RESEARCH ARTICLE SUMMARY

BIOPHYSICS

Sequence-based prediction of intermolecular interactions driven by disordered regions

Garrett M. Ginell, Ryan J. Emenecker, Jeffrey M. Lotthammer, Alex T. Keeley, Stephen P. Plassmeyer, Nicholas Razo, Emery T. Usher, Jaqueline F. Pelham, Alex S. Holehouse*

Full article and list of author affiliations: https://doi.org/10.1126/science.adq8381

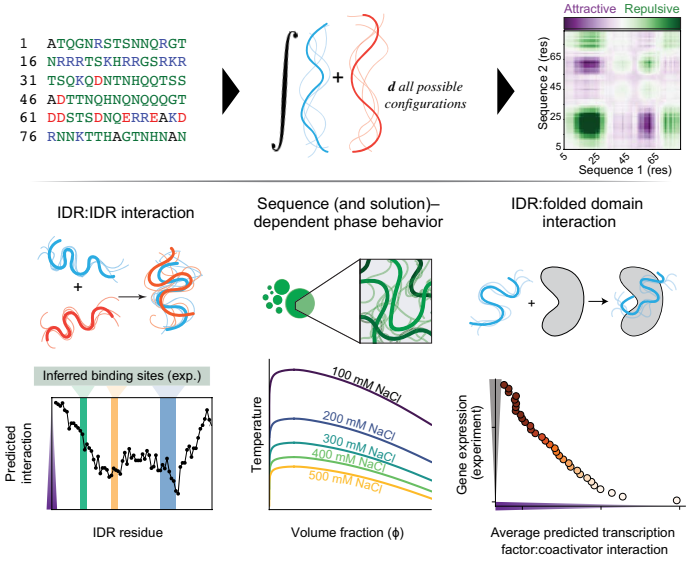
INTRODUCTION: Intrinsically disordered regions (IDRs) are found in >70% of human proteins and can play crucial roles in many cellular processes. IDRs lack a stable three-dimensional structure yet often play key roles in mediating complex interactions with various cellular partners. These interactions can be sequence specific, leading to bound states with a structured interface, or chemically specific, leading to an ensemble of bound configurations. Although deep-learning models can predict sequence-specific interactions, predicting chemically specific molecular recognition remains challenging.

RATIONALE: The past few years have seen substantial progress in the accuracy of simple (coarse-grained) molecular force fields for describing the biophysical properties of IDRs. Force fields are a set of equations and numbers that capture the chemical physics representing how residues in a disordered protein interact. Force fields were developed to perform molecular simulations, which can be slow and challenging, especially for larger IDRs. We reasoned that it might be possible to repurpose force fields to predict intermolecular interactions without performing simulations for cases where bound states lack residual structure. To test this idea, we developed a computational framework, FINCHES, which enables different force field functional forms and parameters to predict IDR-mediated chemical specificity. FINCHES enables direct predictions of which residues or regions in an IDR are expected to provide attractive and repulsive interactions with a partner, be that the surface of a folded domain or another IDR.

RESULTS: We tested FINCHES-based predictions by focusing on three types of behavior: predicting IDR-IDR interactions, predicting phase separation propensity from sequence, and predicting IDR–folded domain interactions. In all cases, we found good qualitative or semiquantitative agreement between predictions and experiments across various proteins and systems. Because these predictions are analytical and based on an underlying energy function, they are fast, tunable, and fully interpretable. This enables proteome-scale interrogation in minutes. We used FINCHES to explore the chemical structure of IDRs across the human proteome, delineate large IDRs into subdomains, and systematically investigate the consequences of phosphorylation on IDR-mediated interactions at a proteome scale. Most importantly, FINCHES provides an easy and fast way to develop precise molecular hypotheses regarding the driving forces for interaction that may underlie IDR-mediated function.

CONCLUSION: Although many caveats are associated with predicting chemical specificity in this way (e.g., sequence-specific interactions will not be captured, so structured interactions mediated by disordered regions are missed), we see FINCHES as providing a new and complementary set of information for understanding IDR-mediated function. FINCHES is fully open source and available as a Python package (https://github.com/idptools/finches) and through an easy-to-use web server (https://www.finches-online.com/). □

*Corresponding author. Email: alex.holehouse@wustl.edu Cite this article as G. M. Ginell *et al.*, *Science* **388**, eadq8381 (2025). DOI: 10.1126/science.adq8381



Intrinsically disordered regions can interact with partners driven by complementary chemistry. FINCHES uses chemical physics from molecular force fields to analytically predict which regions and residues in an IDR can drive attractive or repulsive interactions with a partner protein (top). This program enables the de novo prediction of regions that drive IDR-mediated binding (bottom left), how sequence and environment cooperate to tune IDR-mediated phase separation (bottom middle), and how IDRs can interact with the surfaces of folded domains (bottom right).

BIOPHYSICS

Sequence-based prediction of intermolecular interactions driven by disordered regions

Garrett M. Ginell[1,2], Ryan J. Emenecker[1,2], Jeffrey M. Lotthammer[1,2], Alex T. Keeley[1,2], Stephen P. Plassmeyer[1,2], Nicholas Razo[1,2], Emery T. Usher[1,2], Jaqueline F. Pelham[1,2], Alex S. Holehouse[1,2]*

Intrinsically disordered regions (IDRs) in proteins play essential roles in cellular function. A growing body of work has shown that IDRs often interact with partners in a manner that does not depend on the precise order of amino acids but is instead driven by complementary chemical interactions, leading to disordered bound-state complexes. However, these chemically specific dynamic interactions are difficult to predict. In this study, we repurposed the chemical physics developed originally for molecular simulations to predict this chemical specificity between IDRs and partner proteins using protein sequence as the only input. Our approach—FINCHES—enables the direct prediction of phase diagrams, the identification of chemically specific interaction hotspots on IDRs, the decomposition of chemically distinct domains in IDRs, and a route to develop and test mechanistic hypotheses regarding IDR function in molecular recognition.

Intrinsically disordered regions (IDRs) in proteins are prevalent across the kingdoms of life (*1*). While folded domains exist in a stable three-dimensional (3D) structure, IDRs exist in a heterogeneous ensemble of conformations. Despite lacking a fixed tertiary structure, disordered regions can play essential roles across many distinct cellular processes, and their function often depends on molecular interactions with a variety of partners (*1*, *2*). While some IDRs fold upon binding, in many cases, some degree of structural disorder is retained in the bound state, leading to so-called "fuzzy" interactions (*3*). IDRs can bind partners through sequence-specific or chemically specific molecular interactions (*1*, *4*). In sequence-specific recognition (also called site-specific recognition), the bound state acquires some amount of order, leading to a conventional (although potentially transient) structured interface with a partner (Fig. 1A, left). In chemically specific molecular interactions, the bound state lacks a stable structure, and attractive interactions are driven by complementary chemistry that enables an ensemble of bound configurations (Fig. 1A, right). While deep-learning models trained on structural information are poised to uncover sequence-specific interactions (*5*), they are less well suited for elucidating chemically specific molecular recognition.

We address this knowledge gap through the lens of molecular biophysics. By repurposing the chemical physics originally developed for molecular force fields and discarding requirements for spatial information, we offer a bottom-up framework for predicting chemically specific IDR-mediated interactions. Our approach—implemented in the FINCHES software package—enables analytical prediction of how IDRs interact with other IDRs or with folded domain surfaces. Our work opens the door to high-throughput, straightforward, and interpretable prediction of IDR-mediated intermolecular interactions to generate hypotheses, predict phase behavior, identify distinct domains, and aid in the interpretation of experimental results.

Repurposing molecular force fields for high-throughput bioinformatic analysis of IDR interactions

Molecular force fields describe the chemical physics of biomolecules through a series of equations and parameters (Fig. 1, B and C). Recent work on coarse-grained models of disordered proteins has led to several force fields that offer accurate predictions of global IDR dimensions, notably among those of the Mpipi and CALVADOS families (*6*–*9*). These models prescribe a set of equations and parameters that quantify the nonbonded interactions between every pair of amino acids to describe the attraction or repulsion of a pair of residues at some arbitrary distance (Fig. 1D). We reasoned that this chemical physics—while generally used for simulation—could be stripped out and repurposed by taking the integral under the pairwise potential as a means to calculate a mean-field interaction parameter between two residues, akin to a Mayer f-function without a volume correction component (see supplementary materials) (Fig. 1D). Using this force field–derived interaction parameter as a starting point and then tuning the interaction of individual amino acids on the basis of their local context (for charged and aliphatic hydrophobic residues, specifically), the resulting inter-residue matrix between a pair of residues can be averaged to obtain a single mean-field interprotein interaction parameter epsilon (ε) or averaged over a sliding window to decode local intermolecular interactions that are attractive or repulsive (Fig. 1E and fig. S1). We refer to the predicted intermolecular interaction maps as "intermaps," as shown at the bottom left of Fig. 1E. In all cases, negative ε values are attractive, and positive ε values are repulsive.

A central assumption in this approach is that the attraction between two IDRs is mediated solely by complementary chemical interactions (i.e., chemical specificity) that emerge in the limit of all possible configurations being sampled, not through precise "structured" interaction between subregions. It also does not allow coordination between distinct regions, which introduces several caveats (see Discussion section). Nevertheless, this approach enables us to easily calculate a mean-field interaction parameter between two sequences as well as identify subregions that are expected to drive attractive and repulsive interactions. Unlike deep-learning approaches, the molecular determinants for such interactions are entirely interpretable and codified by the force field's underlying functional form and associated parameters.

This study uses the Mpipi-GG and CALVADOS2 force fields in FINCHES, although other force fields can be implemented (*7*, *9*). These force fields allow us to modulate the solution environment in terms of salt by means of a Debye-Hückel term, allowing us to investigate salt-dependent effects. They also enable high-throughput prediction (>1000 sequences per second for a 100-residue IDR; fig. S2). While we emphasize that the resulting interaction scores are not expected to offer high-resolution, quantitative predictions, they do enable rapid semiquantitative descriptions of likely IDR-associated interactions.

Validation against molecular interaction

We first investigated whether the calculated mean-field interaction parameter (ε) between two sequences is proportional to experimentally measurable values for intermolecular interaction. Osmotic and light-scattering second virial coefficients (B_2 and A_2, respectively) are experimental measurements that report on the deviation from the so-called "ideal behavior" of noninteracting molecules (*10*). A negative B_2 or A_2 implies net attractive intermolecular interactions, whereas a positive B_2 or A_2 implies a net repulsive intermolecular interaction. We considered two systems where second virial coefficients have previously been characterized for IDRs: variants of the low-complexity domain of the RNA binding protein FUS and the RGG domain from the DEAD-box helicase LAF-1 (*11*, *12*). For FUS, A_2 values were calculated

[1]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO, USA. [2]Center for Biomolecular Condensates (CBC), Washington University in St. Louis, St. Louis, MO, USA. *Corresponding author. Email: alex.holehouse@wustl.edu

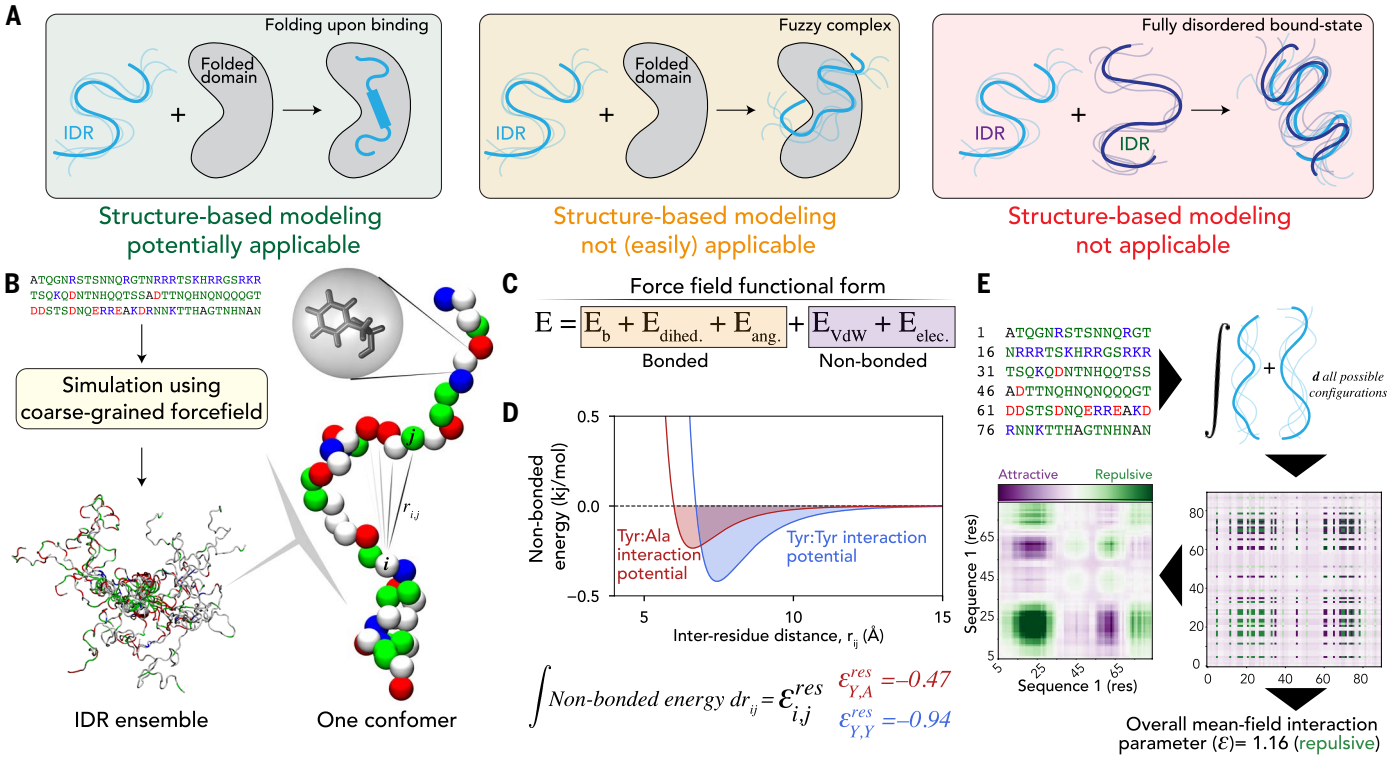


Fig. 1. Coarse-grained force field parameters can be repurposed for informatic analysis. (**A**) IDR intermolecular interaction can occur through a conventional structured interface (left), a disordered bound-state complex with a folded partner (middle), or a disordered bound-state complex where both partners are disordered (right). (**B**) Coarse-grained force fields can represent amino acids as individual beads and generate IDR conformational ensembles by sampling energetically accessible conformers in 3D space. (**C**) Force fields describe bonded and nonbonded components, where nonbonded reflects short-range and long-range interactions that determine attraction or repulsion between individual residues. (**D**) Nonbonded interactions are defined by distance-dependent potentials that describe the relationship between interbead distance and instantaneous potential energy. Integrating under these potentials yields a parameter proportional to the overall attraction or repulsion between those beads. (**E**) By assuming that two proteins can interact through all possible configurations without concern for chain connectivity, we can calculate inter-residue preferential interaction coefficients and then use local sequence context to convert these into smoothed predicted intermolecular interaction maps (intermaps) or a single mean-field intermolecular interaction parameter (epsilon, ε).

for a series of mutants (Fig. 2A), with values correlating well with predicted ε values (fig. S3). Despite distinct scales, the value of 0 should be equivalent in ε and A_2 space, a prediction confirmed by the fact that the best-fit line travels through (0,0) (Fig. 2A). Intermaps comparing wild-type FUS with the tyrosine-to-serine mutant (Y2S; in which all tyrosines are replaced by serine) illustrates the complete suppression of attractive interactions (Fig. 2B). For the RGG domain, we calculated the salt-dependent ε values and compared them against NaCl-dependent B_2 values, yielding a 1:1 correspondence between measured values and predictions (Fig. 2C). While there are many caveats associated with relating second virial coefficients to our mean-field interaction parameter, the trends here gave us confidence that our underlying assumptions were reasonable.

Direct prediction of phase diagrams from sequence

Recent interest in biomolecular phase separation has led to many experimental studies characterizing full phase diagrams of disordered proteins in vitro. Predicting phase behavior from sequence has been a goal for many predictors, but conceptual and technical challenges have limited their generalizability and scope (see Discussion section).

We first sought to use ε values as input to Flory-Huggins theory, from which phase diagrams can be predicted. Flory-Huggins theory is a simple mean-field solution mixing theory that considers the balance between entropy and enthalpy to determine whether a solution of a given temperature and composition will exist in a single phase or multiple phases (*13*). Although Flory-Huggins theory is reductive, our goal was not to quantitatively reproduce coexistence curves to match 1:1 with experiments but to provide qualitative predictions for how changes in environment or sequence are expected to alter phase diagrams. To achieve this, we calculated homotypic ε values for proteins where full phase diagrams have previously been measured, converted the (extensive) ε into an (intensive) Flory χ parameter by dividing by the sequence length, and used the recently developed analytical solution to the Flory-Huggins model of Qian *et al.* to solve full phase diagrams for a series of systems (Fig. 3A) (*14*). For a detailed overview of how to read phase diagrams, see fig. S4A. Our predicted phase diagrams report temperature normalized by the critical temperature of a reference sequence (T/T_C) and concentration as volume fraction (ϕ). Phase diagrams here were predicted with the Mpipi-GG–based ε analysis, but equivalent results are obtained using CALVADOS2-based analysis.

To assess how strongly our predictions depend on primary structure—that is, the precise order of amino acids in the sequence—we implemented a random shuffle null model. For this null model, we generated large numbers of randomly shuffled variants, computed the average interaction score across those variants, and then compared those interaction scores (or intermaps) with the specific sequence of interest (fig. S4B). These analyses confirm that intermap predictions tightly depend on the primary structure, consistent with the remainder of our work.

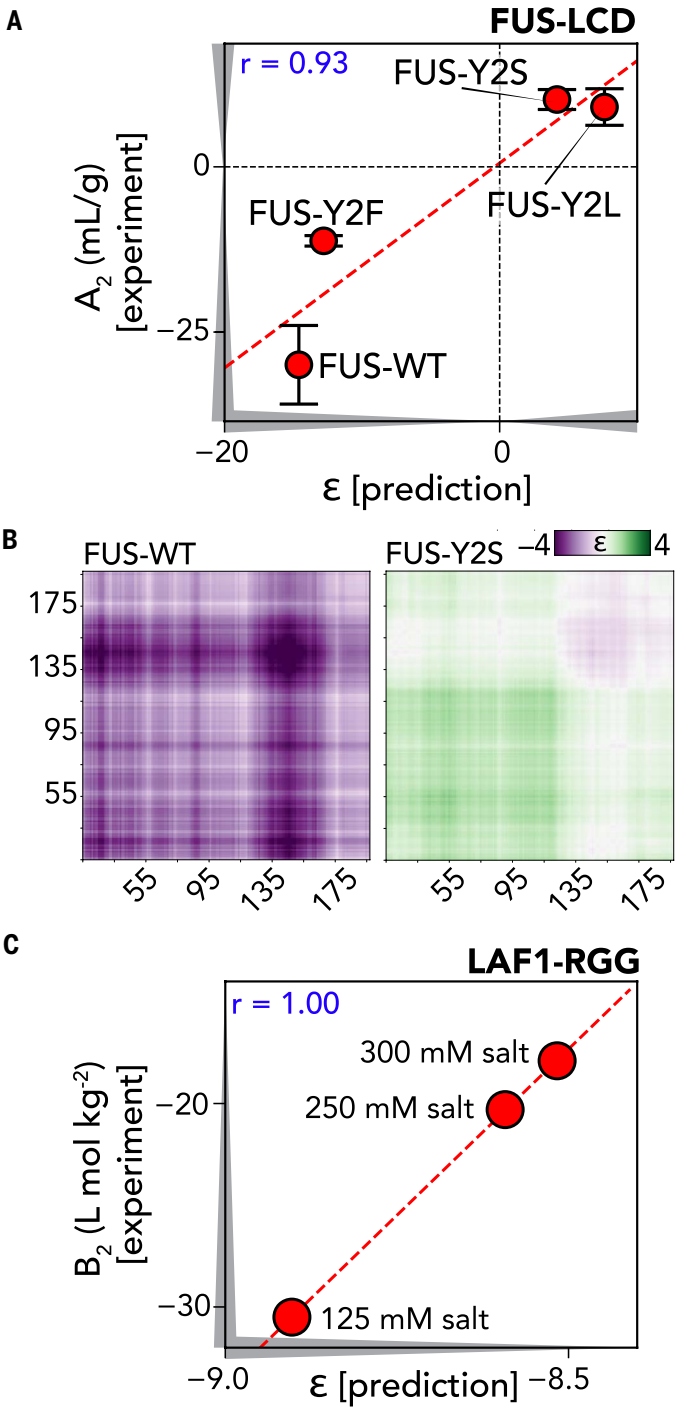


Fig. 2. Intermolecular interaction coefficients depend on sequence and environment. (**A**) Comparison of the light-scattering second virial coefficient (A_2, which under dilute conditions is strongly correlated with the osmotic second virial coefficient) with overall ε value for different IDRs [data taken from (*11*)]. Note that experiments were performed with FUS and its variants tethered to a common scaffold system (see supplementary materials). (**B**) Intermaps for FUS and FUS-Y2S, illustrating how intermolecular interactions vary. (**C**) Comparison of the osmotic second virial coefficient with overall ε value for the same IDR under different solution conditions. Here, salt weakens intermolecular interactions (ε becomes less negative).

We began with previously characterized aromatic variants of the low-complexity domain of the RNA binding protein hnRNPA1 (heterogeneous nuclear ribonucleoprotein A1), a 135-residue low-complexity prion-like domain (Fig. 3B) (*15*). These variants increase (Aro^+) or decrease (Aro^-, Aro^-) the number of aromatic residues in the sequence. Our approach yielded full phase diagrams that show good agreement with respect to the relative impact of aromatic mutations (Fig. 3C). Prior work to elucidate these phase diagrams combined simulations and experiments to arrive at conclusions regarding the impact of aromatic residues on phase behavior. In contrast, the major benefit of our approach is that these phase diagrams can now be predicted directly from the sequence in seconds.

We next asked how well our approach could capture the solution environment, sequence patterning, and differences between charged residues. Sequence patterning refers to the relative positions of amino acids along the sequence, where the patterning of charged residues, in particular, has been shown to modulate phase behavior in various systems (*16*–*18*). The N-terminal intrinsically disordered domain of the RNA helicase DDX4 (DDX4-NTD) has been extensively studied in this context (Fig. 3D) (*19*, *20*). Intermaps identify both charge clusters and aromatic residues as key drivers of intermolecular interaction (Fig. 3D, bottom). Prior work measured full phase diagrams as a function of NaCl (Fig. 3E), which are correctly reproduced with our approach (Fig. 3F). Moreover, charge shuffle (CS) variants that maintain the same composition but reposition a small number of charged residues lead to changes in the phase diagram that are correctly recapitulated by our approach, as are arginine-to-lysine and phenylalanine-to-alanine mutants that entirely suppress phase behavior (Fig. 3, G and H, and fig. S4, B and C). These results illustrate that our approach can capture effects driven by sequence patterning, changes in the identity of cationic residues, and the solution environment.

Beyond these examples, we predicted full phase diagrams for various previously examined systems, including variants of hnRNPA1-LCD, FUS, RLP, and LAF-1 (fig. S5) (*11*, *15*, *21*–*24*). These predictions recapitulate a variety of previously examined aspects of the molecular grammar of phase separation, including differences imparted by aromatic residues, aromatic versus aliphatic hydrophobes, charge shuffling, the role of amyloidogenic regions on phase separation, and positional sensitivity. We also performed an extensive investigation into the low-complexity domain of TDP-43, highlighting that our approach correctly integrates the density of aliphatic residues in the TDP-43 conserved region, recovering experimentally reported consequences of changing a variety of different chemistries (fig. S6) (*25*, *26*). In all cases tested, our predictions (at least qualitatively) capture the effects of sequence chemistry on phase behavior and offer clear explanatory power for how sequence changes are expected to alter intermolecular interactions in the context of IDR-mediated phase separation.

We next wondered how prevalent sequences with the potential to phase separate are across the human proteome. To explore this, we calculated homotypic ε values for all long (>100 residues) IDRs in the human proteome using Mpipi-GG and CALAVDOS2 (tables S1 and S2). We saw differences in the fraction of IDRs with attractive homotypic ε values ($\varepsilon < 0$), with ~10% of IDRs using Mpipi-GG and ~15% using CALVADOS2 (Fig. 3I). An attractive homotypic ε value does not necessarily mean the IDR will undergo homotypic phase separation in a biochemically relevant context (see fig. S7, A and B). However, it does imply the IDR has the potential to self-interact. Gene Ontology (GO) analysis of proteins with IDRs that have attractive homotypic ε values (in comparison to all proteins with long IDRs) identifies RNA-associated processes, morphogenesis, and development as key biological processes associated with these proteins (Fig. 3J and fig. S7C). The enrichment for RNA binding proteins qualitatively agrees with recent proteome-wide analyses on IDR compaction, highlighting the symmetry between intramolecular interactions

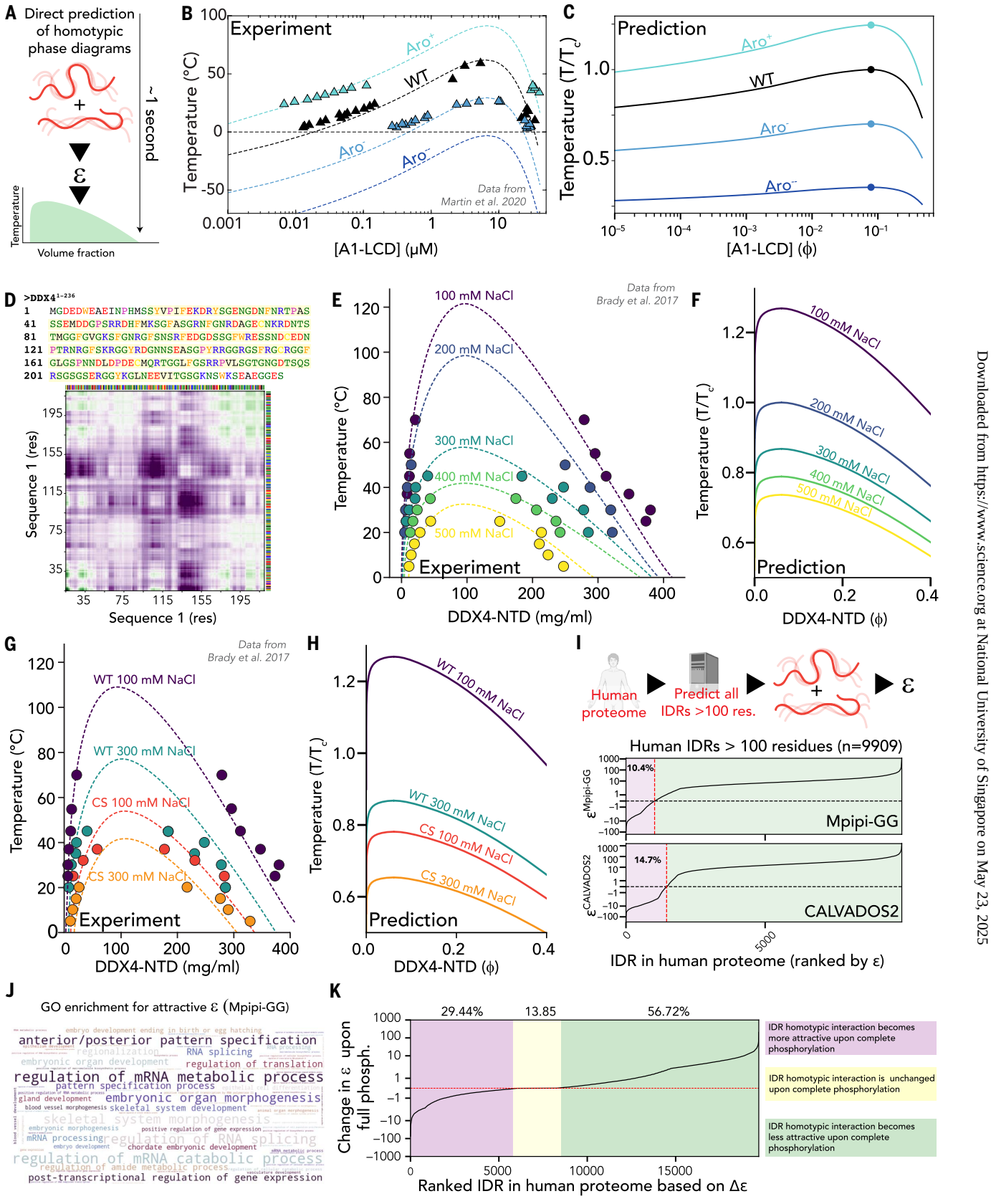


Fig. 3. Full phase diagrams and the regions that drive phase separation can be predicted directly from sequence. (**A**) Using ε as a means to parameterize a Flory-Huggins description of phase behavior, our ε-based approach enables the prediction of sequence- and solution-dependent full phase diagrams in seconds. (**B**) Experimentally measured phase diagrams for four variants of the low-complexity domain (LCD) from the RNA binding protein hnRNPA1. These measurements were originally reported by Martin *et al*. and are reproduced here for easy comparison with predictions (*15*). (**C**) Direct predictions of full phase diagrams using only sequence as input. (**D**) The N-terminal IDR from the

DEAD-box helicase protein DDX4 (DDX4-NTD) is a polyampholytic disordered protein that undergoes phase separation in vitro. Intermaps identify subregions that drive phase separation. Alongside each intermap dimension is a chemical representation of the sequence (green, polar; red, acidic; blue, basic; orange, aromatic; black, aliphatic). (**E**) Experimentally measured phase diagrams for wild-type DDX4-NTD under five different salt concentrations. These measurements were originally reported by Brady *et al*. and are reproduced here for easy comparison with predictions (*19*). (**F**) Direct predictions of salt-dependent full phase diagrams using only sequence and salt concentration as input. (**G**) Experimentally measured phase diagrams for wild-type and charge shuffle (CS) DDX4-NTD variants under two different salt concentrations. These measurements were originally reported by Brady *et al*. and are reproduced here for easy comparison with predictions (*19*). (**H**) Direct predictions of sequence and salt-dependent full phase diagrams using only sequence and salt concentration as input. Note that in all cases, the composition of the sequence is identical here, and salt and the order of amino acids are varying. (**I**) Proteome-wide analysis of homotypic ε values for all IDRs in the human proteome of >100 amino acids. (**J**) Gene Ontology enrichment for proteins with attractive ε compared with all human proteins with one or more IDRs that are longer than 100 amino acids. RNA-associated biological processes are strongly enriched (see also fig. S7C). (**K**) Proteome-wide analysis for change in homotypic ε values for all IDRs upon complete phosphorylation.

and homotypic intermolecular interactions (fig. S7) (*9*, *15*, *27*). Moreover, intermaps enable direct interrogation of the sequence features that underlie an attractive homotypic ε value, as highlighted for the hub protein HAX1 (fig. S7D). As a final note, we emphasize explicitly and unconditionally that we make no claims whatsoever as to the physiological relevance of these predicted attractive homotypic interactions.

Given that IDR-mediated chemical specificity depends on the amino acid–encoded chemistry, posttranslational modifications offer one route to recode that chemistry. To this end, we calculated all homotypic ε values for all human IDRs with one or more phosphosites (19,703 IDRs) before and after making phosphomimetic mutations. Only experimentally reported phosphosites (Ser, Thr, or Tyr) were used, and in all cases, they were converted to Glu (glutamic acid). Notably, ~57% of IDRs that undergo phosphorylation showed a reduction in homotypic attractive interaction upon phosphorylation (Fig. 3K and fig. S8), whereas ~30% showed an increase in homotypic attractive interaction (table S3). Our analysis predicts phosphorylation of the C-terminal IDR from the tight junction protein Zonula Occludens (ZO1) will drastically reduce homotypic interaction, a result strongly supported by extant data (fig. S8, G and H) (*28*). Overall, these analyses suggest many IDRs poised for homotypic (and likely heterotypic) interactions that can be rewired through posttranslational modifications.

Organizing IDRs in intermolecular chemical space

Given that IDRs often appeared poorly conserved as assessed by multiple sequence alignments, we next investigated whether we could group IDRs in terms of their chemical interactions (as opposed to sequence similarity). We identified all IDRs in the human proteome between 100 and 150 residues in length (3414 IDRs), focusing on this specific size to ensure intermolecular pairs were approximately equal in length. We then computed all possible pairwise interactions (around 12 million calculations), allowing us to map the heterotypic interaction landscape at the proteomic scale (Fig. 4A).

Hierarchical clustering into 24 clusters revealed subsets of IDRs that showed similar chemical specificity interaction fingerprints, despite being highly diverse in terms of their primary structure (Fig. 4B). Of note, only 15% of heterotypic IDR-IDR interactions are attractive, with the majority being repulsive. While overall repulsive ε does not mean two IDRs cannot interact, our work here illustrates two key points: First, when classified in terms of mean-field intermolecular interaction, naturally occurring IDRs fall into distinct chemical niches. Second, sequence chemistry determines whether an IDR is poised for homotypic or heterotypic attractive interactions, and some chemistries appear much more promiscuous than others.

We next sought to explore the idea of chemical promiscuity further. We define chemical promiscuity as the tendency for an IDR to have attractive ε values for many different potential partners. To quantify this, we ranked each IDR (as defined in Fig. 4A) by the number of attractive heterotypic ε values found (Fig. 4C). This analysis reveals many IDRs have the potential to be highly promiscuous. Excising the top 100 most promiscuous IDRs, we identified a range of molecular functions, notably RNA binding, but also proteins involved in cellular homeostasis (e.g., TRIM41, DCAF1, ANAPC15), apoptosis (ANP32B, ANP32A, SET, CLK2, GRINA), transcriptional regulation (ANP32A, SET), and histone chaperoning (ANP32E, SET) (table S6). Moreover, taking protein copy number information into account, high-abundance and promiscuous IDR-containing proteins are almost universally RNA binding proteins (*29*) (fig. S9). In short, while most possible IDR-IDR interactions are repulsive (i.e., most rows in Fig. 4B are largely green), all IDRs do interact favorably with at least one other IDR (i.e., each row in Fig. 4B has at least one purple pixel), and many IDRs are in principle highly promiscuous (i.e., some rows in Fig. 4B are largely purple, for example, clusters 23 and 24).

The ability to segment IDRs on the basis of intermolecular chemical specificity is not limited to proteome-scale analyses. While identifying IDRs in proteins is now relatively straightforward, subclassifying internally distinct subdomains has been historically challenging. A major confounding factor here is that IDR function is context specific; as such, a "functional domain" only makes sense to define in the context of some function. If we restrict ourselves to chemically specific molecular interactions as our function of interest, it becomes possible to define distinct subdomains within an IDR in the context of some interaction partner. For a given pair of IDRs, we can segment subregions into chemically distinct domains, offering clear guidelines for subdomain deletion studies beyond arbitrary cut-off points (Fig. 3E and fig. S10). This is illustrated here in proteins with IDRs with clear compositional biases (e.g., the N-terminal half of the yeast prion protein Sup35) but is perhaps most useful for segmenting large IDRs where interaction partners are not yet known, using a limited set of chemical fingerprints (figs. S10 and S11) (e.g., the highly disordered protein BRCA2).

Decoding chemical specificity for IDR-mediated interactions

Finally, we examined binary interactions between IDRs and their partners to establish whether FINCHES-based analyses can guide and interpret detailed biophysical investigations. Prothymosin alpha (ProTα) and histone H1 (H1.0) coassemble into a fully disordered complex with picomolar affinity (Fig. 5A) (*30*). ProTα is entirely disordered, whereas H1 contains a small globular domain flanked by N- and C-terminal disordered regions (fig. S12). We wondered whether FINCHES would allow us to discern specific subregions that contribute more or less to binding. Extant binding data obtained from single-molecule Förster resonance energy transfer (smFRET) showed that the C-terminal half of H1 binds ProTα with a dissociation constant (K_d) of 0.04 nM, while the N-terminal region binds much less tightly, with a K_d of 173 nM (Fig. 5B). By calculating the per-residue sum of ProTα:H1 intermaps (fig. S12), we find a stark difference in the predicted interaction strength between the two halves (Fig. 5C). This effect is captured for both Mpipi- and CALAVDOS-based models, with both predicting that the C-terminal half will bind much more strongly than the N-terminal half (Fig. 5D), in good agreement with recent deep-learning–based approaches to predict disordered bound-state ensembles (*31*). In addition, a strong salt dependence on this interaction is predicted, in line with published work (fig. S12B). Finally, we recapitulate per-residue interaction profiles from the perspective of ProTα with histone

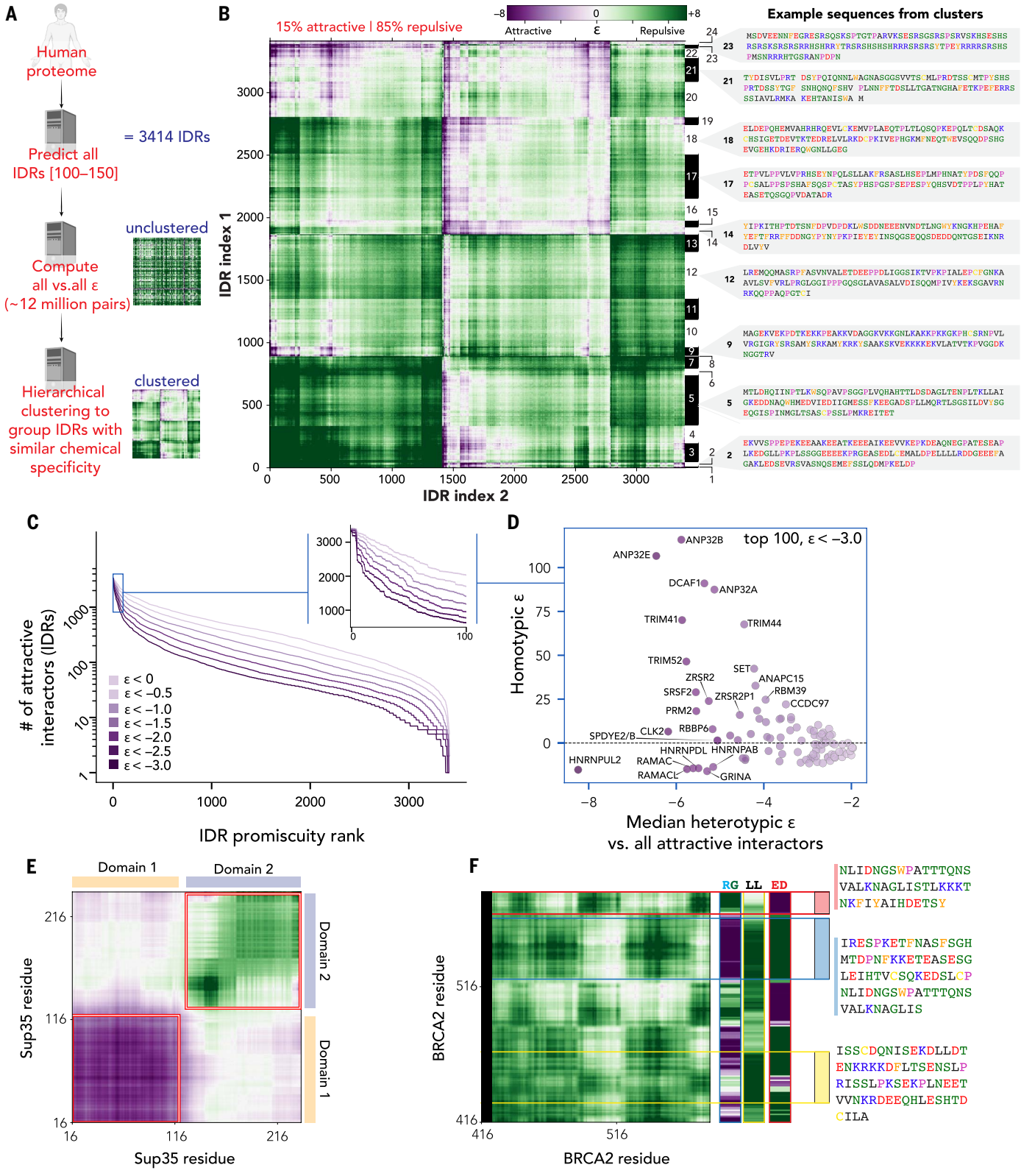


Fig. 4. Proteome-wide analysis reveals chemical structure of disordered regions. (**A**) Overview of workflow for construction of heterotypic chemical interaction map. (**B**) IDRs (3414) clustered by hierarchical clustering into groups that show similar global intermolecular interaction fingerprints (table S4). The *x* and *y* indices reflect the identity of those 3414 IDRs. The average chemical properties of sequences in each cluster are shown in table S5. (**C**) All IDRs are ranked by the number of attractive interactions, where an "attractive interaction" is defined as ε below some threshold value indicated by the purple hue (table S6). Inset shows zoom-in on the top 100 proteins. (**D**) Comparison of median heterotypic ε versus homotypic interaction ε for the top 100 proteins reveals some highly promiscuous proteins that are also predicted to interact strongly homotypically (homotypic ε < 0), while others are predicted to be obligatory heterotypic interactors (homotypic ε > 0). Gene names for a subset of the proteins are shown when graphically convenient. (**E**) Automatic domain decomposition of disordered regions in the yeast prion protein Sup35 based on homotypic intermaps. (**F**) Domain decomposition of a BRCA2 subregion using a subset of predefined chemically distinct chemical fingerprints.

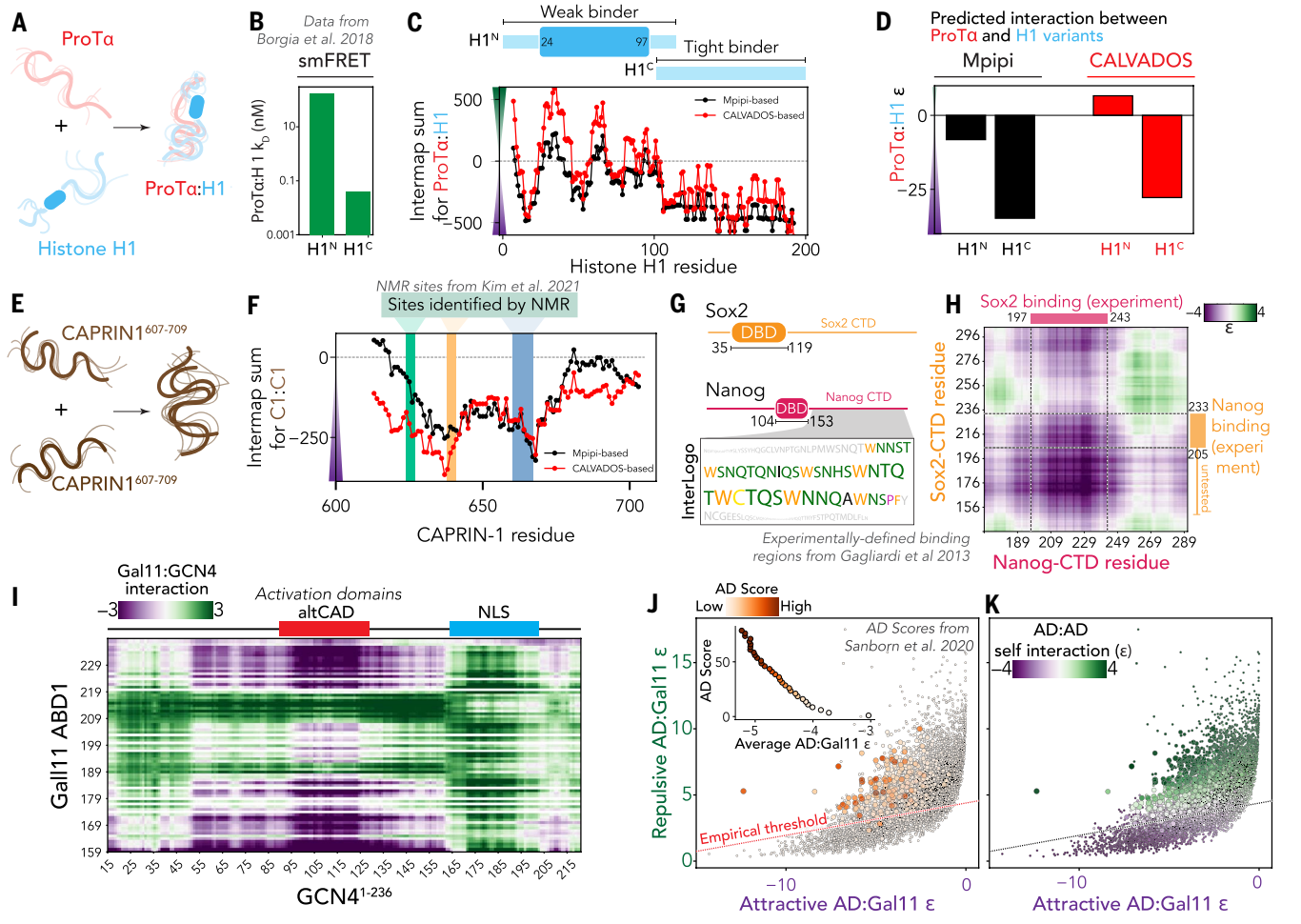


Fig. 5. IDR-mediated intermolecular interactions can be predicted from sequence. (**A**) Overview of the ProTα and H1 system. (**B**) Previously measured binding data for the N- and C-terminal halves of ProTα. These measurements were originally reported by Borgia *et al*. and are reproduced here for easy comparison with predictions (*30*). (**C**) Per-residue intermap sum for H1 with ProTα, where values come from summing over all rows in each H1 residue column from the intermap (fig. S12). More-negative values mean more attractive. (**D**) Sum of N- and C-terminal intermap sums to capture gross relative affinities of the two halves in both Mpipi and CALVADOS. (**E**) Overview of the CAPRIN-1$^{607–709}$ homotypic interaction. (**F**) A comparison of per-residue intermap sum for CAPRIN-1$^{607–709}$ homotypic interaction with NMR hotspots is highlighted. NMR hotspots were originally reported by Kim *et al*. and are reproduced here for easy comparison with predictions (*33*). (**G**) (Top) Domain schematic of transcription factors Sox2 and Nanog with DNA binding domains highlighted. The remainder of the proteins are disordered (fig. S14). (Bottom) InterLogo for Nanog C-terminal repeat domain (CTD) interaction with Sox2. Here, letter size is proportional to attractive interactions in the region with the most-attractive residues colored. The colored logo encompasses residues 201 to 240. (**H**) The intermap between Sox2 and Nanog CTDs, with previously identified binding region highlighted. Sox2$^{156–205}$ has not been tested as mediating Nanog interaction but is strongly predicted to be critical for this binding. In both cases, the specific residues that are predicted to mediate this interaction (Trp in Nanog and Tyr in Sox2; fig. S14) also underlie this interaction experimentally (*34*). (**I**) Intermap between GCN4 disordered region (*x* axis) and Gal11 ABD1 revealing which regions in GCN4 are predicted to interact with which surface regions on Gal11. Experimentally identified activation domains (*46*), which align well with the core region predicted to interact strongly with GCN4, are shown at the top. The experimentally verified nuclear localization signal (NLS) also emerges as a chemically distinct module within the IDR. (**J**) Scatterplot comparing attractive and repulsive ε values for activation domain (AD)–(Gal11$^{158–238}$) interaction for all tiles measured by Sanborn *et al*. (*40*), where marker size and color report on activation domain score. The empirical threshold line separates regions where pairs are predicted to have strongly attractive AD-Gal11 interaction, yet no strong activation domain tiles are reported. Inset shows a strong correlation between tiles with a strong AD score and tiles that interact favorably with Gal11$^{158–238}$. (**K**) Same points and sizes as in panel (J), with markers colored by homotypic ε value. Markers below the empirical threshold show strong homotypic interaction.

H1 and protamine, in line with experimental and computational work (fig. S12E) (*32*). These results illustrate that even for relatively low-complexity, high-affinity interactions, local chemical specificity is, in principle, predictable.

Another example of a dynamic interaction is the homotypic interaction between a fragment from the C-terminal IDR of the stress granule–associated protein CAPRIN-1 (Fig. 5E and fig. S13) (*33*). Recent NMR work characterized three interaction hotspots (residues 624–626, 638–640, and 660–666, highlighted as green, yellow, and blue, respectively) that contribute key interactions to CAPRIN-1 intermolecular behavior (Fig. 5F). In line with this, two of the three hotspots are clearly predicted from the sequence, with a smaller peak shown for the first. Moreover, unlike DDX4 (for which salt suppresses phase separation), CAPRIN-1 homotypic phase separation is enhanced at

higher salt concentrations, an effect also recapitulated by our approach (fig. S13). This result highlights how the complex interplay of charged and aromatic residues can be appropriately captured.

Finally, the master transcription factors Sox2 and Nanog have previously been shown to interact through their C-terminal IDRs (Fig. 5G) (*34*). We calculated the heterotypic intermap between these two C-terminal IDRs, identifying specific regions and residues predicted to drive intermolecular interaction (Fig. 5H). These predictions align well with previously identified binding regions and extend those regions beyond what was experimentally tested. To help visualize interacting regions, we mapped interaction strength directly onto the amino acid sequence (Fig. 5G,bottom). These interaction logos ("InterLogos") use the FINCHES-derived attractive interaction scores to define residue size, with the most strongly interacting residues colored by physicochemistry. Finally, FINCHES predicts that the C-terminal IDR of Nanog will undergo robust homotypic phase separation, in agreement with recently published work (fig. S14) (*35*). This result highlights how—given two proteins of interest—likely chemically specific sites of intermolecular interaction can be readily predicted.

Having examined several specific examples of IDR-IDR interaction, we next asked whether interactions between an IDR and a folded domain could be investigated. Transcription factors are DNA binding proteins typically consisting of a folded DNA binding domain and one or more intrinsically disordered regions (fig. S15A) (*36*, *37*). Among various functions, transcription factor IDRs can recruit coactivators (e.g., Med15, Gal11 in yeast), which, in turn, drive transcription (*38*–*40*). Recent work from several groups has reported detailed high-throughput studies to identify sequence features associated with so-called activation domains (ADs)—subregions within transcription factor IDRs that drive gene expression by recruiting coactivators (*40*–*46*).

We first investigated whether FINCHES could identify activation domains directly from sequence. To do this, we decomposed the surface residues on the Gal11 activation domain binding domain (ADB1; fig. S15B) to identify local residue communities (*47*). We then calculated an average interaction between each community of residues and a local window along the disordered region from the yeast transcription factor GCN4. This analysis identified specific subregions in the Gal11 IDR that show good overlap with the strongest experimentally determined AD (Fig. 5I).

We next investigated whether we could use this approach to interpret prior large-scale experiments relating amino acid sequence to activation domain function. We calculated average IDR–folded domain surface attractive and repulsive values for each of the 7577 40-residue tiles measured previously by Sanborn *et al.* in a high-throughput screen, relating the IDR-Gal11 interaction back to measured activation domain function (i.e., the ability to drive gene expression, quantified as an AD score) (*40*). This analysis shows an excellent correlation between the average IDR-Gal11 interaction and AD score; on average, sequences with a strong AD score show strong attractive interactions (Fig. 5J, inset). However, unexpectedly, when individual IDRs were decomposed into attractive and repulsive interactions with Gall11, we found many IDRs that were strongly attractive and weakly repulsive for Gal11 yet showed very poor AD scores (Fig. 5J). This apparent paradox was resolved by realizing that those IDRs are also predicted to engage in strong homotypic interaction (Fig. 5K). We interpret this result to mean that those IDRs may (i) undergo strong intramolecular interaction and/or (ii) undergo strong "off-target" intermolecular interaction (either through self-assembly or with other cellular components). In both cases, these "off-target" interactions may prevent Gal11 interaction, attenuating transcription (fig. S15D). These results are broadly consistent with data interpreted by the acidic exposure model proposed initially by Staller and Cohen and appear to offer a possible explanation for recent work exploring the link between condensate formation and transcriptional output (*43*, *48*). Overall, our work illustrates how FINCHES enables the molecular basis for IDR function to be rationally interpreted directly from sequence.

Discussion and conclusions

By excising the chemical physics developed for molecular simulations, we have repurposed the analytical forms and molecular parameters to describe inter–amino acid interaction and estimate IDR-associated attractive and repulsive interactions directly from the sequence. While it serves no one to overstate a method's efficacy, accuracy, and generalizability (see caveats discussed below), our approach is fast and simple and enables proteome-scale intermolecular interaction predictions.

FINCHES enables predictions of chemical specificity but does not account for sequence-specific (also known as site-specific) interactions. Sequence-specific interactions are driven by a defined structured interface. Consequently, these are—at least in principle—amenable to more conventional "structure-centric" approaches (e.g., AlphaFold, RoseTTAFold). While we emphasize that we do not expect to capture sequence-specific binding interfaces (e.g., short linear motifs), we see our approach as complementary to existing methods. FINCHES also does not encode mutual information between distal regions, such that the apparent valency between two IDRs is always maximal, regardless of whether, in principle, there should be a smaller number of mutually exclusive modes of interaction. Finally, the approach does not consider the intrinsic competition between intramolecular and intermolecular interactions, although this effect can be accounted for by considering homotypic and heterotypic interactions.

The predictions made by FINCHES rely on parameters obtained from coarse-grained molecular mechanics force fields—in this case, CALVADOS (CALVADOS2) and Mpipi (Mpipi-GG) (*6*–*9*). The two force fields have been well vetted but also, unavoidably, have limitations. As discussed previously, Mpipi-GG appears to underestimate aliphatic hydrophobic interactions, whereas both will fail to capture emergent effects driven by secondary structure. Similarly, the treatment of electrostatics through a Coulomb potential combined with a Debye-Hückel screening term will fail to recapitulate bona fide salt effects (*49*). Finally, the temperature dependence of the hydrophobic effect is absent from the functional forms of the force fields used here, and while hydrophobic and charged residues would drive lower critical solution temperature (LCST)–type attractive interactions at higher temperatures, this kind of behavior is not currently captured (*50*). That said, all of these limitations could, in principle, be parameterized and addressed by altering the underlying model. Moreover, we see the application of FINCHES-based predictions as a powerful route to identify force field shortcomings and even aid in reparameterization. Additionally, because of the underlying architecture of the FINCHES software, if and when new parameters become available (e.g., improved models, new residues, etc.), they can be immediately introduced and used.

Finally, we caution that the calculation of heterotypic ε values should not be in isolation inferred as predicting likely interactomes between IDRs in the cell. FINCHES predictions are facsimiles of in vitro experiments in which the only two proteins present are those considered in the calculation. However, specificity is determined by the combination of pairwise affinities of different partners and the availability (i.e., concentration) of those partners in binding-competent states (*51*). As such, FINCHES is most useful for taking two proteins known to interact and identifying the residues or regions that likely underlie that interaction. Those predictions can then readily be tested through mutagenesis.

One possible application of our approach is predicting homotypic phase diagrams directly from sequence. While one could envisage applying FINCHES to investigate the underlying molecular grammar associated with phase separation across disordered regions, prior and ongoing work from other groups has fairly exhaustively

explored the underlying chemical principles encoded by coarse-grained force fields (*6*, *8*, *52*–*55*). We build on this prior work, moving away from the need to extrapolate general principles to specific systems and instead enable direct predictions for how mutations are predicted to affect phase diagrams—at least qualitatively—on a case-by-case basis.

There has been recent interest in predicting phase separation from sequence using extant experimental data for training. A challenge for this type of machine learning–based predictor is a relative paucity of experimental data, compounded by the fact that those data are collected under various conditions (salt, temperature, pH, concentration, crowder, nucleic acid). This is more than just an inconvenience; phase separation is a solution-dependent and concentration-dependent phenomenon and is often driven by heterotypic interactions. As such, a binary classification for proteins or protein regions as phase separation competent (or not) misses the caveats of (i) at what concentration, (ii) under what conditions, and (iii) with what partner(s)?

Our work complements recent work by several groups in the context of disordered proteins and phase separation. Houston *et al.* illustrate how combining machine learning enables the prediction of intramolecular interactions from sequence (*56*). While focused on intramolecular interactions, recent work has shown how intramolecular interactions can be repurposed to investigate intermolecular interactions (*31*). Von Bülow *et al.* combined active learning with coarse-grained simulations to generate the requisite data to train a machine learning model to quantitatively predict homotypic phase diagram properties (i.e., concentrations in the dense and dilute phase) from sequence (*55*). Similarly, Adachi and Kawaguchi take an approach comparable to ours, illustrating how one can leverage molecular dynamics simulations to parameterize an effective interaction model, incorporating both pairwise amino acid interactions and sequence context effects, to estimate critical parameters such as the Boyle temperature and virial coefficients (*57*). Their work elegantly demonstrates how these ideas can be used to understand and even design sequences that undergo selective demixing for complex multiphase behaviors.

While much of the comparison in this manuscript is in the context of phase separation, looking forward, we anticipate that our approach will be most useful in four distinct areas unrelated to biomolecular phase separation. First, we see the most immediate impact of our approach in the guidance and interpretation of experiments examining IDR-mediated intermolecular interactions. We and others subscribe to an emerging model whereby IDR-mediated interactions are driven by a combination of sequence-specific and chemical-specific interactions (*1*, *4*, *51*, *58*). Our approach provides a means for anyone to easily and quickly quantify chemical specificity between an IDR and a partner. Importantly, because intermaps offer chemically specific insight into how an IDR may interact, they are effectively instantaneous hypothesis generators for understanding the mapping between IDR sequence and molecular interaction. Finally, applying our approach to investigate IDR–folded domain interactions suggests that this approach can be used to investigate folded domain surface chemistry, although this is an area of active work. We recently applied FINCHES to investigate folded domain surface chemistries to explore the determinants of desiccation resistance at proteome scale (*59*). Beyond binary interactions, our approach offers a route to aid in interpreting techniques that generate small interactomes (e.g., affinity purification mass spectrometry and proximity labeling).

Second, we are actively investigating the application of our approach to better understand IDR conservation and functional annotation. We and others have historically leaned heavily on sequence features (e.g., amino acid composition and patterning) to understand conservation in IDRs where primary structure is poorly conserved (*4*, *15*, *21*, *60*–*62*). In some cases, we anticipate that the conservation of sequence features reflects the preservation of intrinsic biophysical properties of an IDR (*9*, *27*, *63*). In other cases, we anticipate that the conservation of sequence features reflects the conservation of chemical specificity (*15*, *21*). Indeed, the classification of IDRs into distinct chemically similar clusters revealed many IDRs with very different sequences that share similar global interaction fingerprints (Fig. 5B). With the methods described here, we now have tools to predict both biophysical properties and chemical specificity directly from sequence, opening the door to new routes for the systematic assessment of conservation in IDRs through the lens of chemical physics (*9*).

Third, we see the ability to define context-dependent domains as a potentially important step toward generalized functional annotation in IDRs. In general, we consider protein domains to be defined by a function. In practice, the tight link between structure and function for folded proteins has enabled "structure" to act as a surrogate for function. As a result, there are many examples of conserved domains where—while function remains elusive—we are comfortable referring to them as domains because it is anticipated that they will have a stand-alone molecular function in some context. In many situations, an IDR's function is determined by its interactions with other cellular components. Moreover, a given IDR may interact with many different cellular partners, and those interactions may, in turn, be mediated by different sets of residues. Consequently, we propose that the definition of a domain in an IDR is, in many cases, unavoidably context dependent, where context here is defined by the partner of interest. Our approach here enables domains to be defined on the basis of intermolecular interaction profiles, allowing the domains to be defined in a context-dependent manner with respect to those putative partners.

Fourth, the throughput and generalizability of our approach lend themselves to the rational design of IDRs with desired intermolecular interaction properties. For example, identifying variants predicted to enhance or suppress phase separation and/or partitioning into an existing condensate with known components becomes trivial. Moreover, rationally designing IDRs that flank binding motifs to assess the role these flanking IDRs have in tuning affinity and specificity is straightforward, as is the design of IDRs that modulate intermolecular repulsion in the context of entropic force generation. In short, we foresee a range of design applications, with many of these methodologies becoming available shortly through our design package GOOSE (*64*).

As a final note, the impact and consequences of the various caveats raised here should always be considered when assessing whether a FINCHES prediction is valid or reasonable. With that caveat in mind, we see FINCHES as an effective way to obtain qualitative (and potentially semiquantitative) insight into how an IDR may interact with a partner, enabling hypothesis generation that can be precisely tested through precise mutagenesis and sequence design.

Materials and methods

The functional forms of nonbonded terms for Mpipi and CALVADOS force fields were reproduced and implemented in the FINCHES Python package (see supplementary materials). For Mpipi, the two components are a Wang-Frenkel and a Debye-Hückel potential (*6*, *65*, *66*). For CALVADOS, this is a shifted and truncated Ashbaugh-Hatch potential with a Debye-Hückel potential using an empirical correction for the temperature-dependence of electrostatic interactions (*7*, *67*, *68*). Force field parameters for CALVADOS2 and Mpipi-GG were taken from their respective publications (*7*, *9*).

Local charge effects were accounted for by considering the local $i+1$ and $i-1$ charge around a charged residue and down-weighting like-charged regions on the basis of local charge density, effectively reducing the repulsion associated with clusters of like-charged residues. Local hydrophobic effects were accounted for by considering contiguous runs of two, three, or more aliphatic residues, scaling up aliphatic-aliphatic attractive interactions by 1.5× and 3.0× for residues embedded within runs of two, three, or more aliphatic residues,

respectively. A more detailed discussion and technical implementation of these corrections are provided in the supplementary materials.

Phase diagrams were calculated using the analytical solution to the Flory-Huggins theory developed and implemented originally by Qian *et al.* (*14*). Solvent-accessible surface areas are calculated using MDTraj (*69*). IDR global dimensions in fig. S7 were predicted using ALBATROSS (*9*). Disorder prediction was calculated using metapredict V2-FF (*9*, *70*). Rational sequence designs used for examining homopolymer versus IDR properties were generated using GOOSE (*64*). Proteome-wide analysis was performed using SHEPHARD, with data obtained from UniProt (*71*, *72*). We make extensive use of previously published experimental data and are indebted to the authors for their previously published careful biophysical and biochemical studies. All sequences reported in this manuscript are defined in table S7.

FINCHES software implementation

The ability to predict ε-based intermolecular interactions is implemented in our software package FINCHES (First-principle INteractions via CHEmical Specificity). FINCHES not only continues our long-standing battle in the fight between acronyms and sanity but implements both Mpipi-GG and CALVADOS2 modules through a common interface that enables identical analysis and predictions. Moreover, the underlying architecture makes it straightforward to implement additional force fields in an object-oriented manner.

FINCHES is fully open-source and hosted at https://github.com/idptools/finches. To facilitate adoption and use, we present core analysis on our webserver (https://www.finches-online.com/). This enables standard analyses for proteins up to 1300 amino acids in length to be conducted in seconds. To facilitate further analysis, we also provide several Colab notebooks that enable standard types of analysis (intermaps, phase diagram prediction, etc.). These are linked from https://github.com/idptools/finches-colab.

FINCHES is implemented in Python (https://www.python.org/) (version 3.7 or higher) with Cython implementations for a subset of performance-sensitive algorithms (https://cython.org/). FINCHES also extensively uses NumPy, SciPy, and Matplotlib (*73*–*75*). Version control is provided through git (https://git-scm.com/). Bugs can be reported and features requested at https://github.com/idptools/finches/. Additional force fields can be implemented in the finches.forcefields module.

The finches-online webserver (https://finches-online.com/) is built using Flask (https://flask.palletsprojects.com/) and uses nginx (https://nginx.org/en/) and gunicorn (https://gunicorn.org/) for backend infrastructure. It provides front-facing services for intermap construction and phase diagram prediction.

Analysis

Analysis for this manuscript relies on metapredict, SHEPHARD, ALBATROSS, and sparrow (*9*, *70*, *72*, *76*). The figures and analysis associated with this manuscript are fully reproducible via Jupyter notebooks shared at https://github.com/holehouse-lab/supportingdata/tree/master/2025/finches_2025.

Calibration of charge prefactor

We calibrated the Mpipi-GG charge prefactor at 0.2 by tuning the self-interaction propensity of the Das-Pappu sequences (fig. S17) (*16*). Briefly, on the basis of single-chain compaction results alongside work from Lin and Chan, we anticipate sequences with large blocks of oppositely charged residues (i.e., high-kappa sequences) to have a negative (attractive) ε in the zero-salt limit (*17*). A charge prefactor of 0.2 defines the boundary between attractive and repulsive ε values for these sequences in a reasonable location, given previous theoretical predictions and simulation work (fig. S18A) (*17*, *77*, *78*). For CALVADOS2, the charge prefactor was calibrated to match the (magnitude-corrected) trends described by Mpipi-GG, giving a charge prefactor of 0.7 (fig. S18, B and C). This calibration is relatively empirical and qualitative. While a more robust calibration is certainly possible, we suggest that attempting to fit a more precise dependency here would overextend the reasonable predictive power of the underlying force field, especially given the layers of simplifying assumptions being made.

Comparison of Mpipi-GG and CALVADOS2-based predictions

We note that almost all results in this manuscript are reproduced by both CALVADOS and Mpipi-GG–based predictions. The singular exception to this is the impact of the conserved region (CR) in TDP-43 (fig. S6). For these, Mpipi-GG–based predictions do not correctly capture the hydrophobic nature of the aliphatic residues in the CR, whereas CALVADOS2 does a good job. To reiterate, neither model has any structural knowledge, so identifying this hotspot reflects the local density of aliphatic residues.

Mpipi-GG and CALVADOS2-based predictions are generally reported in their natural units, which place CALVADOS2 scores ~4× larger than Mpipi-GG scores. The exception to this is in Fig. 5, C and F, where Mpipi scores are multiplied by a scalar to better overlap with the CALVADOS2 scores; this is solely for visual purposes, in that numerical comparison of absolute values between the two models is not relevant or informative.

Sensitivity analysis

To assess model sensitivity, we calculated ε values for several sequences while varying the underlying force field parameters, allowing us to calculate the dependencies of the ε on each individual inter-residue pair and back-calculate sensitivity in terms of the absolute percentage change in ε compared with the percentage change applied to the underlying parameter. These analyses are shown in fig. S16. In all cases examined, a change of 1% in any of the underlying force field parameters has a <1% change in the resulting predictions, but there is a measurable change (typically up to ~0.5% for the largest effects). This suggests that the conclusions we arrive at do depend on the underlying force field parameters, but there is reasonable local robustness that, if we made seemingly small updates to the parameters, our conclusions would remain qualitatively similar. We also found that, as expected, the strongest dependence of a given sequence was scaled by the fraction of the amino acids present in that sequence.

Default intermap construction

Intermaps are calculated using a sliding window between all possible intermolecular fragments of a specified size. In general, for larger IDRs, we default to a size of 31 residues, although for smaller IDRs, this leads to a loss of fine-grain detail. For example, intermaps associated with the analysis in Fig. 5 use a 13-residue sliding window.

Clustering of IDRs in chemical space (human proteome)

Chemical clustering of natural IDRs (Fig. 4) was performed by calculating all possible heterotypic ε values between all 3413 IDRs in the human proteome that are between 100 and 150 residues in length. This consists of 11,648,569 individual ε calculations, generating the complete nonredundant matrix of intermolecular ε scores. Having generated this matrix, we use Ward's method for hierarchical clustering, followed by the optimal ordering of the resulting clusters to minimize the distance between adjacent clusters in chemical space. We arrived at 24 clusters as a reasonable number that provided chemically interpretable and distinct groups. Cluster assignment for each IDR is provided in table S4.

Clustering of IDRs in chemical space (chemical fingerprints)

Chemical clustering of all possible dipeptides (fig. S11) was performed by calculating all possible heterotypic ε values between all 210 nonredundant dipeptides. We used a repeat of 20 amino acids (10 repeats) for ε calculations. As before, having generated the matrix of all intermolecular ε values, we used Ward's method for hierarchical clustering,

followed by the optimal ordering of the resulting clusters to minimize the distance between adjacent clusters in chemical space. We specified 36 clusters for dipeptide fingerprint chemical decomposition by varying the number of clusters and observing when intuitively consistent chemical groups emerged. Full cluster memberships are provided in table S8.

Folded domain surface interactions

Folded domain–IDR interactions were determined using the following approach. (i) Surface residues were identified by taking a 3D Protein Data Bank (PDB) file and using SOURSOP and MDTraj to identify residues in which the solvent-accessible surface area (SASA) is above a threshold value (*69*, *79*). (ii) Once surface residues were identified, we constructed a meshwork graph whereby the over-surface nearest neighbors were identified, allowing us to calculate the distance between any given residues that approximates the distance via those surface residues, as opposed to the 3D geometric distance, which may bisect the protein structure. This shortest-path construction was achieved using Dijkstra's algorithm as implemented in NetworkX (*80*). (iii) Having established the surface residues and precalculated all inter-residue distances over the surface, we then defined each residue's local neighborhood by defining a maximum inter-residue distance below which residues are considered part of a local patch. Multiple residues can belong to the same patch, effectively defining a local chemical environment. IDR interactions were then computed as the average across all patches (as shown in Fig. 5) or using a sliding window that extracts local fragments from the IDR to construct a folded domain–IDR intermap where surface inaccessible residues are presented as empty vectors (as shown in fig. S15).

Null model construction

We provide a default null model in FINCHES by generating a large set of random shuffles and calculating the average intermap associated with these random shuffles. This makes it possible to normalize intermap interactions simply as a consequence of the composition versus those expected to drive attractive or repulsive interactions beyond merely the expectation of the sequence composition. Practically, this null model can be invoked by setting the null_model parameter in the Mpipi or CALVADOS front-end objects to the number of variants requested in the underlying prediction.

Identifying subdomains in IDRs

Intermaps were segmented to identify and extract subdomains with contiguous interaction chemistry. A thresholding mask was first applied to score columns in the intermap on the basis of the prevalence of the desired interaction characters. These scores were then analyzed using a peak-finding algorithm (SciPy), where boundaries between peaks were used to segment the array over one dimension (*74*). The same approach was then repeated for each subarray over the second dimension and again over the first dimension to remove off-target space. This triple segmentation approach was then repeated over the second dimension to account for bias, was subject to several user-defined parameters governing the masking threshold and peak identification characteristics, and provided an output list of rectangular regions defined as the residues covered in each of two input protein sequences.

Specifically, users can specify a precise interaction value upon which to threshold the scoring mask and whether to mask for values less than or greater than this value, parameters encoded as the arguments "criteria_threshold" and "criteria" respectively for the function "get_bidirectional_interaction_regions." Additionally, users can define a baseline coefficient of peak height at which the bounds of a given peak are calculated, encoded as the argument "baseline," and a minimum size for identified regions of interaction character (defined as the region's rectangular area) is encoded as the argument "min_region_size."

InterLogo construction

InterLogos are constructed by rescaling the per-residue attractive vector obtained from the summed intermap between 0 and 1 and rescaling each amino acid size between maximum and minimum font size. Then, residues in the top 70% in terms of attractiveness score are colored using the standard amino acid coloring (this is the default behavior; the fraction of colored residues can be defined by an input argument). InterLogo construction is implemented in FINCHES and can be done using the finches.frontened.interlogo module.

Phosphomimetic mutations

Experimentally verified phosphosites were taken from ProteomeScout and parsed using SHEPHARD as described previously (*72*, *81*). For every site, we ensured the corresponding position matched the expected unmodified residue (i.e., pSer = Ser, pThr = Thr, and pTry = Tyr) and converted every phosphosite to glutamic acid (E). This involved 130,606 phosphoserine sites, 53,744 phosphothreonine sites, and 38,106 phosphotyrosine sites. We note that a glutamic acid phosphomimetic likely does not fully capture the physical chemistry associated with phosphorylation, but previous work examining phosphomimetic mutations in the context of phase separation suggests that these are reasonable approximations, at least in some cases (*26*, *82*). This analysis was performed over all IDRs without length filtering.

Protein copy number analysis

Protein copy number information was taken from quantitative mass spectrometry analysis on HeLa cells performed by Hein *et al.* (*29*).

Gene Ontology analysis

GO analysis was performed using PANTHER-DB (*83*, *84*). Enrichment analysis was performed in terms of biological processes with *P* values computed on the basis of Fisher's exact test, and multiple hypothesis correction was done in terms of false discovery rate (default settings). This analysis was performed over proteins where one (or more) IDR was 100 residues or longer.

For Mpipi-GG and CALVADOS2 enrichment analysis, we compared proteins with an attractive ε value ($\varepsilon < 0$) with all proteins with an IDR of more than 100 residues. This is the appropriate background to compare against to answer the question, "For proteins with long IDRs, what types of biological processes are enriched for when considering those IDRs that are homotypically attractive?"

For WordCloud generation, we filtered GO hits with the following criteria: >40 different proteins must have been associated with the label, *P* value of less than 0.001, and enrichment of >1.25 versus the background. The WordCloud was generated using the "wordcloud" Python package (https://pypi.org/project/wordcloud/).

REFERENCES AND NOTES

1. A. S. Holehouse, B. B. Kragelund, The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* **25**, 187–211 (2024). doi: 10.1038/s41580-023-00673-0; pmid: 37957331
2. P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015). doi: 10.1038/nrm3920; pmid: 25531225
3. P. Tompa, M. Fuxreiter, Fuzzy complexes: Polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008). doi: 10.1016/j.tibs.2007.10.003; pmid: 18054235
4. I. Langstein-Skora *et al.*, Sequence- and chemical specificity define the functional landscape of intrinsically disordered regions. bioRxiv 2022.02.10.480018 [Preprint] (2024); https://doi.org/10.1101/2022.02.10.480018.
5. T. R. Alderson, I. Pritišanac, Đ. Kolarić, A. M. Moses, J. D. Forman-Kay, Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2304302120 (2023). doi: 10.1073/pnas.2304302120; pmid: 37878721
6. J. A. Joseph *et al.*, Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat. Comput. Sci.* **1**, 732–743 (2021). doi: 10.1038/s43588-021-00155-3; pmid: 35795820

7. G. Tesei, K. Lindorff-Larsen, Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res. Eur.* **2**, 94 (2023). doi: 10.12688/openreseurope.14967.1; pmid: 37645312
8. G. Tesei, T. K. Schulze, R. Crehuet, K. Lindorff-Larsen, 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**, e2111696118 (2021). doi: 10.1073/pnas.2111696118; pmid: 34716273
9. J. M. Lotthammer, G. M. Ginell, D. Griffith, R. J. Emenecker, A. S. Holehouse, Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods* **21**, 465–476 (2024). doi: 10.1038/s41592-023-02159-5; pmid: 38297184
10. P. R. Wills, D. J. Scott, D. J. Winzor, The osmotic second virial coefficient for protein self-interaction: Use and misuse to describe thermodynamic nonideality. *Anal. Biochem.* **490**, 55–65 (2015). doi: 10.1016/j.ab.2015.08.020; pmid: 26344712
11. Y. Lin, S. L. Currie, M. K. Rosen, Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. *J. Biol. Chem.* **292**, 19110–19120 (2017). doi: 10.1074/jbc.M117.800466; pmid: 28924037
12. M.-T. Wei *et al.*, Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nat. Chem.* **9**, 1118–1125 (2017). doi: 10.1038/nchem.2803; pmid: 29064502
13. M. Rubinstein, R. H. Colby, *Polymer Physics* (Oxford Univ. Press, 2003).doi: 10.1093/oso/9780198520597.001.0001
14. D. Qian, T. C. T. Michaels, T. P. J. Knowles, Analytical solution to the Flory-Huggins model. *J. Phys. Chem. Lett.* **13**, 7853–7860 (2022). doi: 10.1021/acs.jpclett.2c01986; pmid: 35977086
15. E. W. Martin *et al.*, Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694–699 (2020). doi: 10.1126/science.aaw8653; pmid: 32029630
16. R. K. Das, R. V. Pappu, 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). doi: 10.1073/pnas.1304749110; pmid: 23901099
17. Y.-H. Lin, H. S. Chan, Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophys. J.* **112**, 2043–2046 (2017). doi: 10.1016/j.bpj.2017.04.021; pmid: 28483149
18. C. W. Pak *et al.*, Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. *Mol. Cell* **63**, 72–85 (2016). doi: 10.1016/j.molcel.2016.05.042; pmid: 27392146
19. J. P. Brady *et al.*, Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E8194–E8203 (2017). doi: 10.1073/pnas.1706197114; pmid: 28894006
20. T. J. Nott *et al.*, Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57**, 936–947 (2015). doi: 10.1016/j.molcel.2015.01.013; pmid: 25747659
21. A. Bremer *et al.*, Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14**, 196–207 (2022). doi: 10.1038/s41557-021-00840-w; pmid: 34931046
22. M. Dzuricky, B. A. Rogers, A. Shahid, P. S. Cremer, A. Chilkoti, De novo engineering of intracellular condensates using artificial disordered proteins. *Nat. Chem.* **12**, 814–825 (2020). doi: 10.1038/s41557-020-0511-7; pmid: 32747754
23. B. S. Schuster *et al.*, Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11421–11431 (2020). doi: 10.1073/pnas.2000223117; pmid: 32393642
24. M. Linsenmeier *et al.*, The interface of condensates of the hnRNPA1 low-complexity domain promotes formation of amyloid fibrils. *Nat. Chem.* **15**, 1340–1349 (2023). doi: 10.1038/s41557-023-01289-9; pmid: 37749234
25. H. B. Schmidt, A. Barreau, R. Rohatgi, Phase separation-deficient TDP43 remains functional in splicing. *Nat. Commun.* **10**, 4890 (2019). doi: 10.1038/s41467-019-12740-2; pmid: 31653829
26. L. A. Gruijs da Silva *et al.*, Disease-linked TDP-43 hyperphosphorylation suppresses TDP-43 condensation and aggregation. *EMBO J.* **41**, e108443 (2022). doi: 10.15252/embj.2021108443; pmid: 35112738
27. G. Tesei *et al.*, Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024). doi: 10.1038/s41586-023-07004-5; pmid: 38297118
28. O. Beutel, R. Maraspini, K. Pombo-García, C. Martin-Lemaitre, A. Honigmann, Phase separation of zonula occludens proteins drives formation of tight junctions. *Cell* **179**, 923–936.e11 (2019). doi: 10.1016/j.cell.2019.10.011; pmid: 31675499
29. M. Y. Hein *et al.*, A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015). doi: 10.1016/j.cell.2015.09.053; pmid: 26496610
30. A. Borgia *et al.*, Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61–66 (2018). doi: 10.1038/nature25762; pmid: 29466338
31. B. Novak, J. M. Lotthammer, R. J. Emenecker, A. S. Holehouse, Accurate predictions of conformational ensembles of disordered proteins with STARLING. bioRxiv 2025.02.14.638373 [Preprint] (2025); https://doi.org/10.1101/2025.02.14.638373.
32. N. Galvanetto *et al.*, Mesoscale properties of biomolecular condensates emerging from protein chain dynamics. arXiv:2407.19202 [physics.bio-ph] (2024).
33. T. H. Kim *et al.*, Interaction hot spots for phase separation revealed by NMR studies of a CAPRIN1 condensed phase. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2104897118 (2021). doi: 10.1073/pnas.2104897118; pmid: 34074792
34. A. Gagliardi *et al.*, A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J.* **32**, 2231–2247 (2013). doi: 10.1038/emboj.2013.161; pmid: 23892456
35. K.-J. Choi *et al.*, NANOG prion-like assembly mediates DNA bridging to facilitate chromatin reorganization and activation of pluripotency. *Nat. Cell Biol.* **24**, 737–747 (2022). doi: 10.1038/s41556-022-00896-x; pmid: 35484250
36. S. A. Lambert *et al.*, The human transcription factors. *Cell* **172**, 650–665 (2018). doi: 10.1016/j.cell.2018.01.029; pmid: 29425488
37. M. V. Staller, Transcription factors perform a 2-step search of the nucleus. *Genetics* **222**, iyac111 (2022). doi: 10.1093/genetics/iyac111; pmid: 35939561
38. I. Jedidi *et al.*, Activator Gcn4 employs multiple segments of Med15/Gal11, including the KIX domain, to recruit mediator to target genes *in vivo*. *J. Biol. Chem.* **285**, 2438–2455 (2010). doi: 10.1074/jbc.M109.071589; pmid: 19940160
39. J. K. Thakur *et al.*, Mediator subunit Gal11p/MED15 is required for fatty acid-dependent gene activation by yeast transcription factor Oaf1p. *J. Biol. Chem.* **284**, 4422–4428 (2009). doi: 10.1074/jbc.M808263200; pmid: 19056732
40. A. L. Sanborn *et al.*, Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *eLife* **10**, e68068 (2021). doi: 10.7554/eLife.68068; pmid: 33904398
41. M. V. Staller *et al.*, Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Syst.* **13**, 334–345.e5 (2022). doi: 10.1016/j.cels.2022.01.002; pmid: 35120642
42. A. Udupa, S. R. Kotha, M. V. Staller, Commonly asked questions about transcriptional activation domains. *Curr. Opin. Struct. Biol.* **84**, 102732 (2024). doi: 10.1016/j.sbi.2023.102732; pmid: 38056064
43. M. V. Staller *et al.*, A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell Syst.* **6**, 444–455.e6 (2018). doi: 10.1016/j.cels.2018.01.015; pmid: 29525204
44. A. Erijman *et al.*, A high-throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning. *Mol. Cell* **79**, 1066 (2020). doi: 10.1016/j.molcel.2020.08.013; pmid: 32946759
45. C. N. J. Ravarani *et al.*, High-throughput discovery of functional disordered regions: Investigation of transactivation domains. *Mol. Syst. Biol.* **14**, e8190 (2018). doi: 10.15252/msb.20188190; pmid: 29759983
46. C. J. LeBlanc *et al.*, Conservation of function without conservation of amino acid sequence in intrinsically disordered transcriptional activation domains. bioRxiv 2024.12.03.626510 [Preprint] (2024); https://doi.org/10.1101/2024.12.03.626510.
47. P. S. Brzovic *et al.*, The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. *Mol. Cell* **44**, 942–953 (2011). doi: 10.1016/j.molcel.2011.11.008; pmid: 22195967
48. A. Bremer *et al.*, Reconciling competing models on the roles of condensates and soluble complexes in transcription factor function. bioRxiv 2024.11.21.624739 [Preprint] (2024); https://doi.org/10.1101/2024.11.21.624739.
49. H. I. Okur *et al.*, Beyond the Hofmeister series: Ion-specific effects on proteins and their biological functions. *J. Phys. Chem. B* **121**, 1997–2014 (2017). doi: 10.1021/acs.jpcb.6b10797; pmid: 28094985
50. G. L. Dignon, W. Zheng, Y. C. Kim, J. Mittal, Temperature-controlled liquid–liquid phase separation of disordered proteins. *ACS Cent. Sci.* **5**, 821–830 (2019). doi: 10.1021/acscentsci.9b00102; pmid: 31139718
51. K. Teilum, J. G. Olsen, B. B. Kragelund, On the specificity of protein-protein interactions in the context of disorder. *Biochem. J.* **478**, 2035–2050 (2021). doi: 10.1042/BCJ20200828; pmid: 34101805
52. M. J. Maristany *et al.*, Decoding phase separation of prion-like domains through data-driven scaling laws. *eLife* **13**, RP99068 (2025). doi: 10.7554/eLife.99068; pmid: 39937084
53. S. Rekhi *et al.*, Expanding the molecular language of protein liquid–liquid phase separation. *Nat. Chem.* **16**, 1113–1124 (2024). doi: 10.1038/s41557-024-01489-x; pmid: 38553587
54. S. Das, Y.-H. Lin, R. M. Vernon, J. D. Forman-Kay, H. S. Chan, Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 28795–28805 (2020). doi: 10.1073/pnas.2008122117; pmid: 33139563
55. S. von Bülow, G. Tesei, F. K. Zaidi, T. Mittag, K. Lindorff-Larsen, Prediction of phase-separation propensities of disordered proteins from sequence. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2417920122 (2025). doi: 10.1073/pnas.2417920122; pmid: 40131954
56. L. Houston, M. Phillips, A. Torres, K. Gaalswyk, K. Ghosh, Physics-based machine learning trains Hamiltonians and decodes the sequence-conformation relation in the disordered proteome. *J. Chem. Theory Comput.* **20**, 10266–10274 (2024). doi: 10.1021/acs.jctc.4c01114; pmid: 39504303
57. K. Adachi, K. Kawaguchi, Predicting heteropolymer interactions: Demixing and hypermixing of disordered protein sequences. *Phys. Rev. X* **14**, 031011 (2024). doi: 10.1103/PhysRevX.14.031011

58. K. Bugge *et al.*, Interactions by disorder – a matter of context. *Front. Mol. Biosci.* **7**, 110 (2020). doi: 10.3389/fmolb.2020.00110; pmid: 32613009
59. P. S. Romero-Pérez *et al.*, Protein surface chemistry encodes an adaptive tolerance to desiccation. bioRxiv 2024.07.28.604841 [Preprint] (2024); https://doi.org/10.1101/2024.07.28.604841.
60. I. Pritišanac *et al.*, A functional map of the human intrinsically disordered proteome. bioRxiv 2024.03.15.585291 [Preprint] (2024); https://doi.org/10.1101/2024.03.15.585291.
61. T. Zarin *et al.*, Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *eLife* **10**, e60220 (2021). doi: 10.7554/eLife.60220; pmid: 33616531
62. M. K. Shinn *et al.*, Connecting sequence features within the disordered C-terminal linker of *Bacillus subtilis* FtsZ to functions and bacterial cell division. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2211178119 (2022). doi: 10.1073/pnas.2211178119; pmid: 36215496
63. N. S. González-Foutel *et al.*, Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat. Struct. Mol. Biol.* **29**, 781–790 (2022). doi: 10.1038/s41594-022-00811-w; pmid: 35948766
64. R. J. Emenecker, K. Guadalupe, N. M. Shamoon, S. Sukenik, A. S. Holehouse, Sequence-ensemble-function relationships for disordered proteins in live cells. bioRxiv 2023.10.29.564547 [Preprint] (2023); https://doi.org/10.1101/2023.10.29.564547.
65. X. Wang, S. Ramírez-Hinestrosa, J. Dobnikar, D. Frenkel, The Lennard-Jones potential: When (not) to use it. *Phys. Chem. Chem. Phys.* **22**, 10624–10633 (2020). doi: 10.1039/C9CP05445F; pmid: 31681941
66. P. W. Atkins, J. de Paula, J. Keeler, *Atkins' Physical Chemistry* (Oxford Univ. Press, 2023).
67. H. S. Ashbaugh, H. W. Hatch, Natively unfolded protein stability as a coil-to-globule transition in charge/hydropathy space. *J. Am. Chem. Soc.* **130**, 9536–9542 (2008). doi: 10.1021/ja802124e; pmid: 18576630
68. G. C. Akerlof, H. I. Oshry, The dielectric constant of water at high temperatures and in equilibrium with its vapor. *J. Am. Chem. Soc.* **72**, 2844–2847 (1950). doi: 10.1021/ja01163a006
69. R. T. McGibbon *et al.*, MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015). doi: 10.1016/j.bpj.2015.08.015; pmid: 26488642
70. R. J. Emenecker, D. Griffith, A. S. Holehouse, Metapredict: A fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* **120**, 4312–4319 (2021). doi: 10.1016/j.bpj.2021.08.039; pmid: 34480923
71. UniProt Consortium, UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023). doi: 10.1093/nar/gkac1052; pmid: 36408920
72. G. M. Ginell, A. J. Flynn, A. S. Holehouse, SHEPHARD: A modular and extensible software architecture for analyzing and annotating large protein datasets. *Bioinformatics* **39**, btad488 (2023). doi: 10.1093/bioinformatics/btad488; pmid: 37540173
73. C. R. Harris *et al.*, Array programming with NumPy. *Nature* **585**, 357–362 (2020). doi: 10.1038/s41586-020-2649-2; pmid: 32939066
74. P. Virtanen *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020). doi: 10.1038/s41592-019-0686-2; pmid: 32015543
75. J. D. Hunter, Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007). doi: 10.1109/MCSE.2007.55
76. A. S. Holehouse, sparrow: a tool for integrative analysis and prediction from protein sequence data, version 0.1, Zenodo (2022); https://doi.org/10.5281/zenodo.6891920.
77. S. Das, A. Eisen, Y.-H. Lin, H. S. Chan, A lattice model of charge-pattern-dependent polyampholyte phase separation. *J. Phys. Chem. B* **122**, 5418–5431 (2018). doi: 10.1021/acs.jpcb.7b11723; pmid: 29397728
78. J. McCarty, K. T. Delaney, S. P. O. Danielsen, G. H. Fredrickson, J.-E. Shea, Complete phase diagram for liquid-liquid phase separation of intrinsically disordered proteins. *J. Phys. Chem. Lett.* **10**, 1644–1652 (2019). doi: 10.1021/acs.jpclett.9b00099; pmid: 30873835
79. J. M. Lalmansingh, A. T. Keeley, K. M. Ruff, R. V. Pappu, A. S. Holehouse, SOURSOP: A python package for the analysis of simulations of intrinsically disordered proteins. *J. Chem. Theory Comput.* **19**, 5609–5620 (2023). doi: 10.1021/acs.jctc.3c00190; pmid: 37463458
80. A. A. Hagberg, D. A. Schult, P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," *Proceedings of the 7th Python in Science Conference (SciPy 2008)* (Los Alamos National Laboratory, 2008), pp. 11–15.
81. M. K. Matlock, A. S. Holehouse, K. M. Naegle, ProteomeScout: A repository and analysis resource for post-translational modifications and proteins. *Nucleic Acids Res.* **43**, D521–D530 (2015). doi: 10.1093/nar/gku1154; pmid: 25414335
82. Z. Monahan *et al.*, Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *EMBO J.* **36**, 2951–2967 (2017). doi: 10.15252/embj.201696394; pmid: 28790177
83. S. Carbon, C. Mungall, Gene Ontology Data Archive (2024-01-17) [Data set], Zenodo (2024); https://doi.org/10.5281/zenodo.10536401.
84. P. D. Thomas *et al.*, PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022). doi: 10.1002/pro.4218; pmid: 34717010
85. G. Ginell *et al.*, Supporting data for: Sequence-based prediction of intermolecular interactions driven by disordered regions, Zenodo (2025); https://doi.org/10.5281/zenodo.15005027.

ACKNOWLEDGMENTS

We thank S. Sukenik, B. Schmidt, M. Staller, and members of the Holehouse lab for feedback and discussion. We thank K. Lindorff-Larsen and S. von Bülow for being willing to—once again—coordinate preprint submission. We are indebted to the various groups who provided high-quality data with which we could directly test predictions. We also extend a warm thank you to J. Joseph, A. Reinhardt, R. Collepardo-Guevara, G. Tesei, and K. Lindorff-Larsen for their prior and ongoing work on the Mpipi and CALVADOS force fields. Additionally, we thank members of the Water and Life Interface Institute (WALII; DBI grant 2213983) for helpful discussion. Finally, we thank members of the Center for Biomolecular Condensates, supported by the Department of Biomedical Engineering at Washington University in St. Louis, for helpful discussion and feedback. **Funding:** A.S.H. was funded by the National Institutes of Health (NIH) through grant DP2-CA290639. J.M.L. was funded by the National Science Foundation Graduate Research Fellowship Program (DGE-2139839). E.T.U. was funded by a Keck Postdoctoral Fellowship. G.M.G. was funded by a MilliporeSigma Fellowship. J.F.P. was funded by a Washington University in St. Louis Department of Biochemistry and Molecular Biophysics Cori Fellowship **Author contributions:** Conceptualization: G.M.G., R.J.E., A.S.H.; Methodology: G.M.G., A.S.H., A.T.K., N.R., S.P.P., R.J.E.; Investigation: G.M.G., A.S.H., E.T.U., J.M.L., R.J.E.; Funding acquisition: A.S.H., G.M.G., E.T.U., J.M.L.; Project administration: A.S.H.; Supervision: A.S.H., J.F.P.; Writing – original draft: A.S.H.; Writing – review & editing: A.S.H., G.M.G., A.T.K., N.R., R.J.E., J.F.P., S.P.P. **Competing interests:** A.S.H. is on the scientific advisory board of Prose Foods. All other authors declare that they have no competing interests. **Data and materials availability:** All data and code associated with this manuscript are deposited as a single snapshot in Zenodo (*85*). In addition, all data and code in this work are provided at either https://github.com/idptools/finches or https://github.com/holehouse-lab/supportingdata/tree/master/2025/finches_2025. FINCHES is implemented as a web server at https://www.finches-online.com/.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adq8381
Supplementary Text; Figs. S1 to S19; Tables S1 to S8; References (*86*–*96*);
MDAR Reproducibility Checklist

Submitted 3 June 2024; accepted 11 March 2025

10.1126/science.adq8381