017).
91. Zhao, Y. *et al.* GIGYF1 loss of function is associated with clonal mosaicism and adverse metabolic health. *Nat. Commun.* **12**, 4178 (2021).
92. Deaton, A. M. *et al.* 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).
93. Chen, G. *et al.* GIGYF1 disruption associates with autism and impaired IGF-1R signaling. *J. Clin. Invest.* **132**, (2022).
94. Calnan, B. J., Tidor, B., Biancalana, S., Hudson, D. & Frankel, A. D. Arginine-mediated RNA recognition: the arginine fork. *Science* **252**, 1167–1171 (1991).
95. Cléry, A., Blatter, M. & Allain, F. H.-T. RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.* **18**, 290–298 (2008).
96. Hall, K. B. RNA–protein interactions. *Curr. Opin. Struct. Biol.* **12**, 283–288 (2002).
97. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Mol. Cell* **78**, 9–29 (2020).
98. Brown, C. J., Johnson, A. K., Dunker, A. K. & Daughdrill, G. W. Evolution and disorder. *Curr. Opin. Struct. Biol.* **21**, 441–446 (2011).
99. Zarin, T., Tsai, C. N., Nguyen Ba, A. N. & Moses, A. M. Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc. Natl. Acad. Sci. U. S. A.*

114, E1450–E1459 (2017).

100. Zarin, T. *et al.* Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife* **10**, (2021).

101. Zeno, W. F. *et al.* Synergy between intrinsically disordered domains and structured proteins amplifies membrane curvature sensing. *Nat. Commun.* **9**, 4152 (2018).

102. 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).

103. 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).

104. Sarni, S. H. *et al.* Intrinsically disordered interaction network in an RNA chaperone revealed by native mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2208780119 (2022).

105. Zúñiga, S. *et al.* Coronavirus nucleocapsid protein is an RNA chaperone. *Virology* **357**, 215–227 (2007).

106. 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).

107. Nott, T. J. *et al.* Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57**, 936–947 (2015).

108. Han, T. W. *et al.* Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell* **149**, 768–779 (2012).

109. Kato, M. *et al.* Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012).

110. 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).

111. Lee, P., Paik, S.-M., Shin, C.-S., Huh, W.-K. & Hahn, J.-S. Regulation of yeast Yak1 kinase by PKA and autophosphorylation-dependent 14-3-3 binding. *Mol. Microbiol.* **79**, 633–646 (2011).

112. Parua, P. K. & Young, E. T. Binding and transcriptional regulation by 14-3-3 (Bmh) proteins requires residues outside of the canonical motif. *Eukaryot. Cell* **13**, 21–30 (2014).

113. Bhat, W., Boutin, G., Rufiange, A. & Nourani, A. Casein kinase 2 associates with the yeast chromatin reassembly factor Spt2/Sin1 to regulate its function in the repression of spurious transcription. *Mol. Cell. Biol.* **33**, 4198–4211 (2013).

114. Warren, C. & Shechter, D. Fly Fishing for Histones: Catch and Release by Histone Chaperone Intrinsically Disordered Regions and Acidic Stretches. *J. Mol. Biol.* **429**, 2401–2426 (2017).

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

116. Meller, A., Bhakat, S., Solieva, S. & Bowman, G. R. Accelerating Cryptic Pocket Discovery Using AlphaFold. *J. Chem. Theory Comput.* (2023) doi:10.1021/acs.jctc.2c01189.

117. Vani, B. P., Aranganathan, A., Wang, D. & Tiwary, P. AlphaFold2-RAVE: From Sequence to Boltzmann Ranking. *J. Chem. Theory Comput.* (2023) doi:10.1021/acs.jctc.3c00290.

118. 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.

119. Tesei, G. *et al.* 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.

120. Mugnai, M. L. *et al.* Sizes, Conformational fluctuations, and SAXS profiles for Intrinsically Disordered Proteins. *bioRxiv* 2023.04.24.538147 (2023) doi:10.1101/2023.04.24.538147.

121. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional

Transformers for Language Understanding. *arXiv [cs.CL]* (2018).

122. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

123. Lin, Z. *et al.* 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.

124. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

125. Elnaggar, A. *et al.* Ankh ☥ : Optimized Protein Language Model Unlocks General-Purpose Modelling. *bioRxiv* 2023–2001 (2023).

126. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: The Long-Document Transformer. *arXiv [cs.CL]* (2020).