PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

A functional map of the human intrinsically disordered proteome

Iva Pritišanac[a,b,c,d,1], T. Reid Alderson[e,f], Đesika Kolarić[e,f], Taraneh Zarin[a,2], Shuting Xie[a], Alex Lu[a,g,3], Aqsa Alam[a], Abdullah Maqsood[b], Ji-Young Youn[b,h], Julie D. Forman-Kay[b,i,1], and Alan M. Moses[a,g,1]

Affiliations are included on p. 11.

Edited by Susan Marqusee, University of California Berkeley, Berkeley, CA; received February 18, 2026; accepted March 28, 2026

Intrinsically disordered regions (IDRs) represent at least one-third of the human proteome and defy the established structure–function paradigm. Because IDRs often have limited positional sequence conservation, the functional classification of IDRs using standard bioinformatics is generally not possible. Here, we show that evolutionarily conserved molecular features of IDRs enable clustering of the human disordered proteome (IDRome) into a map with strong functional enrichments. We quantify how conserved IDR features correlate with functional terms and, for a subset of terms, provide proteome-wide predictions of annotations for IDRs. Further, we show that conserved features of IDRs can predict protein localization to different biomolecular condensates and underlie elevated intracluster connectivity in condensate-associated IDRs, as well as enrich for short-linear motif-binding domains among interaction partners. We highlight patterns of conservation in disordered proteins with unknown function and in clusters enriched for proteins encoded by disease-risk genes. Our map of the human IDR-ome should be a valuable resource that aids in the discovery of new IDR biology.

intrinsically disordered proteins | molecular features | biomolecular condensates | interaction networks | protein functional prediction

Significance

Much of the human proteome lacks stable structure and consists of intrinsically disordered regions (IDRs). IDRs have key roles in cellular signaling, gene expression, and cellular organization, but their rapid sequence evolution has made them notoriously difficult to study using standard tools. This work offers insights into human disordered proteins by focusing on conserved bulk molecular features of the sequence rather than positional sequence conservation. By mapping these features across thousands of human IDRs, the study reveals which conserved aspects of disorder contribute to specific protein functions or interaction networks and which are associated with disease-risk genes. This resource charts the hidden logic of information encoded in disorder, a long-standing frontier in proteome science.

The sequence–structure–function paradigm in molecular biology postulates that the amino acid sequence of a protein encodes its three-dimensional structure, which determines the function of the protein. The close relationships between sequence, structure, and function are routinely exploited to infer function from sequence or structural data (1–5), trace the evolutionary history of protein–protein interactions (6), design de novo proteins with desired folds or functions (7–9), and predict the pathogenicity of sequence variants in the human genome (10–13). Indeed, structural information recovered from amino acid sequence alignments is central to state-of-the-art protein structure prediction methods (14, 15). However, the sequence–structure–function paradigm does not apply to the approximately one-third of residues in the human proteome that map to intrinsically disordered regions (IDRs), which lack stable secondary and tertiary structure and exhibit poor positional sequence conservation (16–18). Despite their lack of ordered structural elements, IDRs function in key cellular processes (19) and frequently act as hubs in protein–protein interaction networks (20), often via transient, multivalent interactions that promote phase separation and involvement in biomolecular condensates (21).

While the presence of IDRs in proteins can generally be predicted with high accuracy from their amino acid sequences (22, 23), the relationship between the sequences and biological functions of IDRs, although understood in many cases, is not understood in general (24–32). Segments of IDR sequences that show strong similarity in sequence alignments (which we refer to as "positional conservation") often point to so-called short-linear motifs (SLiMs) and Molecular Recognition Features (MoRFs) (20, 33–36). However, positionally conserved elements typically constitute only a minor fraction of an IDR sequence, and many of the experimentally characterized SLiMs are not positionally conserved (35, 37–39). More recently, we and others showed that approximately 15% of human IDRs contain significant positional alignment due to the acquisition of a conditional fold in particular functional contexts (40, 41). Nevertheless, it is appreciated that the majority of positions in the sequences of IDRs appear to evolve more rapidly relative to ordered regions in the same proteins (39, 42). Rapid evolution in IDRs reflects the absence of stable folded structure, since positional conservation is directly linked to evolutionary pressure to maintain a three-dimensional fold (25). Thus, because IDRs exhibit limited positional conservation in multiple sequence alignments, these alignments provide limited insight into the functional roles of IDRs (25–27). For intrinsically disordered proteins (IDPs), which are fully disordered and make up ~5% of the human proteome (*ca.* 1,000 proteins) (43), predictions of function are even more limited due to the lack of any folded domains (24).

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: iva.pritisanac@helmholtz-munich.de, forman@sickkids.ca, or alan.moses@utoronto.ca.

[2]Present address: Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona 08003, Spain.

[3]Present address: Microsoft Research New England, Cambridge, MA 02139.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2604562123/-/DCSupplemental.

Published May 18, 2026.

The functional importance of IDRs and IDPs is increasingly appreciated, especially in the context of biomolecular condensates (44–49). More generally, the biological functions of IDRs often relate to protein localization (both subcellular and extracellular), cell signaling interactions modulated by posttranslational modifications, and other aspects of protein regulation. IDR-containing proteins are often dysregulated in diseases such as cancer, amyotrophic lateral sclerosis, and other neurological disorders (43, 44, 50), with a recently increased focus on disease-associated sequence variants that map to IDRs (48, 51, 52). Interpreting the effects of mutations in IDRs remains challenging, as most variant effect predictors compute the impacts on fold stability and other structural features, e.g., changes to enzyme active sites or interfaces (53). Indeed, recent reports show that prediction methods are typically less accurate for disease variants in IDRs compared to folded domains (54). Thus, an understanding of how the sequences of human IDRs relate to biological function is urgently needed.

Several recent efforts aim to predict biological function of IDRs without relying on multiple sequence alignments (26, 27, 55–62). IDRs generally show strong evolutionary conservation of sequence-derived molecular features that are not positionally constrained (26, 27, 57, 60, 63–65). In a series of earlier studies, we showed how evolutionary properties of bulk molecular features that are computable from IDR sequences can be used to cluster and classify yeast IDRs into an unexpectedly large number of functional groups (26, 27).

Building on this framework, we show here that human IDRs are amenable to systematic classification based on evolutionary conservation of a broad set of bulk molecular features. We provide a comprehensive functional map of IDRs within the human proteome (IDR-ome). We obtain estimates for the proportion of human IDRs correlated with a broad range of Gene Ontology (GO) terms, and we train classifiers to predict association with those terms for unannotated IDPs and IDRs from sequence alone. Since the functional map is based on evolutionary conservation of simple molecular features, we can determine which features are associated with different groups of IDRs, such as those involved in the formation of biomolecular condensates or those associated with disease-risk genes. We demonstrate how combinations of conserved molecular features correlate with diverse biological roles and localizations of IDRs within the context of full-length proteins. Our map of the IDR-ome, GO-based classification of IDRs, and predictions of association with cellular localizations and disease-risk represent a critical resource for understanding IDR biology. Our map is readily searchable and retrievable to support validation and hypothesis-driven studies of the molecular basis of IDR function and dysfunction.

Results

Intrinsic Disorder Is Abundant in Human Proteins. To build a global functional map of the human IDRs, we first identified the boundaries of IDRs in the human proteome using the SPOT-Disorder (SPOTD) predictor (66) followed by filtering to retain 21,252 sequences (43) (*SI Appendix*, *SI Methods*). Intrinsic disorder is abundant in the proteome (22), with IDRs found in approximately 60% of proteins and exhibiting diverse length distributions (median 74 residues) and positional contexts (*SI Appendix*, Fig. S1 *A*–*C*). Predominantly disordered proteins, or those that contain more than 50% disordered residues, amount to nearly 20% of the proteome, whereas entirely disordered proteins (i.e., IDPs) account for 5% (*SI Appendix*, Fig. S1*A*). The average IDR in our dataset is approximately 200 residues long, and 280 IDRs/IDPs have more than 1,000 consecutively disordered residues (*SI Appendix*, Fig. S1*B*). We confirmed that, as expected, the predicted IDRs show generally lower levels of positional sequence-similarity in alignments of homologs from the Ensembl database (67) as compared to folded protein domains (*SI Appendix*, Fig. S1*D*). Moreover, most human IDRs are not easily assigned to protein families using sequence alignments, with only 21% of human IDRs showing significant sequence similarity (BLAST e-value < 1e-6) to any other human IDR (*SI Appendix*, Fig. S2).

A Global Map of the Human IDR-ome Based on Evolutionary Conservation of Molecular Features. Next, we estimated the evolutionary conservation of molecular features in the human IDR-ome without relying on conventional multiple sequence alignments. To this end, we first compiled a comprehensive list of 144 bulk molecular features that have been shown to be important for function of IDRs in different studies, including known short linear interaction motifs (SLiMs), physicochemical properties (e.g., hydrophobicity, polarity, charge, charge patterning), residue composition, and (homo-)repeats (*SI Appendix*, *SI Methods* and Dataset S1) (35, 68–75). We focus our analysis on these bulk, sequence-derived properties of IDRs, as we specifically aim to understand how IDR properties that correlate with protein function are encoded in the IDR amino acid sequence. We then developed a computational framework to estimate the evolutionary conservation of molecular features (*SI Appendix*, *SI Methods* and Figs. S3 and S4). While our method is conceptually based on previous work (26), our approach demanded significant reengineering and methods development to address the complexity and scale of the human IDR-ome, as detailed in the Methods (*SI Appendix*, Figs. S5 and S6). We use standard Z-scores to compare the observed distributions of molecular features in homologous IDR sequences to those expected from simulations of IDR sequences under a null hypothesis, which assumes no evolutionary constraints on molecular features (*SI Appendix*, Figs. S3 and S4). We refer to a set of Z-scores for all molecular features as an evolutionary signature of an IDR, which represents the pattern of conserved molecular features. Positive Z-scores indicate a feature value greater than expected under the null hypothesis (i.e., the simulations), while negative Z-scores indicate a feature value smaller than expected (*SI Appendix*, Figs. S3 and S4). Negative Z-scores can suggest either depletion of a feature (e.g., selection against hydrophobic residues) or a strongly negative value (e.g., selection for a net charge far below the expectation).

After filtering the human IDR-ome, we computed evolutionary signatures for 19,032 IDRs (*SI Appendix*, *SI Methods*) and performed hierarchical clustering to identify groups of IDRs that share patterns of conservation (Fig. 1*A*) (76). In this global map of the IDR-ome (Fig. 1*A* and *SI Appendix*, Fig. S7), IDRs that have similar evolutionary signatures are placed closer to one another. To test the importance of evolutionary conservation, we also generated a "features-only" signature (FS) representation in which we computed and normalized features of the human IDR sequences only, without any considerations of orthologous IDR sequences and evolutionary conservation of features (*SI Appendix*, Fig. S8). Clustering this features-only representation produced a sparser and more fragmented map than the map based on evolutionary conservation (*SI Appendix*, Fig. S8). We hypothesize that similarity in the evolutionary patterns of molecular features is analogous to sequence similarity detected in alignments for folded protein regions [e.g., by using PSI-BLAST (77)]. Our IDR-ome map reveals many clusters of IDRs, which are defined by distinct patterns of conserved molecular features, i.e., evolutionary signatures (Fig. 1 and *SI Appendix*, Fig. S7).

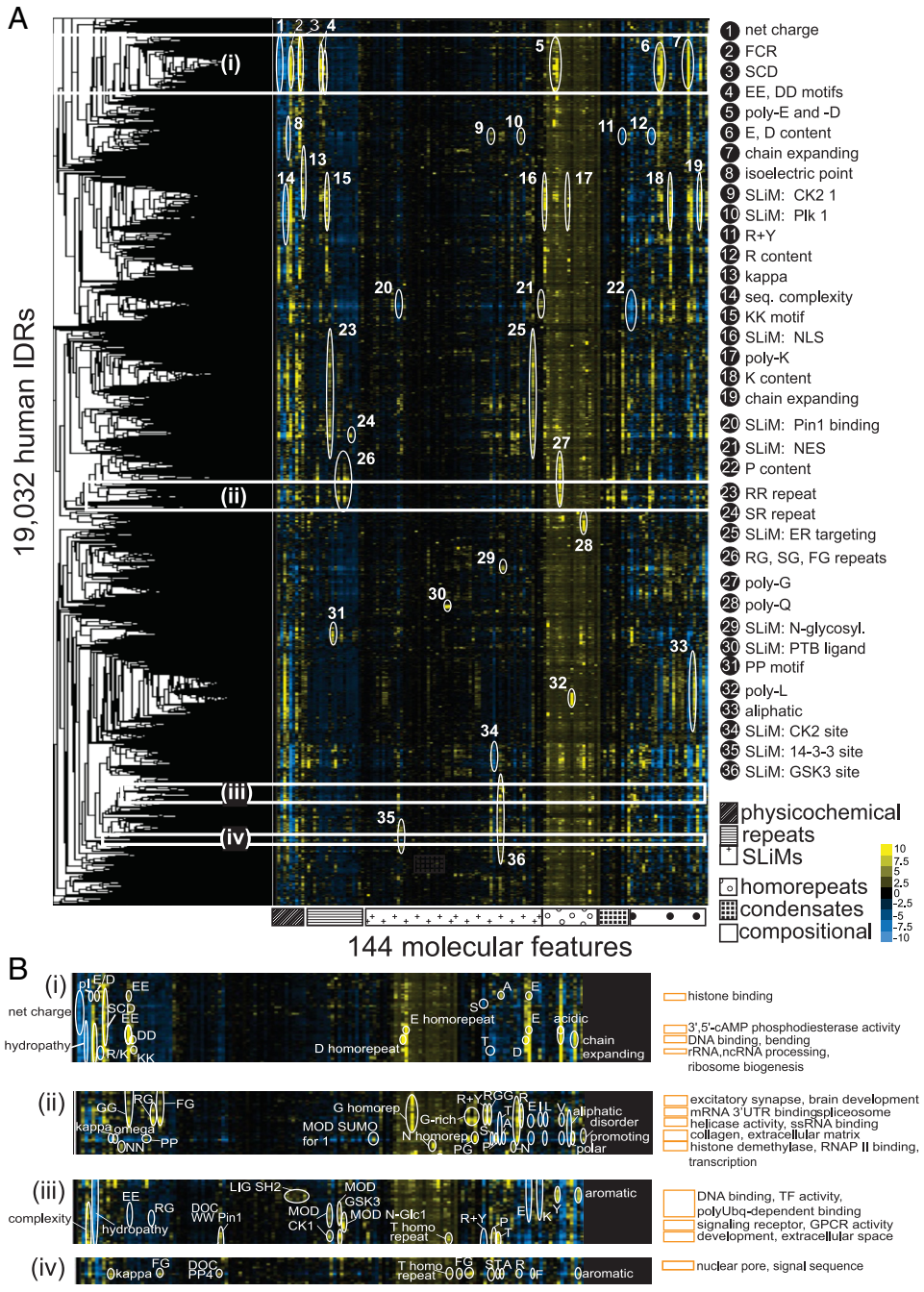

Fig. 1. A global map of human IDRs obtained through clustering of evolutionary signatures. (*A*) Hierarchical clustering of 19,032 human IDRs (y-axis) based on the evolutionary conservation of 144 different molecular features (x-axis). The molecular features are grouped into six different categories (physicochemical, repeats, SLiMs, homorepeats, condensates, or compositional biases). This global map of the human IDR-ome shows conservation Z-scores, with some of the dominant molecular features annotated with white circles and numbers, described in the legend (*Right*). A positive or negative Z-score, respectively, is defined by a higher or lower value of a mean of a molecular feature over orthologous IDRs than expected based on a simulation of an absence of evolutionary conservation. White rectangles indicate areas of selected clusters featured in panel *B*. (*B*) Clusters that are defined by strong patterns of Z-scores often contain a statistically significant overrepresentation of GO-term molecular functions, biological processes, and/or subcellular localizations, as listed here for select examples in areas (*i*), (*ii*), (*iii*), and (*iv*) from the panel *A*. A detailed view of statistically overrepresented terms and features for the rest of the map are available in *SI Appendix*, Fig. S7. Complete information on statistics of functional overrepresentation associated with each cluster selected manually or extracted using an automatic protocol are available in Dataset S2.

Assigning Gene Ontology to Clusters of IDRs. To test if the patterns of conservation of molecular features are associated with specific biological functions of proteins, we performed standard enrichment analysis for the GO annotations on the proteins containing IDRs that cluster together on the map (*SI Appendix*, *SI Methods*). First, we manually selected 93 clusters from the map, focusing on patterns of Z-score signals. Among these clusters, 53 (i.e., 57% of the clusters) exhibited overrepresentation of at least one GO term. The 53 clusters amounted to 9,294 IDRs (i.e., 49% of the human IDR-ome), and represent diverse GO terms (Dataset S2 and *SI Appendix*, Fig. S7). In contrast, in the randomized GO-assignment control, only 16 clusters (4%) showed any GO term overrepresentation. To confirm that the widespread overrepresentation of GO terms was not due to bias in manual selection of clusters, we repeated the analysis using automatically defined clusters and found qualitatively similar overrepresentations to those observed in the manually identified clusters (Datasets S2 and S3 and *SI Appendix*, Figs. S9 and S10 and *SI Methods*). This analysis revealed that evolutionary conserved IDR properties are consistently found in proteins associated with specific GO annotations, indicating proteome-wide association between IDR feature patterns and the biological roles of their host proteins. We also tested whether features obtained from human sequences alone were associated with GO terms and found broadly similar fractions of GO-term enrichments (*SI Appendix*, Fig. S8). This is consistent with recent studies that find human IDR feature analysis is sufficient to classify many types of IDRs (78, 79) (*Discussion*).

To obtain a global picture of the types of biological functions that human proteins bearing IDRs with specific properties have, we defined around 20 broader categories that, with some overlap, cover most overrepresented GO terms linked to various clusters

(*SI Appendix*, Fig. S9 and Dataset S2). We defined those categories by grouping related overrepresented GO terms, as detailed in the *SI Appendix*, *SI Methods*. Some of the most populated clusters are associated with DNA binding (23%), Chromatin/Chromatin binding (33%), RNA metabolism (22%), Cytoskeleton (12%), Signaling (11%), Transmembrane transport (7%), and Reproduction (7%) (*SI Appendix*, Fig. S9), which are molecular functions and cellular processes frequently attributed to proteins containing IDRs (e.g., transcription factors, splicing factors, and signaling proteins). We also take note of some less widespread, but significantly overrepresented terms from the IDR-ome-based map, such as those associated with "histone modifications" (4%), "cell morphogenesis" (2%), "innate immune response" (2%), "nuclear pore complex" (1%) and "clathrin binding" (1%) (*SI Appendix*, Fig. S9). Focusing on overrepresented molecular features in the GO-term enriched clusters (Dataset S2), we find that the IDRs of RNA-associated proteins are enriched in conserved Arg-Gly/Arg-Gly-Gly (RG/RGG) motifs, Lys (K) content and K homorepeats, Arg (R) and Arg+Tyr (R + Y) content, as well as homorepeats of acidic residues [Asp (D), Glu (E)], all of which is in line with features of IDRs typically associated with phase separation, as many RNA-associated proteins interact with RNA in the context of biomolecular condensates (44, 80–82). In this case, the function of these IDRs is likely due to their direct role in RNA binding. For the IDRs associated with transmembrane transport protein GO terms, we note that many of these IDRs belong to G protein–coupled receptors (GPCRs) (Dataset S2). While these IDRs may not be directly related to receptor function, we believe that they are likely involved in trafficking of the receptors to the membrane. Consistent with this, GPCR IDRs show strong signals for glycosylation motifs (83). We also find clusters showing enrichment for categories not typically linked to IDRs, e.g., development, cell morphogenesis, extracellular space, and lipase activity (*SI Appendix*, Figs. S7 and S9), suggesting that there may be specific roles for IDRs in these biological processes that are not currently appreciated. To assess whether these results are sensitive to the 30-residue threshold used to define IDRs, we repeated our analysis using a cut-off of at least 40 consecutive disordered residues and found similar results (*SI Appendix*, Fig. S11). We also observed consistent patterns of GO term overrepresentation when applying alternative clustering methods (average and complete linkage; *SI Appendix*, Fig. S12).

Finally, the IDR-ome map offers functional hypotheses for 878 IDR-containing proteins of unknown function, nearly half of which cluster with proteins that share significant GO-term overrepresentations, suggesting a sequence-based strategy for their characterization (*SI Appendix*).

IDR Evolutionary Signatures Predict Protein GO Functions and Subcellular Localization. Next, we tested whether a systematic classification of protein function and subcellular localization across the human proteome could be achieved based on evolutionary conserved molecular features of IDRs. To this end, we applied a machine-learning approach termed Feature Analysis of IDR (FAIDR) (27) (Fig. 2). FAIDR assigns individual IDRs to specific GO terms, subcellular localizations, or other, user-defined categories, and simultaneously identifies a sparse set of IDR molecular features that are predictive for each category. We trained FAIDR on evolutionary signatures of human IDRs and GO-terms for molecular functions, biological processes, and cellular localizations, which were available on a per-protein level. When a protein contains multiple IDRs, FAIDR can be used to single out the IDR with the highest posterior probability of being responsible for the prediction of a particular functional category. We note that these annotations of function to IDRs are correlative. For example, Q-rich IDRs are often annotated to "transcription factor activity" and "sequence specific DNA-binding." In this case, the Q-rich IDRs are very likely to be the activation domains of transcription factors and sequence-specific DNA binding is annotated to the entire proteins because these proteins also contain DNA-binding domains. Similarly, SR-rich IDRs are often annotated to "RNA-splicing" or "nuclear speckle." While the IDRs in these proteins are unlikely to carry out the enzymatic steps of RNA splicing, they can facilitate localization to the nuclear speckle (84).

We first obtained predictions for protein association with GO terms based solely on evolutionary signatures of IDRs by aggregating IDR-specific predictions to a per-protein prediction using the FAIDR framework (27). To get an unbiased estimate of the predictive power, we did a protein-level cross-validation such that all IDRs belonging to a subset of proteins were held out during model training. Protein-level scores were then derived by aggregating these out-of-fold IDR predictions, which ensured that the evaluation was independent of the training data. Protein-level GO associations showed strong classification performance for certain terms, such as "nucleolus," "spliceosomal complex," and "GPCR activity" (AUC values of 0.7 or higher, with corresponding PPVs at 0.35 or above) (Fig. 2 *A* and *B*). Furthermore, "histone binding" was one of the top-performing terms with an AUC of 0.83 (Fig. 2*B*). As expected, the positively predictive molecular features of the "histone binding" model align with those overrepresented in the clusters enriched for this term, but these molecular features are distributed across multiple clusters, with alanine content overrepresented in one "histone binding" cluster and dilysine repeats, along with aspartate- and glutamate-homorepeats, dominating signals in other cluster(s). This underscores the complementary insights that can be derived from an unbiased, unsupervised clustering analysis (Fig. 1 and *SI Appendix*, Fig. S7) and the supervised FAIDR approach (Fig. 2). Strong protein-level associations were also found for "extracellular matrix" (AUC=0.82), "mRNA processing" (0.78), "GTPase regulator activity" (0.77), "centrosome" (0.75), "actin binding" (0.75), "endopeptidase activity" (0.73), and "mitotic cell cycle" (0.70) (Fig. 2 *A* and *B*). The GO terms that we could not reliably predict based on evolutionary signatures of IDRs, were not considered further (Dataset S6).

To better understand the combinations of molecular features of IDRs that are associated with protein GO annotations, we clustered the FAIDR t-statistic to examine the predictive molecular features for various GO terms (Fig. 2*C*) and those that are shared between the most commonly co-occurring terms (*SI Appendix*, Fig. S13). For instance, among several molecular features that are positively correlated with and predictive of GO terms associated with transcription, we note a high positive Z-score for Pin1 WW domain-binding motifs (DOC_WW_Pin1_4) and for specific SUMOylation motifs, such as KEPE and SUMO-1 (Fig. 2*D*, cluster 7, and *SI Appendix*). We discuss these features and other specific examples in the *SI Appendix*.

Next, for 146 high-confidence protein-level GO terms (AUC ≥ 0.7; PPV ≥ 0.35, *SI Appendix*, *SI Methods*), we provide IDR-specific annotations of function as a resource (Fig. 2*A* and Dataset S6). To generate these, we computed IDR-level probabilities for association with each GO term using FAIDR (*SI Appendix*, *SI Methods*). For each term, we chose a probability threshold cut-off so that a maximum of 1% of IDRs found in proteins not annotated to that term were above the cut-off. To estimate internal consistency of the IDR-level predictions, we retrained the FAIDR models 100 times for each term and report the fraction of runs in

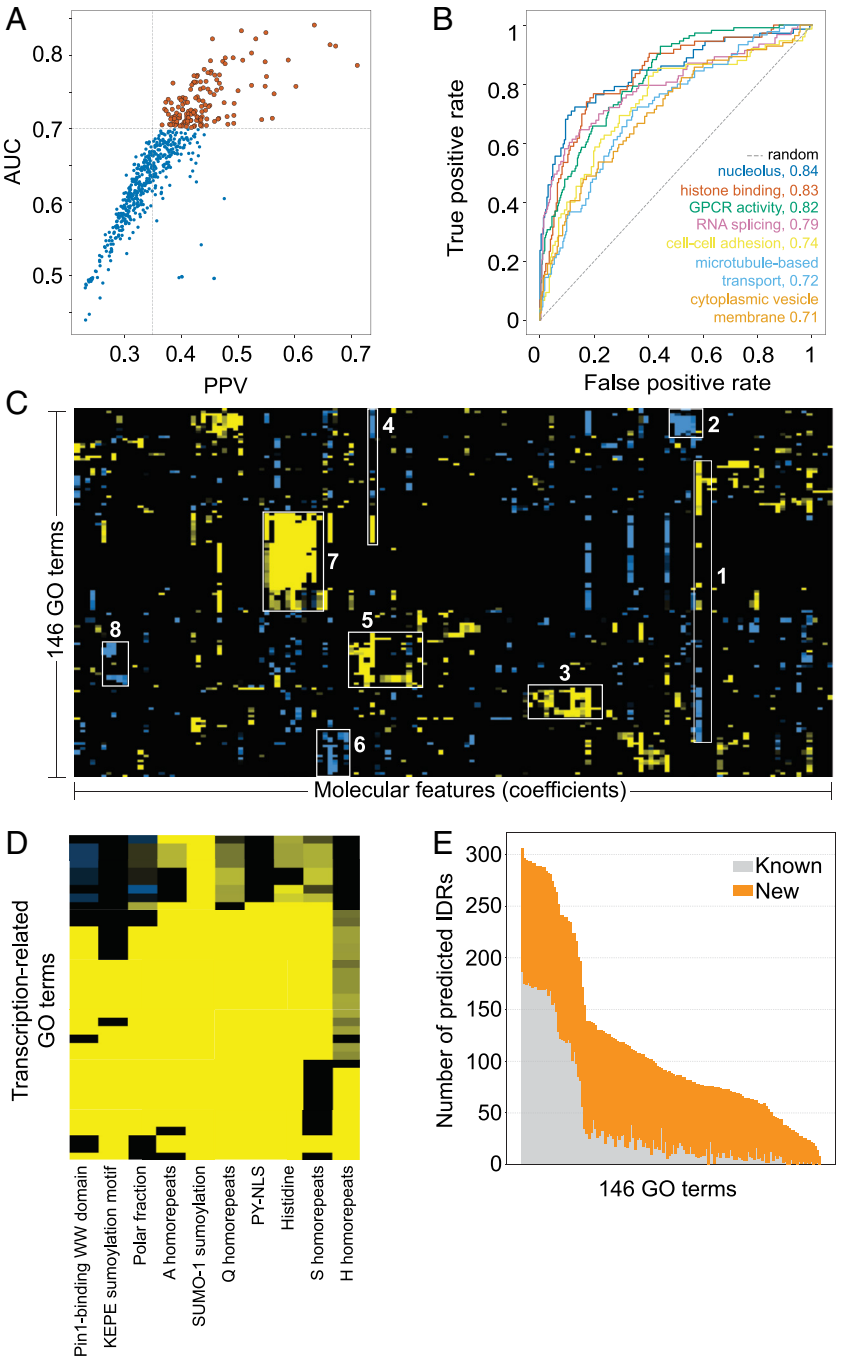

Fig. 2. Predicting GO-term association of proteins based on evolutionary signatures of IDRs. (*A*) Each point represents held-out data performance of a classifier for one of 601 GO terms covering a broad range of molecular functions, biological processes, and cellular localizations (Dataset S6). The X-axis represents PPV and the y-axis represents area under the receiver operating curve (AUC) on the held-out data for protein-level classification based solely on IDR evolutionary Z-scores. The models corresponding to data points in orange (PPV > 0.35 and AUC > 0.7) were deemed sufficiently reliable to train IDR-level classifiers (Dataset S6). (*B*) Receiver operating characteristic (ROC) curve for classification of human proteins to a representative set of GO terms based on IDR evolutionary signatures. The performance is shown on held-out data. The terms were selected to display a variation in the FAIDR performance on the binary classification tasks at the protein level. (*C*) Identifying the molecular features underlying FAIDR predictions of protein-level GO-terms and localizations. The heat map represents t-statistics summarizing the predictive importance of different molecular features (x-axis) across 146 GO terms (y-axis). Rows and columns have been organized by hierarchical clustering. Selected regions, indicated with white rectangles and numbers, are expanded in *SI Appendix*, Fig. S13. The clusters reveal how different combinations of conserved molecular features of IDRs underlie protein association with different GO terms. (*D*) Expansion of region 7 from (*C*), highlighting "transcription related" GO terms. Specific IDR features significantly associated with these GO terms were identified based on their FAIDR t-statistics. Positive t-statistics (shown in yellow) indicate features overrepresented in IDRs of proteins associated with transcription. (*E*) IDR-level annotations for 146 selected GO terms (*A*). IDRs were only annotated to a GO term if they exceeded the category-specific probability threshold consistently across 90 or more of 100 independent FAIDR repetitions. Gray bars indicate the number of "Known" annotations. An annotation of an IDR to a GO term is considered as "Known" if the protein from which the IDR originates has been previously associated with the term. If the GO term was predicted for an IDR of a protein not previously annotated with that term, the annotation was considered "New" (*SI Appendix*, Fig. S14).

which each GO term was assigned to each IDR (Dataset S6—Tab D). When an IDR is consistently associated with a GO term missing from the protein's current GO annotation, we define this as a "new" prediction for that IDR. For IDRs belonging to proteins already linked to the GO term, the IDR-level prediction remains valuable, because it provides subprotein localization of the annotation. Nonetheless, we consider such IDR annotations as "known" for the following analysis. We find that for most categories, we can predict new GO term associations for 5 to 10-fold more IDRs than currently annotated proteins, opening avenues for future hypothesis-driven validation studies (Fig. 2*E* and Dataset S6—Tab E). Based on our estimated upper bound on the false discovery rate (FDR) of the predictions, even if none of our new predictions were correct, the false discovery rates fall between 0.8 and 0.9 for most terms (Dataset S6—Tab E). This implies that testing 5 to 10 IDRs would likely yield at least one functional validation. Even with these FDRs, we believe that the IDR-level predictions will be useful to biologists, because there are currently few sequence-based approaches to obtain hypotheses about the diverse biological functions of IDRs (*Discussion*). Finally, we condensed the 146 GO terms into a smaller, nonredundant set of a few representative terms by clustering the IDR-level GO-term predictions based on Jaccard similarity (Dataset S6—Tab E and *SI Appendix*, Fig. S14 and *SI Methods*). We identified 25 distinct functional clusters and selected a representative term for each based on the lowest FDR (*SI Appendix*, Fig. S14 and Dataset S6—Tab F). This representative set of GO terms spans a broad functional landscape, including processes such as "RNA binding," "mitotic cell cycle," "G-protein coupled receptor activity," and "cytoskeleton organization." These high-confidence IDR-level

predictions (≥0.9 consistency) expand the annotated sets by sixfold on average. The prioritized assignments in Dataset S6—Tab D provide a starting point for IDR-specific experimental validation.

Disease-Associated IDRs in the Global Human IDR-ome Map. The map of the human IDR-ome also reveals molecular features of IDRs encoded by genes associated with different pathologies. Previously, we reported the enrichment of intrinsic disorder in genes related to complex diseases, such as autism-spectrum disorder (ASD) and cancer (43). Here, we asked whether our map of the human IDR-ome contains regions in which proteins encoded by genes associated with these diseases are overrepresented and, if so, what types of conserved sequence features characterize their IDRs. We found several clusters of different sizes that contain significant overrepresentation of ASD-risk and cancer genes (Dataset S7). At least eight of the clusters contain 10-fold or higher overrepresentation in ASD-risk or cancer-associated genes (P-values < 0.05, Dataset S7), and most of these clusters are not associated with any overrepresented GO terms.

Increased conservation of Q and H residue content is evident in several clusters of IDRs encoded by disease-associated genes, some of which are known to be involved in transcriptional regulation. Consistent with this, conservation of the same features was found to be predictive of autism-risk genes by FAIDR (see below). We also note a link between clusters showing enrichment in cancer census genes and stress-granules (Dataset S7), indicating a possible association between dysregulated stress granule formation and cancer. Moreover, some ASD-linked IDRs cluster closely with cancer-linked ones, which likely reflects shared processes of these IDR-containing proteins, such as transcriptional regulation (43). Our results suggest that evolutionary signatures of IDRs could improve our understanding of disease-linked genes that encode proteins with substantial intrinsic disorder.

Identifying the Molecular Features that Impart Specificity to Biomolecular Condensates. In cells, many different proteins partition into distinct biomolecular condensates (85–87). Although classifiers have been developed to predict protein association with biomolecular condensates (88–92), the molecular properties that drive the specificity and composition of different condensates are less well understood. We asked whether we could identify the molecular features of human IDRs that are predictive of specific condensate localization. We trained FAIDR on IDR-containing proteins that are known to associate with stress granules (n = 229), nuclear speckles (n = 165), or nucleoli (n = 519) based on experimentally derived datasets (*SI Appendix, SI Methods*). Examining performance across distinct datasets, obtained from Gene Ontology annotations (571 IDRs from 553 proteins, *SI Appendix, SI Methods*), revealed that evolutionary signatures of IDRs predict compartmentalization with moderate power (Fig. 3*A*), yielding AUC values of 0.69, 0.66, and 0.73 for stress granules, nuclear speckles, and nucleoli, respectively. The top 10% of IDR-containing proteins that we predict to associate with stress granules, nucleoli, and nuclear speckles are provided in Dataset S7—Tab E.

We aim to reveal features of IDRs that are characteristic of specific condensates. Accordingly, our method relies exclusively on IDR sequences for these predictions, in contrast to state-of-the-art condensate classification methods that use entire protein sequences (79, 93, 94). To explore the underlying molecular properties of IDRs that might underlie the specificity for different condensates, we compared the features of IDRs selected by the FAIDR classifiers for nuclear speckles and nucleoli (Fig. 3*B*). As expected, both models identified that nuclear localization signals (NLS) are important for IDRs that localize to both the nucleolus and nuclear speckle (Fig. 3*B*), as well as an increased presence of SUMOylation sites and a decreased overall hydropathy. For nuclear speckles, we find pY ligands of the SH2 domain NCK-1, Ser/Arg repeats, CDK and PKB phosphorylation sites, DYRK kinase ligands, and homorepeats of Gly, Gln, and Pro (Fig. 3*B*). Indeed, nuclear speckle proteins are highly phosphorylated and enriched in Ser/Arg repeats (95, 96). Thus, our FAIDR model for nuclear speckle correctly identifies Ser/Arg repeats, and multiple phosphorylation motifs as distinguishing features of IDRs found in nuclear speckle-associated proteins, although some of these are likely to be correlates, while others may be sufficient determinants of partitioning to nuclear speckles (Fig. 3*B*).

For IDRs of proteins that localize to stress granules, the overall pattern of conserved molecular features differs from those of the nuclear condensates (Fig. 3*B*). Our model identifies strong enrichments in RGG motifs, FG-rich motifs, PDZ domain ligands, and KEAP1-binding degrons (Fig. 3*B*). Multiple sets of experimental evidence confirm that these motifs are abundant in stress granule-containing IDRs (81, 97), such as PRRC2A, which contains several RG and FG motifs. KEAP1 is an adaptor protein that associates with the E3 ubiquitin ligase CUL3, and the positive association of KEAP1-binding degrons is particularly interesting in the context of recent works reporting on roles of ubiquitylation on stress granule dynamics (98, 99). Other enrichments in bulk properties include isoelectric point; sequence charge decoration; content of Trp, Asn, Ala, and Gln; dipeptide repeats NN and DD; and homorepeats of Gln, Gly, and Thr (Fig. 3*B*). A strong negative signal for the property omega (100) suggests selection for well-mixed patterning of charged and Pro residues relative to all other residues (as opposed to blocky patterning) in stress granule-associating IDRs (Fig. 3*B*). Interestingly, while a significant depletion in classical NLS is detected for stress granule IDRs, a strong enrichment is found for Pro-Tyr NLS (PY-NLS) (Fig. 3*B*), which at first glance appears counterintuitive with the cytoplasmic localization of stress granules. However, the stress granule-associated proteins FUS, EWS, and TAF-15, which all harbor PY-NLS motifs that are adjacent to RGG motifs, shuttle between the nucleus and cytoplasm in an Arg methylation-dependent manner (101). Other IDR-rich proteins with PY-NLSs also undergo nucleocytoplasmic shuttling, including hnRNPA1 and hnRNPA2 that harbor RGG motifs near the PY-NLS (102). Thus, evolutionarily conserved molecular features within IDRs are predictive of the differential localization of proteins to distinct condensates, but our models, like any machine-learning approach, identify statistical associations rather than causal determinants. IDR-mediated enrichment can arise from multiple mechanisms, including minimizing exclusion from a given environment, complementing interactions mediated by structured domains, or contributing to multivalent scaffolding, rather than from one specificity code encoded in IDRs alone. Consequently, the FAIDR-identified features should be interpreted as correlates of compartmentalization rather than sufficient determinants of recruitment, particularly in multilayered condensates with distinct microenvironments such as the nucleolus, nuclear speckle, and stress granule (103).

Leveraging Evolutionary Signatures to Discover Condensate- and Disease-Associated Proteins. To test the predictive power of FAIDR to discover new condensate-associated proteins based on molecular features of IDRs, we filtered our predictions of proteins associated with the nuclear speckle, nucleolus, and stress granule (Fig. 3*C*), and focused on proteins that were not involved in training. We correctly identify 14 (PPV 70%), 18 (PPV 90%),

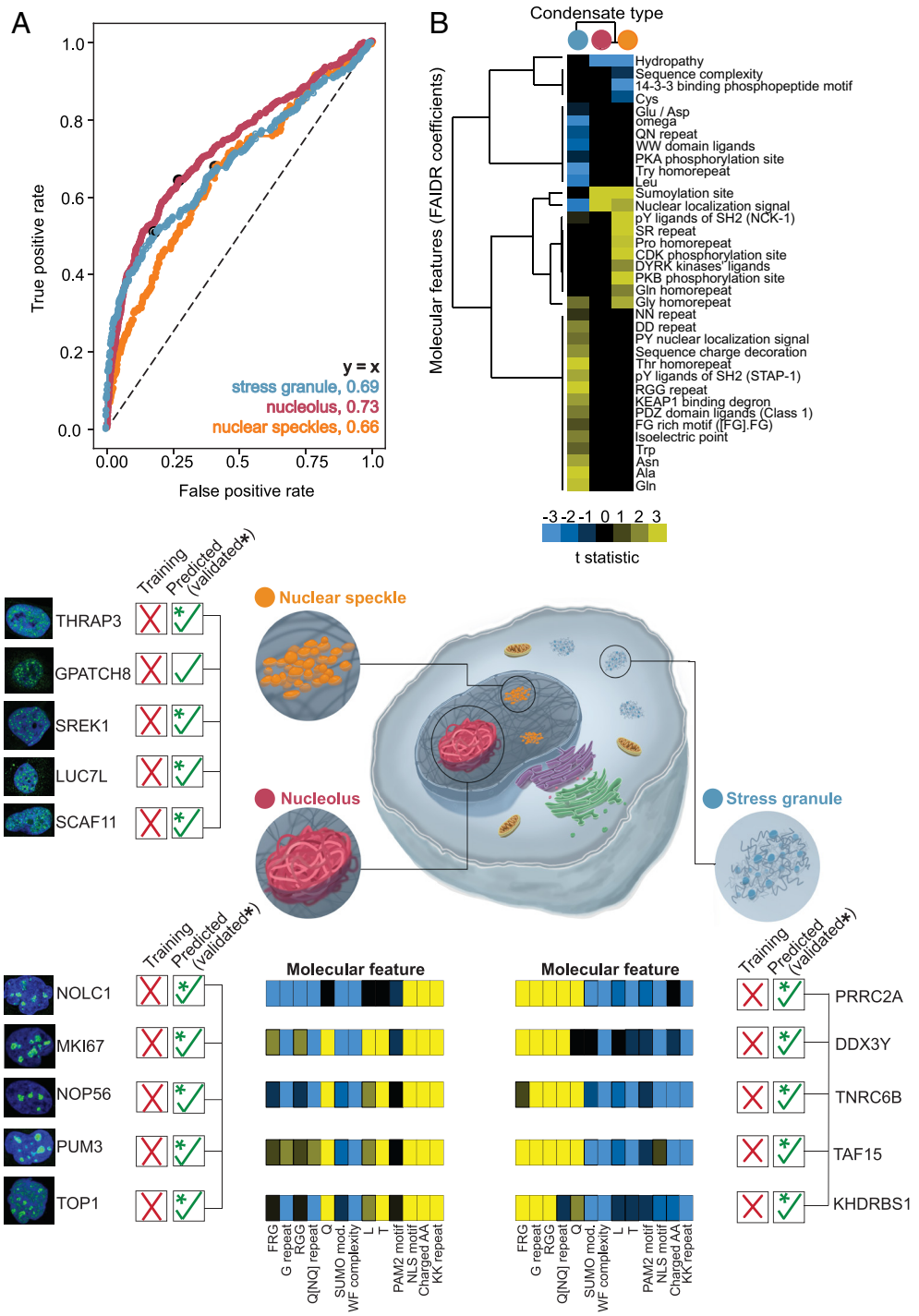

Fig. 3. Predictions of association with different bimolecular condensates. (*A*) ROC curve for classification of human IDRs to different cellular bimolecular condensates: stress granules (blue), the nucleolus (red), or nuclear speckles (orange). The performance was tested on independent and nonoverlapping datasets with condensate annotations (*SI Appendix*, *SI Methods*). AUC values are shown in the *Lower Right*. (*B*) Hierarchical clustering of the t-statistics as in *SI Appendix*, Fig. S13. A negative or positive association of a conserved molecular feature with a particular biomolecular condensate is given in blue or yellow, respectively. The condensate type is shown at the *Top* (color scheme as in *A*). (*C*) Select examples from the top predictions of association to stress granules, the nucleolus, or nuclear speckles (Dataset S7—Tab E). The examples include proteins that were not used in training and for which an association with the indicated condensate was not previously reported. For the nucleolar and stress granule-associated proteins, the evolutionary signatures (Z-scores) for selected molecular features are shown in the same representation as in Fig. 1. In *B*, FAIDR t-statistics are shown, which measure how strongly each feature contributes to class separation and depend on coefficient uncertainty. In *C*, the per-IDR evolutionary Z-scores are displayed. Therefore, the patterns in panels *B* and *C* are not directly comparable. Note that high coefficient uncertainty (e.g., due to correlations among features) can reduce t-statistics even when evolutionary Z-scores are consistent across IDRs. The micrographs were obtained from the Human Protein Atlas and were cropped to focus on regions of interest.

and 12 (PPV 60%) of the top-20 scoring proteins associated with the nucleolus, stress granule, and nuclear speckle, respectively. This performance is comparable to the 87.5% of condensate-associated proteins predicted by the PICNIC model that were experimentally validated (94). Experimental evidence in The Human Protein Atlas (104) or elsewhere in the literature provides independent validation of FAIDR predicted condensate localization (Fig. 3*C*). For example, based on our FAIDR predictions among the top scoring proteins in the nuclear speckle are THRAP3, GPATCH8, SREK1, LUC7L, and SCAF11, all of which are annotated with nuclear speckle localization by The Human Protein Atlas (Fig. 3*C*). For the nucleolus, the IDR-containing proteins NOLC1, MKI67, NOP56, PUM3, and TOP1 are all predicted as nucleolar and validated by literature reports (105–109) but were not in our training data. Finally, for the stress granule, FAIDR gives high predictive scores to the proteins PRRC2A, DDX3Y, TNRC6B, TAF15, and KHDRBS1 (Fig. 3*C*), all of which are listed as "gold standard" category (tier 1) components of stress granules in the RNA Granule Database (110) but were not in our training data.

We visualized the evolutionary signatures of IDRs for the top-scoring proteins predicted to localize to the nucleolus or stress granule (Fig. 3*C* and Dataset S7—Tab E). In this representation, we compare the molecular features of individual IDRs that FAIDR assigns the highest posterior probability of association with a given condensate. For instance, [FR]G motifs in stress granule IDRs are strongly enriched overall (Fig. 3*B*) and in four of the five examples

in Fig. 3*C*. Even though TNRC6B exhibits no evolutionary selection on (FR)G motifs, the remaining molecular features are highly similar to those of IDRs of other stress granule-localizing proteins (Fig. 3*C*). Presumably, the absence of evolutionary selection for [FR]G motifs in TNRC6B does not preclude its stress granule localization; instead, it is likely that other molecular properties function in a compensatory manner, e.g., the observed enrichments in G repeats or RGG motifs. Similar trends are seen for the nucleolus-localizing IDRs, where the overall pattern of molecular features in each IDR is similar, even if slight differences exist (e.g., no enrichment in Gln content for NOLC1, Fig. 3*C*). As an example, the nucleolar protein GTF2F1 contains a strong enrichment in RGG motifs, which are selected for in stress granule-localizing IDRs (Fig. 3*B*) and could hint toward an alternative localization for GTF2F1. Indeed, recent experimental evidence confirms that an interacting partner of GTF2F1, GTF2B, shuttles between the nucleus and stress granules (111). Thus, examination of the molecular signatures for individual IDRs can provide additional insight into the molecular properties and localization of these IDR-containing proteins.

To test whether we could apply a similar approach to disease-associated IDRs, we trained FAIDR using a curated set of IDRs from ASD-risk genes identified by Satterstrom et al. (112), and then applied the model to predict ASK-risk genes proteome-wide (*SI Appendix*, *SI Methods*). Through leave-one-out validation, FAIDR could retrieve 34% of known ASD-risk genes (recall = 0.34), with a PPV of 0.4. Notably, among the top 10% of predicted risk genes across the proteome, we identified several genes newly added to the Simons Foundation Autism Research Initiative (SFARI) database in 2023 (113) (Dataset S7). Remarkably, these and other novel predictions indicate that a model based on IDR features alone could offer predictive power for identifying new ASD-risk genes.

IDR Evolutionary Signatures Distinguish Patterns in Protein–Protein Interaction Networks. We next investigated whether clustering of IDRs based on evolutionary signatures provides insight into patterns of protein–protein interactions (Fig. 4*A*). We first focused on the conservation of SLiMs, which are typically <10 amino acid segments within IDRs that mediate transient, often low-affinity interactions with peptide-binding domains (114). We hypothesized that proteins bearing IDRs with conserved SLiMs would be more likely to interact with proteins containing canonical SLiM-binding domains. To test this, we restricted our analysis to SLiM conservation Z-scores, which analyzes the presence of the motif over the whole IDR rather than its position within the IDR. Based on the SLiM Z-scores, we reclustered the IDRs into a "SLiM conservation map." We selected 29 clusters strongly dominated by the conservation of specific SLiMs. Each cluster was labeled by the predominant conserved SLiM (Dataset S8).

For each SLiM cluster, we retrieved the interaction partners of the cluster members from the BioGRID (115) database and tested whether these interactors were enriched in canonical binding domains for the conserved SLiMs, relative to randomly drawn protein sets (Fig. 4*B*). In 17 of the 29 SLiM-based clusters, we observed significant enrichment (Benjamini–Hochberg adjusted P-value < 0.05) for the corresponding binding domains (Fig. 4*C* and Dataset S8). For example, clusters dominated by proline-rich SH3-binding motifs showed significant enrichment for SH3 domains among their interactors, with SH3 domains occurring nearly twice as frequently as in random sets. Similar enrichment was observed for clusters with several SH2-binding motifs and SH2 domains, and for clusters with phosphotyrosine-binding (PTB) motifs and PTB domains (Fig. 4*C*).

Additional enrichments included WD domains in interactors of IDRs with conserved phosphodegrons (e.g., SCF and APC/C recognition motifs), and protein kinase domains in interactors of IDRs with conserved MAP kinase docking motifs. In contrast, we did not observe significant enrichment for interactors bearing domains associated with SLiMs implicated in N-glycosylation, N-myristoylation, SUMOylation, or phosphatase docking. We note that, for these motifs, the total number of domain-containing interactors was very low, resulting in insufficient statistical power to detect significant enrichments (Fig. 4*C* and Dataset S8). This result may also reflect limited coverage of modifying enzymatic domains in protein–protein interaction database due to the transient nature of interactions with modifying enzymes, such as glycosyltransferases or phosphatases. In contrast to canonical SLiM-binding domains like SH3, SH2, and PTB, which mediate repeated motif recognition in scaffolding or signaling contexts, enzyme–substrate interactions are often short-lived and may not drive the same patterns of interaction modes when compared to classical SLiM-binding domains.

Next, we asked whether proteins in the same cluster are i) more likely to share interaction partners and ii) more likely to interact with each other, compared to random protein samples of the same size as the clusters (Fig. 4 *A* and *D*–*F*). We assessed this across all clusters extracted from the full IDR-ome map in an automatic fashion (Dataset S2 and Fig. 4*D*), the clusters from the SLiM conservation map (Fig. 4*E*), and clusters enriched in proteins known to localize to specific condensates (Dataset S7). We found that while both SLiM-based and condensate-related clusters tend to share interactors, proteins in condensate-related clusters were more likely to interact with each other. This supports our previous results indicating that we are capturing conserved features that distinguish IDRs localized to particular biomolecular condensates. Repeating this analysis across all automatically defined clusters from the functional map (distance cut-off 0.7), only a smaller subset of clusters showed a significant excess of within-cluster interactions (Z-score ≥ 3) relative to expectation. These clusters were frequently enriched for GO terms related to biomolecular condensates, such as RNA processing, chromatin regulation, and stress granule assembly, consistent with analysis of specific condensate-related clusters (Fig. 4*F* and Dataset S7).

In conclusion, evolutionary signatures of SLiM conservation in IDRs are associated with an increased prevalence of SLiM-binding domains among the interaction partners, while clusters of IDRs related to biomolecular condensates feature many shared interactions. These different patterns of protein interactions for SLiM-recognition domains (IDR-to-folded partner) and condensates (IDRs to other similar IDRs) could reflect the different types of protein complexes these IDRs form. Although both types of protein interactions are likely to be transient and relatively weak, we speculate that the difference in pattern arises from either the multivalency or lower specificity of the condensate IDR-mediated interactions such that they interact with many similar partners.

Discussion

We measured evolutionary conservation of nearly 150 bulk molecular features (26) (Dataset S1), including motif and repeat content and diverse physicochemical properties, in nearly 20,000 human IDRs (76). Using both clustering (Fig. 1 and *SI Appendix*, Fig. S7) and classification (Figs. 2 and 3 and *SI Appendix*, Fig. S13), we show that combinations of the conserved IDR molecular features correlate with specific protein-level functional annotations and subcellular localizations. Hence, by recasting the sequences of IDRs into evolutionary conserved molecular features, we identify

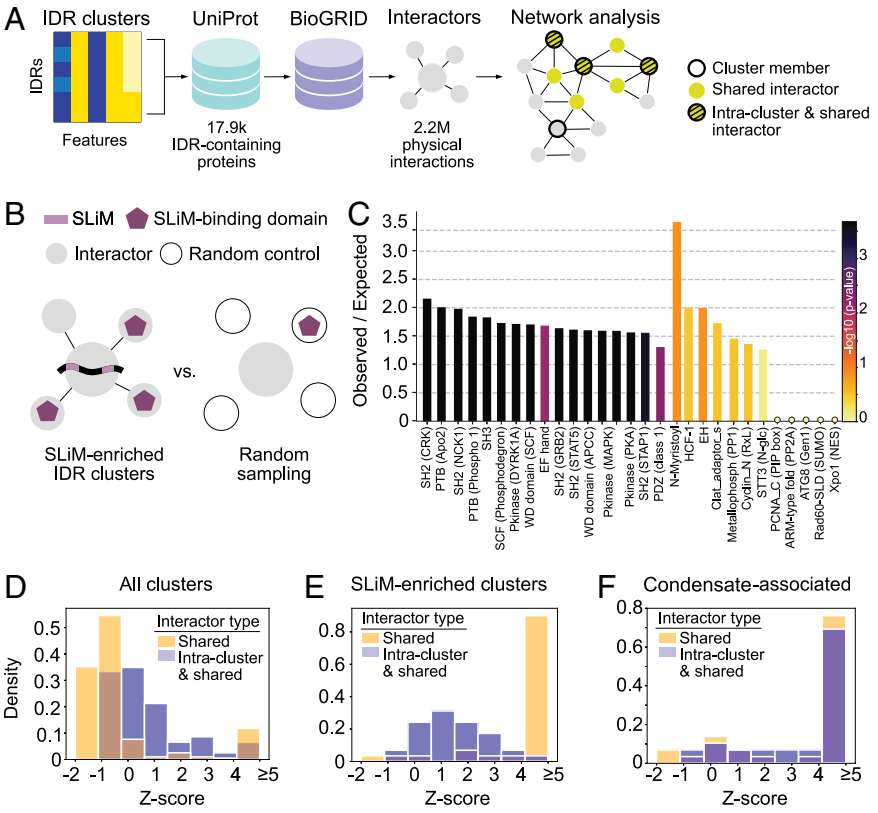

Fig. 4. Evolutionary signatures of IDRs underlie interaction network features. (*A*) Schematic outline of the analysis. Protein–protein interaction data from BioGRID (115) were used to extract interaction partners for proteins in each cluster, with UniProt IDs used for cross-referencing. Shared and intracluster interactions among cluster members were quantified. Network analysis (*Right*): Proteins containing IDRs that are cluster members are illustrated within the interaction network as circles with thick black outline. Interactors are classified as shared between cluster members (yellow circles) and as shared and cluster members (yellow circles with black crosslines). (*B*) Schematic illustrating the analysis of interaction partners of SLiM-based IDR clusters. IDRs were clustered based on conservation of SLiMs, and each cluster was labeled by its predominant conserved SLiM. For each SLiM cluster, we quantified how many interaction partners contained at least one canonical SLiM-binding domain. Enrichment for canonical SLiM-binding domains was tested relative to 10,000 randomly sampled protein sets. (*C*) Enrichment of canonical SLiM-binding domains across SLiM-dominated clusters (n = 29). Bars indicate log10-transformed empirical *P*-values, corrected for multiple testing. Color scale reflects significance; yellow and orange bars (nonsignificant) are shown after significant bars. (*D*–*F*) Z-score distributions for shared and intracluster interactors across different IDR-cluster subsets. (*D*) Automatically defined clusters from the full IDR-ome map (Dataset S2—Tab G). (*E*) SLiM-based clusters from analyses in (*B* and *C*). (*F*) Clusters associated with biomolecular condensates (Dataset S7). Enrichment of intracluster interactions is observed specifically in condensate-associated clusters, suggesting that conserved IDR features contribute to shared network context.

proteome-wide associations between IDR features and the biological roles of their host proteins. Our results lend further support to the idea that selection for or against specific features suggests a link between biological function and IDR sequence, as previously established for budding yeast and *Drosophila* IDRs (26, 116). Indeed, our feature-based concept has been increasingly recognized and used to gain further insight into function and localization of IDR-containing proteins (58, 97, 117–119). We note here that even though we identify IDR properties correlated with specific functional terms, we cannot infer from the GO associations alone that IDRs independently carry out the respective functions. In some cases, the GO terms can describe protein functions that are expected to be largely mediated by the IDR (e.g., FG-repeats of nucleoporins in nuclear transport). However, in other cases, the terms describe functionalities that likely arise from contributions of both the IDR and structured domains (e.g., terms related to transcriptional regulation).

Using the patterns of conservation in IDRs, we established a "map" of the human IDR-ome, in which we can explore groups of IDRs with similar function or location. The map of the human IDR-ome introduced here represents a resource for discovery of functional elements for vast parts of the human proteome that have thus far eluded standard bioinformatic approaches (*SI Appendix*, Fig. S1). We find that the map of the human IDR-ome recapitulates some known biological functions or processes mediated by IDR-containing proteins, such as overrepresentation of GO terms related to DNA- and RNA-binding, but also sheds light on new or underappreciated functions of IDRs, including their involvement in development and transmembrane transport. For around 40% of the clusters, there are no known GO term enrichments, which likely reflects some of the biases and difficulties associated with functional annotation (120), particularly for proteins having a large fraction of disordered residues. Importantly, the "unexplored" clusters of IDRs with similar conserved molecular features but no enriched functional terms are prime candidates for discovering new biology.

Although certain IDRs cluster together and patterns of evolutionary conservation are associated with biological functions, the clusters alone do not select the features that are most associated with each function. Which of the evolutionary conserved molecular features are most associated with function? In full-length proteins with multiple IDRs, do one or more of the IDR contribute to the biological function? To answer these questions, we used FAIDR (27) to predict association with 146 different GO terms across the human IDR-ome (Fig. 2, *SI Appendix*, Fig. S14, and Dataset S6). Our predictions reflect the rich complexity of IDR-associated functions and support discovery and contrasting of IDR features strongly associated with different IDR functional categories (Fig. 2*C* and *SI Appendix*, Fig. S13).

A limitation of our approach [and other current IDR classification efforts (78, 79, 121)] is that most IDR predictors provide information for binary classification, i.e., considering each residue as either disordered or not. Thus, disorder predictors usually do not discriminate between IDR segments with distinct properties that could support separate functions. Proximal, adjacent IDRs are expected to be merged into one long IDR, which is expected to "dilute" the sequence properties and create complicated "hybrid" IDR signatures. Because the number of possible combinations of different kinds of adjacent IDRs is large, it is unlikely that such "hybridization" would drive strong clustering. In contrast "mixed signatures" are more likely to be unique rather than forming coherent clusters. We also find that longer IDRs accumulate more extreme Z-scores simply due to increased statistical power to detect deviations from our simulations, but we mitigate this issue by using uncentered correlation (cosine similarity)

during clustering, which reduces sensitivity to Z-score magnitude. We therefore see broad mixing of IDRs of different lengths across clusters (*SI Appendix*, Fig. S15). More principled segmentation of IDRs that can distinguish between the many types of functional IDRs is an important direction for future work (122).

Our initial functional map of the human IDRs stands to be improved in additional ways. First, our current map is based on a curated list of molecular features that is limited in scope. We note that widespread association of bulk molecular properties in IDRs with diverse biological functions suggests that the evolutionary characterization of IDRs could be expanded in scope with additional biophysical properties. Recent investigations into the structural ensembles of the human IDR-ome through coarse-grained molecular dynamics simulations have revealed correlations between chain compaction and biological function (123, 124). Here, by considering the evolutionary conservation of sequence-based molecular features, we observed significant and widespread overrepresentation of GO terms in nearly 50% of human IDRs. We anticipate that placing IDRs within a higher-dimensional evolutionary and biophysical space will advance our comprehension of the molecular bases of cellular function of IDRs (19). The list of relevant molecular features will likely increase in the future, and efforts have already been taken to discover functionally relevant features in a systematic and unbiased way using self-supervised deep learning approaches (29). Second, we used a combination of unsupervised (clustering) and supervised (classification using FAIDR) analyses to make predictions about IDR function. In part, we rely on this two-stage approach because the number of IDRs with some known function (such as those found in the nuclear pore, Fig. 1) is too small (n = 64) to train a standard supervised classifier in a space of nearly 150 features. In fact, for most GO categories due to a small number of positive examples relative to the scale of the proteome, we found that the false discovery rates for individual IDR predictions at the proteome scale might be too high. Future approaches such as semisupervised, transfer-learning, or data augmentation (29, 125–127) approaches will likely address these challenges.

Understanding the impact of disease mutations in IDRs is a key area of research. Outside of IDRs with strong positional alignments, which often conditionally fold (40, 41), it is challenging to interpret disease-associated mutations that map to IDRs, which have no stable tertiary structure and, by corollary, limited positional sequence conservation. Here, we looked at overrepresentation of genes involved in two diseases in which IDRs feature prominently, autism spectrum disorder (ASD) and cancer (43). The map of the human IDR-ome reveals specific clusters that show significant enrichments in ASD-risk and cancer census genes. Based on these results, we hypothesize that mutations that disrupt conserved features of IDRs in those clusters are more likely to have a pathological impact, a focus of our future research. Functional prediction within IDRs at the residue level is a rapidly growing research area (30, 128). However, the efforts thus far focused on relatively few broad functions, such as protein binding, DNA binding, RNA binding, and linker or "entropic chain," "assembler," "scavenger," "effector," "display site," "chaperone" (28), Residue-level prediction approaches that can more closely approach the diversity of IDR function we observed in the proteome (Figs. 2 and 3) hold potential to improve the resolution of the initial map presented here, leading to insight into the functional impact of disease mutations.

An active area of IDR research focuses on the role of particular IDRs in phase separation and formation of biomolecular condensates (21, 129). How condensates achieve specificity and why certain proteins localize to certain condensates are key questions. We use a supervised classification approach to predict which proteins will localize to the nucleolus, nuclear speckle, or stress granule (AUC values of ca. 0.7 on independent test sets) and reveal which conserved molecular features of IDRs underlie those predictions. An example showcasing the utility of our feature-based approach is provided by the protein GTF2F1. While predicted to localize to the nucleolus, GTF2F1 exhibits an enrichment in "RGG motifs," typically associated with stress granule-localizing IDRs. The dynamic nature of GTF2F1's binding partner, GTF2B, which shuttles between the nucleus and stress granules (111) (c.f. *SI Appendix*, Datasets S2 and S7 herein), suggests that GTF2F1 may also undergo similar shifts in localization. The example illustrates how our predictions could guide future experimental inquiries.

Our interaction network analyses revealed that IDRs with conserved SLiMs are more likely to interact with proteins containing canonical SLiM-binding domains, while clusters of IDRs associated with biomolecular condensates showed an increased tendency for within-cluster interactions, consistent with the hypothesis that their IDRs contribute to the multivalent interactions required for condensate assembly and maintenance.

In summary, the human IDR-ome map presented here represents a vital resource for classifying vast regions of the human proteome that have evaded systematic characterization. Our resource facilitates the exploration of IDRs with similar functional annotations or subcellular localization, elucidates the conserved features underlying these annotations, and generates hypotheses for testing both known and novel IDR functions. IDR clusters that have shared conserved features but lack enriched GO terms are candidate groups for further investigation. Our predictions of IDR associations with various GO terms, along with the underlying features, provide a complementary resource for understanding the relationship between IDR sequence and function and for guiding further hypothesis-driven research efforts. Finally, prediction of ASD-risk genes underscores the predictive power of conserved IDR features and suggests a new direction for discovering the molecular basis for IDR dysregulation in this complex disease.

Together, our work provides a comprehensive functional map of the human IDR-ome based on evolutionary signatures. The map reveals known and novel combinations of specific molecular features that drive the rich complexity and promiscuous nature of IDR functions.

Methods

Prediction of Intrinsic Disorder and Boundary Definition. The reference human proteome assembly was downloaded from UniProt (Proteome: UP000005640) in August 2019. We note that "miniprotein" products of short open reading frames, many of which are likely to contain IDRs, are increasingly recognized as functionally important constituents of the human proteome (118). However, such "miniproteins" are not yet included in the reference proteome and were thus not considered here. SPOTD predictor v1.0 (66) was used to predict the per-residue probability of intrinsic disorder for every protein sequence in the human proteome. We used SPOTD v1.0 because it provided the closest agreement with NMR-determined disordered content (130, 131) and is among the most accurate predictors overall (22). A disorder probability above 0.5 was used to define disordered residues. Only protein regions with 30 or more consecutive residues that were predicted to be intrinsically disordered were considered as IDRs in all subsequent analyses. The remaining methods are described in *SI Appendix*, *SI Methods*.

Data, Materials, and Software Availability. This study made use of UniProt, ENSEMBL, PANTHER, SFARI, BioGRID, RNA Granule Database, and Human Protein Atlas databases, as specifically referenced throughout. Code and example files to compute all the steps described in the methods are available on GitHub (https://github.com/IPritisanac/IDR_ES/) (76). The hierarchically clustered evolutionary Z-scores of human IDRs (i.e., the functional map), tutorial on the exploratory and automatic analysis of the map, IDR clusters, IDR-ome sequence and alignment

files, FAIDR t-statistic, and target files for top predicted GO terms are available at Zenodo (https://zenodo.org/records/10812875) (132). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. I.P. and T.R.A. were supported by a LiUNA! Fellowship for Research Innovation from The Hospital for Sick Children and a Banting Postdoctoral Fellowship from the Canadian Institutes of Health Research (CIHR), respectively. J.D.F.-K. and A.M.M. acknowledge support from the CIHR (PJT-148532 to J.D.F.-K. and A.M.M.; FDN-148375 and PJT-190060 to J.D.F.-K.) and the Canada Foundation for Innovation for funding to A.M.M. J.D.F.-K. holds a Canada Research Chair in Intrinsically Disordered Proteins. A.M.M. holds a Canada Research Chair in Computational Biology. We thank Ozren Kisić Morduš for creating the artwork in Fig. 3. We thank Marc Singleton for his feedback on the manuscript.

Author affiliations: [a]Department of Cell and Systems Biology, University of Toronto, Toronto, ON M5S 3G5, Canada; [b]Program in Molecular Medicine, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada; [c]Department of Medicinal Chemistry, Otto Loewi Research Center, Medical University of Graz, Graz 8043, Austria; [d]Helmholtz Munich, Computational Health Center, Institute of Computational Biology, Neuherberg 85764, Germany; [e]Helmholtz Munich, Molecular Targets and Therapeutics Center, Institute of Structural Biology, Neuherberg 85764, Germany; [f]Technical University of Munich School of Natural Sciences, Department of Bioscience, Bavarian Nuclear Magnetic Resonance Center, Technical University of Munich, Garching 85747, Germany; [g]Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada; [h]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada; and [i]Department of Biochemistry, The University of Toronto, Toronto, ON M5S 1A8, Canada

Author contributions: I.P., J.D.F.-K., and A.M.M. designed research; I.P., T.R.A., Đ.K., T.Z., S.X., A.L., A.A., A.M., J.D.F.-K., and A.M.M. performed research; I.P., T.Z., J.-Y.Y., and A.M.M. contributed new reagents/analytic tools; I.P., T.R.A., Đ.K., and A. Maqsood analyzed data; T.Z. and J.-Y.Y. contributed additional data, supervision, and critical input; and I.P., T.R.A., J.D.F.-K., and A.M.M. wrote the paper.

The authors declare no competing interest.

1. M. Ashburner *et al.*, Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000), 10.1038/75556.
2. D. Lee, O. Redfern, C. Orengo, Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* **8**, 995–1005 (2007), 10.1038/NRM2281.
3. P. Radivojac *et al.*, A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013), 10.1038/NMETH.2340.
4. T. Yu *et al.*, Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023), 10.1126/SCIENCE.ADF2465.
5. T. Sanderson, M. L. Bileschi, D. Belanger, L. J. Colwell, Proteinfer, deep neural networks for protein functional inference. *Elife* **12**, e80942 (2023), 10.7554/ELIFE.80942.
6. N. Steube *et al.*, Fortuitously compatible protein surfaces primed allosteric control in cyanobacterial photoprotection. *Nat. Ecol. Evol.* **7**, 756–767 (2023), 10.1038/S41559-023-02018-8.
7. P. S. Huang, S. E. Boyken, D. Baker, The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016), 10.1038/NATURE19946.
8. B. Kuhlman, P. Bradley, Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019), 10.1038/s41580-019-0163-x.
9. A. H. W. Yeh *et al.*, De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023), 10.1038/S41586-023-05696-3.
10. F. Luppino, I. A. Adzhubei, C. A. Cassa, A. Toth-Petroczy, DeMAG predicts the effects of variants in clinically actionable genes by integrating structural and evolutionary epistatic features. *Nat. Commun.* **14**, 2230 (2023), 10.1038/S41467-023-37661-Z.
11. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010), 10.1038/NMETH0410-248.
12. T. A. Hopf *et al.*, Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017), 10.1038/NBT.3769.
13. J. Frazer *et al.*, Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021), 10.1038/S41586-021-04043-8.
14. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021), 10.1038/s41586-021-03819-2.
15. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021), 10.1126/SCIENCE.ABJ8754.
16. J. D. Forman-Kay, T. Mittag, From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure* **21**, 1492–1499 (2013), 10.1016/J.STR.2013.08.001.
17. P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015), 10.1038/nrm3920.
18. R. Lee *et al.*, Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014), 10.1021/cr400525m.
19. 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 (2023), 10.1038/S41580-023-00673-0.
20. P. Tompa, N. E. Davey, T. J. Gibson, M. M. Babu, A million peptide motifs for the molecular biologist. *Mol. Cell* **55**, 161–169 (2014), 10.1016/J.MOLCEL.2014.05.032.
21. W. Borcherds, A. Bremer, M. B. Borgia, T. Mittag, How do intrinsically disordered protein regions encode a driving force for liquid-liquid phase separation? *Curr. Opin. Struct. Biol.* **67**, 41–50 (2021), 10.1016/J.SBI.2020.09.004.
22. M. Necci *et al.*, Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* **18**, 472–481 (2021), 10.1038/s41592-021-01117-3.
23. 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), 10.1016/J.BPJ.2021.08.039.
24. S. Basu, J. Gsponer, L. Kurgan, DEPICTER2: A comprehensive webserver for intrinsic disorder and disorder function prediction. *Nucleic Acids Res.* **1**, gkad330 (2023), 10.1093/NAR/GKAD330.
25. I. Pritišanac, R. M. Vernon, A. M. Moses, J. D. Forman Kay, Entropy and information within intrinsically disordered protein regions. *Entropy* **21**, 662 (2019), 10.3390/e21070662.
26. T. Zarin *et al.*, Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife* **8**, e46883 (2019), 10.7554/eLife.46883.
27. T. Zarin *et al.*, Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife* **10**, 1–36 (2021), 10.7554/eLife.60220.
28. Y. Pang, B. Liu, DMFpred: Predicting protein disorder molecular functions based on protein cubic language model. *PLoS Comput. Biol.* **18**, e1010668 (2022), 10.1371/JOURNAL.PCBI.1010668.
29. A. X. Lu *et al.*, Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning. *PLoS Comput. Biol.* **18**, e1010238 (2022), 10.1371/JOURNAL.PCBI.1010238.
30. G. Hu *et al.*, flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.* **12**, 4438 (2021), 10.1038/s41467-021-24773-7.
31. B. Zhao *et al.*, DescribePROT: Database of amino acid-level protein structure and function predictions. *Nucleic Acids Res.* **49**, D298–D308 (2021), 10.1093/NAR/GKAA931.
32. C. F. W. Chow, S. Ghosh, A. Hadarovich, A. Toth-Petroczy, SHARK enables sensitive detection of evolutionary homologs and functional analogs in unalignable and disordered sequences. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2401622121 (2024), 10.1073/pnas.2401622121.
33. A. Mohan *et al.*, Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* **362**, 1043–1059 (2006), 10.1016/J.JMB.2006.07.087.
34. N. E. Davey, L. Simonetti, Y. Ivarsson, The next wave of interactomics: Mapping the SLiM-based interactions of the intrinsically disordered proteome. *Curr. Opin. Struct. Biol.* **80**, 102593 (2023), 10.1016/J.SBI.2023.102593.
35. M. Kumar *et al.*, The eukaryotic linear motif resource: 2022 release. *Nucleic Acids Res.* **50**, D497–D508 (2022), 10.1093/NAR/GKAB975.
36. N. Malhis, J. Gsponer, Computational identification of MoRFs in protein sequences. *Bioinformatics* **31**, 1738–1744 (2015), 10.1093/BIOINFORMATICS/BTV060.
37. A. N. Nguyen Ba *et al.*, Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Sci. Signal.* **5**, rs1 (2012), 10.1126/SCISIGNAL.2002515.
38. K. Van Roey *et al.*, Short linear motifs: Ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.* **114**, 6733–6778 (2014), 10.1021/CR400585Q.
39. N. E. Davey *et al.*, Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281 (2012), 10.1039/C1MB05231D.
40. T. R. Alderson, I. Pritišanac, D. Kolaric, 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), 10.1073/PNAS.2304302120.
41. D. Piovesan, A. M. Monzon, S. C. E. Tosatto, Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci.* **31**, e4466 (2022), 10.1002/PRO.4466.
42. C. J. Brown *et al.*, Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55**, 104–110 (2002), 10.1007/S00239-001-2309-6.
43. B. Tsang, I. Pritišanac, S. W. Scherer, A. M. Moses, J. D. Forman-Kay, Phase separation as a missing mechanism for interpretation of disease mutations. *Cell* **183**, 1742–1756 (2020), 10.1016/j.cell.2020.11.050.
44. S. Alberti, D. Dormann, Liquid-liquid phase separation in disease. *Annu. Rev. Genet.* **53**, 171–194 (2019), 10.1146/annurev-genet-112618-043527.
45. S. Basu *et al.*, Unblending of transcriptional condensates in human repeat expansion disease. *Cell* **181**, 1062–1079.e30 (2020), 10.1016/J.CELL.2020.04.018.
46. A. Molliex *et al.*, Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015), 10.1016/J.CELL.2015.09.015.
47. A. Patel *et al.*, A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**, 1066–1077 (2015), 10.1016/J.CELL.2015.07.047.
48. M. A. Mensah *et al.*, Aberrant phase separation and nucleolar dysfunction in rare genetic diseases. *Nature* **614**, 564–571 (2023), 10.1038/S41586-022-05682-1.
49. T. Nakamura *et al.*, Phase separation of FSP1 promotes ferroptosis. *Nature* **619**, 371–377 (2023), 10.1038/S41586-023-06255-6.
50. V. N. Uversky, C. J. Oldfield, A. K. Dunker, Intrinsically disordered proteins in human diseases: Introducing the D2 concept. *Annu. Rev. Biophys.* **37**, 215–246 (2008), 10.1146/ANNUREV.BIOPHYS.37.032807.125924.
51. V. Vacic *et al.*, Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Comput. Biol.* **8**, e1002709 (2012), 10.1371/JOURNAL.PCBI.1002709.
52. T. R. Alderson *et al.*, A weakened interface in the P182L variant of HSP27 associated with severe Charcot-Marie-Tooth neuropathy causes aberrant binding to interacting proteins. *EMBO J.* **40**, e103811 (2021), 10.15252/EMBJ.2019103811.
53. L. Backwell, J. A. Marsh, Diverse molecular mechanisms underlying pathogenic protein mutations: Beyond the loss-of-function paradigm. *Annu. Rev. Genomics Hum. Genet.* **23**, 475–498 (2022), 10.1146/ANNUREV-GENOM-111221-103208.
54. F. Luppino, S. Lenz, C. F. W. Chow, A. Toth-Petroczy, Deep learning tools predict variants in disordered regions with lower sensitivity. *BMC Genomics* **26**, 367 (2025), 10.1186/s12864-025-11534-9.
55. 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), 10.1016/J.CELS.2018.01.015.
56. I. Langstein-Skora *et al.*, Sequence- and chemical specificity define the functional landscape of intrinsically disordered regions. *Nat. Cell. Biol.* **28**, 323–337 (2026).
57. 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), 10.1016/J.CELS.2022.01.002.
58. M. C. Cohan, M. K. Shinn, J. M. Lalmansingh, R. V. Pappu, Uncovering non-random binary patterns within sequences of intrinsically disordered proteins. *J. Mol. Biol.* **434**, 167373 (2022), 10.1016/J.JMB.2021.167373.
59. 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), 10.1073/PNAS.2211178119.

60. T. Zarin, C. N. Tsai, A. N. N. Ba, A. M. Moses, Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E1450–E1459 (2017), 10.1073/PNAS.1614787114.
61. R. M. C. Vernon *et al.*, Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**, e31486 (2018), 10.7554/eLife.31486.
62. A. K. Lancaster, A. Nutter-Upham, S. Lindquist, O. D. King, PLAAC: A web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* **30**, 2501 (2014), 10.1093/BIOINFORMATICS/BTU310.
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), 10.1038/S41594-022-00811-W.
64. L. Y. Beh, L. J. Colwell, N. J. Francis, A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E1063–71 (2012), 10.1073/PNAS.1118678109.
65. J. J. Alston, A. Soranno, A. S. Holehouse, Conserved molecular recognition by an intrinsically disordered region in the absence of sequence conservation. *Biophys. J.* **123**, 26a (2024), 10.1016/j.bpj.2023.11.260.
66. J. Hanson, Y. Yang, K. Paliwal, Y. Zhou, Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **33**, 685–692 (2017), 10.1093/BIOINFORMATICS/BTW678.
67. K. L. Howe *et al.*, Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021), 10.1093/NAR/GKAA942.
68. C. N. Ravarani *et al.*, High-throughput discovery of functional disordered regions: Investigation of transactivation domains. *Mol. Syst. Biol.* **14**, e8190 (2018), 10.15252/MSB.20188190.
69. C. Warren, D. Shechter, Fly fishing for histones: Catch and release by histone chaperone intrinsically disordered regions and acidic stretches. *J. Mol. Biol.* **429**, 2401–2426 (2017), 10.1016/J.JMB.2017.06.005.
70. A. Schlessinger *et al.*, Protein disorder–a breakthrough invention of evolution? *Curr. Opin. Struct. Biol.* **21**, 412–418 (2011), 10.1016/J.SBI.2011.03.014.
71. A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, R. V. Pappu, Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 8183–8188 (2010), 10.1073/PNAS.0911107107.
72. S. C. Strickfaden *et al.*, A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. *Cell* **128**, 519–531 (2007), 10.1016/J.CELL.2006.12.032.
73. S. Chavali, A. K. Singh, B. Santhanam, M. M. Babu, Amino acid homorepeats in proteins. *Nat. Rev. Chem.* **4**, 420–434 (2020), 10.1038/S41570-020-0204-1.
74. R. Gemayel *et al.*, Variable glutamine-rich repeats modulate transcription factor activity. *Mol. Cell* **59**, 615–627 (2015), 10.1016/J.MOLCEL.2015.07.003.
75. S. Chavali *et al.*, Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. *Nat. Struct. Mol. Biol.* **24**, 765–777 (2017), 10.1038/NSMB.3441.
76. I. Pritišanac, Codes for 'A functional map of the human intrinsically disordered proteome.' GitHub. https://github.com/IPritisanac/IDR_ES/. Deposited 19 March 2024.
77. S. F. Altschul *et al.*, Gapped BLAST and PSI-blast: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997), 10.1093/NAR/25.17.3389.
78. K. M. Ruff *et al.*, Molecular grammars of predicted intrinsically disordered regions that span the human proteome. *Cell* **189**, 323–342.e17 (2025), 10.1016/j.cell.2025.10.019.
79. H. R. Kilgore *et al.*, Protein codes promote selective subcellular compartmentalization. *Science* **387**, 1095–1101 (2025), 10.1126/science.adq2634.
80. P. A. Chong, R. M. Vernon, J. D. Forman-Kay, RGG/RG motif regions in RNA binding and phase separation. *J. Mol. Biol.* **430**, 4650–4665 (2018), 10.1016/j.jmb.2018.06.014.
81. J. Y. Youn *et al.*, High-density proximity mapping reveals the subcellular organization of mRNA-associated granules and bodies. *Mol. Cell* **69**, 517–532.e11 (2018), 10.1016/J.MOLCEL.2017.12.020.
82. J. Y. Youn *et al.*, Properties of stress granule and P-body proteomes. *Mol. Cell* **76**, 286–294 (2019), 10.1016/J.MOLCEL.2019.09.014.
83. S. E. Bondos, A. K. Dunker, V. N. Uversky, Intrinsically disordered proteins play diverse roles in cell signaling. *Cell Commun. Signal.* **20**, 20 (2022), 10.1186/s12964-022-00821-7.
84. J. F. Cáceres, T. Misteli, G. R. Screaton, D. L. Spector, A. R. Krainer, Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity. *J. Cell Biol.* **138**, 225–238 (1997), 10.1083/jcb.138.2.225.
85. S. F. Banani, H. O. Lee, A. A. Hyman, M. K. Rosen, Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017), 10.1038/NRM.2017.7.
86. A. S. Lyon, W. B. Peeples, M. K. Rosen, A framework for understanding the functions of biomolecular condensates across scales. *Nat. Rev. Mol. Cell Biol.* **22**, 215–235 (2021), 10.1038/S41580-020-00303-Z.
87. J. D. Forman-Kay, J. A. Ditlev, M. L. Nosella, H. O. Lee, What are the distinguishing features and size requirements of biomolecular condensates and their implications for RNA-containing condensates? *RNA* **28**, 36–47 (2022), 10.1261/RNA.079026.121.
88. A. Hadarovich *et al.*, PICNIC accurately predicts condensate-forming proteins regardless of their structural disorder across organisms. *Nat. Commun.* **15**, 10668 (2024).
89. R. M. Vernon, J. D. Forman-Kay, First-generation predictors of biological protein phase separation. *Curr. Opin. Struct. Biol.* **58**, 88–96 (2019), 10.1016/J.SBI.2019.05.016.
90. M. Vendruscolo, M. Fuxreiter, Towards sequence-based principles for protein phase separation predictions. *Curr. Opin. Chem. Biol.* **75**, 102317 (2023), 10.1016/J.CBPA.2023.102317.
91. X. Chu *et al.*, Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics* **23**, 72 (2022), 10.1186/S12859-022-04599-W.
92. H. Cai, R. M. Vernon, J. D. Forman-Kay, An interpretable machine-learning algorithm to predict disordered protein phase separation based on biophysical interactions. *Biomolecules* **12**, 1131 (2022), 10.3390/BIOM12081131.
93. K. L. Saar *et al.*, Protein condensate atlas from predictive models of heteromolecular condensate composition. *Nat. Commun.* **15**, 5418 (2024), 10.1038/s41467-024-48496-7.
94. A. Hadarovich *et al.*, PICNIC accurately predicts condensate-forming proteins regardless of their structural disorder across organisms. *Nat. Commun.* **15**, 10668 (2024), 10.1038/s41467-024-55089-x.
95. A. Krämer, The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu. Rev. Biochem.* **65**, 367–409 (1996), 10.1146/ANNUREV.BI.65.070196.002055.
96. H. Li, P. M. Bingham, Arginine/serine-rich domains of the su(wa) and tra RNA processing regulators target proteins to a subnuclear compartment implicated in splicing. *Cell* **67**, 335–342 (1991), 10.1016/0092-8674(91)90185-2.
97. S. R. Millar *et al.*, A new phase of networking: The molecular composition and regulatory dynamics of mammalian stress granules. *Chem. Rev.* **123**, 9036–9064 (2023), 10.1021/ACS.CHEMREV.2C00608.
98. Y. Gwon *et al.*, Ubiquitination of G3BP1 mediates stress granule disassembly in a context-specific manner. *Science* **372**, eabf6548 (2021), 10.1126/SCIENCE.ABF6548.
99. B. A. Maxwell *et al.*, Ubiquitination is essential for recovery of cellular activities after heat shock. *Science* **372**, eabc3593 (2021), 10.1126/SCIENCE.ABC3593.
100. E. W. Martin *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), 10.1021/JACS.6B10272.
101. D. Dormann *et al.*, Arginine methylation next to the PY-NLS modulates transportin binding and nuclear import of FUS. *EMBO J.* **31**, 4258–4275 (2012), 10.1038/EMBOJ.2012.261.
102. L. Guo *et al.*, Nuclear-import receptors reverse aberrant phase transitions of RNA-binding proteins with prion-like domains. *Cell* **173**, 677–692.e20 (2018), 10.1016/J.CELL.2018.03.002.
103. A. S. Holehouse, S. Alberti, Molecular determinants of condensate composition. *Mol. Cell* **85**, 290–308 (2025), 10.1016/j.molcel.2024.12.021.
104. M. Uhlén *et al.*, Tissue-based map of the human proteome. *Science* **347**, 394 (2015), 10.1126/science.1260419.
105. P. Rallabhandi *et al.*, Sumoylation of topoisomerase I is involved in its partitioning between nucleoli and nucleoplasm and its clearing from nucleoli in response to camptothecin. *J. Biol. Chem.* **277**, 40020–40026 (2002), 10.1074/JBC.M200388200.
106. H. Y. Chang *et al.*, hPuf-A/KIAA0020 modulates PARP-1 cleavage upon genotoxic stress. *Cancer Res.* **71**, 1126–1134 (2011), 10.1158/0008-5472.CAN-10-1831.
107. S. Singh, A. V. Broeck, L. Miller, M. Chaker-Margot, S. Klinge, Nucleolar maturation of the human small subunit processome. *Science* **373**, eabj5338 (2021), 10.1126/SCIENCE.ABJ5338.
108. Y. Ahmad, F. M. Boisvert, E. Lundberg, M. Uhlen, A. I. Lamond, Systematic analysis of protein pools, isoforms, and modifications affecting turnover and subcellular localization. *Mol. Cell. Proteomics* **11**, 013680 (2012), 10.1074/MCP.M111.013680.
109. C. Y. Pai, H. K. Chen, H. L. Sheu, N. H. Yeh, Cell-cycle-dependent alterations of a highly phosphorylated nucleolar protein p130 are associated with nucleologenesis. *J. Cell Sci.* **108**, 1911–1920 (1995), 10.1242/JCS.108.5.1911.
110. J. Y. Youn *et al.*, High-density proximity mapping reveals the subcellular organization of mRNA-associated granules and bodies. *Mol. Cell* **69**, 517–532.e11 (2018), 10.1016/J.MOLCEL.2017.12.020.
111. W. Qin *et al.*, Dynamic mapping of proteome trafficking within and between living cells by TransitID. *Cell* **186**, 3307–3324.e30 (2023), 10.1016/j.cell.2023.05.044.
112. F. K. Satterstrom *et al.*, Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e23 (2020), 10.1016/J.CELL.2019.12.036.
113. B. S. Abrahams *et al.*, SFARI gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4**, 36 (2013), 10.1186/2040-2392-4-36.
114. N. E. Davey *et al.*, Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281 (2012), 10.1039/C1MB05231D.
115. R. Oughtred *et al.*, The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021), 10.1002/PRO.3978.
116. M. Singleton, M. Eisen, Evolutionary analyses of IDRs reveal widespread signals of conservation. *PLoS Comput. Biol.* **25**, e1012028. (2024).
117. M. R. King, K. M. Ruff, R. V. Pappu, Emergent microenvironments of nucleoli. *Nucleus* **15**, 2319957 (2024), 10.1080/19491034.2024.2319957.
118. E. E. Duffy *et al.*, Developmental dynamics of RNA translation in the human brain. *Nat. Neurosci.* **25**, 1353–1365 (2022), 10.1038/S41593-022-01164-9.
119. L. O. Loureiro *et al.*, A recurrent SHANK3 frameshift variant in autism spectrum disorder. *NPJ Genom. Med.* **6**, 91 (2021), 10.1038/S41525-021-00254-0.
120. G. Kustatscher *et al.*, Understudied proteins: Opportunities and challenges for functional proteomics. *Nat. Methods* **19**, 774–779 (2022), 10.1038/s41592-022-01454-x.
121. G. Tesei *et al.*, Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024), 10.1038/s41586-023-07004-5.
122. A. G. Sangster *et al.*, Zero-shot segmentation using embeddings from a protein language model identifies functional regions in the human proteome. *PLoS Comput. Biol.* **21**, e1012929 (2025), 10.1371/journal.pcbi.1012929.
123. 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).
124. G. Tesei *et al.*, Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation. *Nature* **626**, 897–904 (2024).
125. K. Lindorff-Larsen, B. B. Kragelund, On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *J. Mol. Biol.* **433**, 167196 (2021), 10.1016/J.JMB.2021.167196.
126. Y. Pang, B. Liu, TransDFL: Identification of disordered flexible linkers in proteins by transfer learning. *Genomics Proteomics Bioinformatics* **21**, 359–369 (2023), 10.1016/J.GPB.2022.10.004.
127. N. K. Lee, Z. Tang, S. Toneyan, P. K. Koo, EvoAug: Improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations. *Genome Biol.* **24**, 105 (2023), 10.1186/S13059-023-02941-W.
128. A. Barik *et al.*, Depicter: Intrinsic disorder and disorder function prediction server. *J. Mol. Biol.* **432**, 3379–3387 (2020), 10.1016/J.JMB.2019.12.030.
129. N. Rostam *et al.*, CD-CODE: Crowdsourcing condensate database and encyclopedia. *Nat. Methods* **20**, 673–676 (2023), 10.1038/S41592-023-01831-0.
130. J. T. Nielsen, F. A. A. Mulder, Quality and bias of protein disorder predictors. *Sci. Rep.* **9**, 5137 (2019), 10.1038/S41598-019-41644-W.
131. R. Dass, F. A. A. Mulder, J. T. Nielsen, ODiNPred: Comprehensive prediction of protein order and disorder. *Sci. Rep.* **10**, 14780 (2020), 10.1038/S41598-020-71716-1.
132. I. Pritišanac, Data repository associated with 'A functional map of the human intrinsically disordered proteome' (1.0). Zenodo. https://doi.org/10.5281/zenodo.10812875. Deposited 13 March 2024.