Disobind: a sequence-based, partner-dependent contact map and interface residue predictor for intrinsically disordered regions

Kartik Majila[1,*], Varun Ullanat[1] and Shruthi Viswanath[1,*,+]

[1]National Center for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, Karnataka, India 560065.

[*]Corresponding authors.
Email: kartikm@ncbs.res.in (KM)
Email: shruthiv@ncbs.res.in (SV)

[+]Lead contact
Shruthi Viswanath. Email: shruthiv@ncbs.res.in

Summary

Intrinsically disordered proteins or regions (IDPs/IDRs) adopt diverse binding modes with different partners, from coupled-folding-and-binding, to fuzzy binding, to fully-disordered binding. Characterizing IDR interfaces is challenging experimentally and computationally. The state-of-the-art AlphaFold-multimer and AlphaFold3 can be used to predict IDR binding sites, although they are less accurate at their benchmarked confidence cutoffs. Here, we developed Disobind, a deep-learning method that predicts inter-protein contact maps and interface residues for an IDR and its partner, given their sequences. It uses sequence embeddings from the ProtT5 protein language model. Disobind outperforms state-of-the-art interface predictors for IDRs. It also outperforms AlphaFold-multimer and AlphaFold3 at multiple confidence cutoffs. Combining Disobind and AlphaFold-multimer predictions further improves the performance. In contrast to current methods, Disobind considers the context of the binding partner and does not depend on structures and multiple sequence alignments. Its predictions can be used to localize IDRs in large assemblies and characterize IDR-mediated interactions.

A record of this paper's transparent peer review process is included in the supplemental information.

Keywords: Intrinsically disordered proteins (IDP), intrinsically disordered regions (IDR), protein structure, deep learning (DL), protein language model (pLMs)

Introduction

Intrinsically disordered proteins or regions (IDPs or IDRs) lack a well-defined three-dimensional structure in their monomeric state [1,2]. They exist as an ensemble of interconverting conformers in equilibrium and hence are structurally heterogeneous [1,3,4]. Here, we use the term IDR for both IDPs and IDRs. The heterogeneity of IDRs provides several functional advantages including the ability to overcome steric restrictions, have a large capture radius, undergo functional misfolding, and interact with multiple partners [5]. They play critical roles in cellular processes including signaling, intracellular transport, protein folding, and condensate formation [4,5].

IDRs are known to mediate a large number of protein-protein interactions [6]. They exhibit considerable diversity in their binding modes within complexes [7,8]. Upon binding to a partner protein, they may adopt an ordered state (disorder-to-order or DOR) or may remain disordered (disorder-to-disorder or DDR) [7,9–11]. Moreover, depending on the binding partner, the same IDR may undergo DOR and DDR transitions [10,11]. For example, IDRs with short linear motifs (SLiMs) may form multivalent interactions with their partner, exchanging between conformations where one or more motifs bind at a time [11]. This heterogeneity in binding makes their structural characterization in complexes challenging.

Here, we develop a deep learning method to determine the binding sites of an IDR and an interacting partner protein given their sequences. Several methods have been developed for predicting interface residues, inter-protein contact maps, and structures of complexes formed by two proteins. Some of them rely on the availability of structures of the input proteins [12–17]. Sequence-based methods that require a multiple sequence alignment (MSA) can be constrained by the low sequence conservation of IDRs [18–21]. Other sequence-based methods for IDRs use physicochemical features or embeddings from protein language models (pLMs) but do not account for the partner protein [22–25]. Yet, the identity of the partner protein is an important determinant of IDR interactions, as these are context-dependent. To this end, FINCHES predicts intermolecular interactions for two IDRs based on their chemical specificity, given their sequences. However, it is currently applicable to IDRs that remain disordered upon binding [26].

State-of-the-art methods like AlphaFold2 (AF2) and AlphaFold3 (AF3) can provide accurate predictions for a range of macromolecules [18,20,27]. However, they tend to provide low-confidence predictions for IDRs. Even though the low-confidence regions predicted by AF2 often correspond to the presence of IDRs, the predicted structure cannot be considered a representative IDR structure [28,29]. Recent studies showed that AlphaFold-multimer (referred to as AF2 hereon) could predict the structures of complexes where the IDR undergoes a DOR transition, including protein-peptide complexes and domain-motif complexes [19,21,30]. However, the results may be sensitive to the inputs, such as the sequence fragment size, the fragment delimitation, and the alignment mode for the MSA [19,21]. In general, predicting the binding interfaces for IDRs at high resolution remains a challenge for current methods [31].

Our method, Disobind, predicts the inter-protein contact maps and interface residues for an IDR and its partner, which could be an IDR or an ordered protein, from their sequences. Disobind is trained on experimental structures of IDRs in complexes. However, predicting inter-protein contact maps for IDRs presents several challenges. The structural data for IDRs in complexes is limited, the IDRs may retain their heterogeneity in complexes, and the inter-protein contact maps can be sparse. Here, we leverage the pLM ProtT5 to train our model with limited data, without the need for large, resource-intensive architectures. To address the problems associated with sparsity and heterogeneity, we predict coarse-grained inter-protein contact maps and interface residues. Coarse-graining reduces the resolution of the contact maps, making the prediction task easier. Disobind performs better than state-of-the-art interface predictors for IDRs and better than AF2 and AF3 across multiple confidence cutoffs used in this study. Combining the Disobind and AF2 predictions further improves the performance. We demonstrate the performance of Disobind+AF2 on several IDR complexes, including DOR and DDR complexes. Predictions from the method can be used to localize IDRs in integrative structures of large assemblies. They can be used to characterize protein-protein interactions involving IDRs, identify novel motifs, and modulate IDR-mediated interactions.

Results

Disobind dataset creation

Given a pair of protein sequences, at least one of which is an IDR, Disobind predicts binary inter-protein contact maps and interface residues for the pair (Fig 1). We compile our dataset by gathering experimental structures of IDR-containing complexes from an array of existing IDR databases and datasets, including DIBS, MFIB, FuzDB, PDBtot, PDBcdr, DisProt, IDEAL, and MobiDB, covering a range of IDR binding modes (Fig 1) (See Methods) [9,32–36]. We first process the chains in the PDB structures of the IDR-containing complexes (Fig. 1). Missing residues in the structure cannot be used to infer contact maps reliably. Therefore, segments corresponding to missing residues are removed, resulting in fragments of the chain. Long chains are also fragmented to 100 residues for memory efficiency. Notably, this does not impose any sequence length restriction for Disobind; it merely determines the crop sizes used in training. At inference, no length restriction exists, and Disobind can be used for sequences of arbitrary lengths.

We then designate the sequence fragments derived from the chains in the PDB structures as IDRs and non-IDRs. We identify disordered residues based on annotations supported by experimental evidence and sequence-based homology obtained from DisProt, IDEAL, and MobiDB. Based on these annotations, sequence fragments with more than 20% disordered residues are considered as IDR fragments, whereas the others are considered as non-IDR fragments (See Methods). The 20% threshold is used solely as a criterion for inclusion in our dataset and is not intended as a general definition of an IDR. This cutoff was chosen to ensure a sufficiently large dataset for training, while limiting the over-representation of ordered residues.

Next, we obtain inter-chain binary complexes from each PDB structure. A binary complex comprises sequence fragments from different chains, of which at least one is an IDR fragment (Fig. 1). Each such binary complex is associated with a contact map; we filter out complexes that do not contain any contacts. The binary complexes

are then grouped based on the UniProt accessions of the corresponding sequence fragment pair.

We then merge all contact maps with overlapping sequences in a group using a logical OR operation to obtain merged contact maps (Fig. 1). A sequence fragment pair, along with its corresponding merged contact map, forms a merged binary complex (See Methods) (Fig. 1). The latter represents contacts formed across all available structures of complexes for a given sequence pair, accounting for the multiple conformations of an IDR in a complex. This simplified output representation avoids making any assumptions about binding mode, allowing it to be applicable to DOR and DDR cases. Further, this provides for a less sparse output representation compared to an ensemble of contact maps, while also mitigating the bias caused by having an unequal number of structures for different sequence fragment pairs. Disobind is trained on a dataset of merged binary complexes with the sequence fragments as input and the merged contact map as output.

We then create an out-of-distribution (OOD) test set comprising merged binary complexes where the input sequence fragments share less than 20% identity with the Disobind dataset and with the AlphaFold2 PDB70 dataset. The remaining complexes are further split into a train, validation, and an in-distribution (ID) test set. The ID and OOD test sets comprise 297 and 52 merged binary complexes respectively. Adding more complexes to the OOD test set may be achieved by increasing the sequence identity cutoff but would also increase similarity to the AF2 training dataset, which we attempted to avoid.

Model architecture

Disobind uses a shallow feedforward neural network architecture with the sequence embeddings obtained from a protein language model as input (Fig 2). The input embeddings are projected to a lower dimension using a projection block. Next, in the interaction block, the projected embeddings are used to compute the outer product and outer difference, which are subsequently concatenated along the feature dimension. This allows the model to capture complementarity between the input

sequences and yields an interaction tensor. This interaction tensor is processed by a multi-layer perceptron (MLP) for contact map prediction. For interface residue prediction, the interaction tensor is first reduced by computing row-wise and column-wise averages and further processed by an MLP. Finally, an output block provides element-wise sigmoid scores, with a score greater than 0.5 representing a contact or interface residue. The outputs are binary contact maps or binary interface residue predictions on the fragments. The same architecture is used to train separate models for the contact map and interface residue prediction tasks. The model is trained with the singularity-enhanced loss (SE loss) [37] using the AdamW optimizer with weight decay (Fig S1). This loss is a modified form of the binary cross-entropy (BCE) loss used for class-imbalanced training. We varied the hyperparameters of the network including the projection dimension, the number of layers in the MLP, and the SE loss parameters (Table S1-S4, Fig S1). We use recall, precision, and F1-score to evaluate model performance (See Methods).

Evaluating Disobind

Inter-protein contact map prediction

We first evaluate Disobind on predicting residue-wise inter-protein contact maps on the ID and OOD test sets. Disobind achieves an F1-score of 0.57 and 0.33 on the ID and OOD test sets respectively (Table 1 and Table S5, contact map prediction at coarse-grained (CG) resolution 1). It performs better than a random baseline (see Methods) on the OOD test set (Table 1). However, predicting residue-wise inter-protein contact maps is challenging due to their sparsity (Table S6). This leads to a class imbalance, biasing the model to predict just the majority class, *i.e.,* 0's or non-contacts. The heterogeneity of IDRs in complexes poses another challenge. For example, the multivalency of IDRs may result in different sets of contacts in different conformations with the same partner [38].

Interface residue prediction

As an alternative to predicting inter-protein residue-wise contact maps, next, we predict interface residues for the IDR and its partner (Fig 3a). For two input sequence fragments of length L_1 and L_2, this reduces the number of predicted elements from $L_1 \times L_2$ in the contact map prediction to $L_1 + L_2$ in the interface residue prediction (Fig 3b). This helps mitigate the class imbalance as the interface residue predictions are less sparse (Fig 3a, Fig 3b Table S6, interface residue prediction at coarse-grained (CG) resolution 1). Interface residue prediction is also less affected by the heterogeneity of IDRs in complexes. The prediction for each residue is simplified to determining whether it binds to the partner, rather than identifying the specific residues within the partner it interacts with.

As expected, Disobind performs better in interface residue prediction than in contact map prediction. It achieves an F1 score of 0.70 and 0.48 on the ID and OOD test sets respectively (Fig S2, Table 1 and Table S5, interface residue prediction at coarse-grained (CG) resolution 1). It performs better than a random baseline evaluated on the OOD test set (Table 1).

Coarse-graining improves the performance

To further improve the performance of our model, we predict coarse-grained contact maps and interface residues (Fig 3a, Fig 3b). Coarse-graining has been widely used for modeling biological systems [39,40]. Coarse-graining over a set of residue pairs in a contact map or a set of residues in the interfaces further reduces the sparsity and class imbalance, and mitigates the problems posed by the multivalency of IDR interactions (Fig 3a, Fig 3b, Table S6). Notably, we coarse-grain over embeddings from pLMs. The latter provide context-aware, abstract representations of the sequence. Further, coarse-graining along the sequence is well-suited for the modular and patch-driven nature of IDR interactions, that are mediated by charge patterning, attractive sticker regions, conserved short linear motifs (SLiMs) and molecular recognition features (MoRFs), and/or less conserved low-complexity regions (LCRs) [7,41–43].

We train Disobind to predict coarse-grained contact maps and coarse-grained interface residues at resolutions of 5 and 10 contiguous residues along the backbone. The coarse-grained embeddings are obtained by averaging the input embeddings [44]. As expected, coarse-graining further improves the performance of Disobind on the ID and OOD test sets for both the contact map and interface residue prediction (Fig S2, Table 1, Table S5). For all cases, Disobind performs better than a random baseline on the OOD test set.

Comparison to AlphaFold2 and AlphaFold3

We further compare the performance of Disobind with the state-of-the-art methods AF2 and AF3 on the OOD test set. For evaluating AF2 and AF3 outputs, we mask out interactions predicted at low confidence based on the interface-predicted template matching (ipTM) score, predicted local distance difference test (pLDDT), and predicted aligned error (PAE) metrics (see Methods).

Using different ipTM cutoffs for AF2 and AF3

The ipTM score is a metric provided by AF2 and AF3 to assess the confidence in the predicted interfaces in a complex [18,27]. A prediction with an ipTM score greater than 0.75 is considered highly confident, while an ipTM lower than 0.6 indicates a likely failed prediction [45,46]. We first evaluate the performance of AF2 and AF3 with an ipTM cutoff of 0.75. Disobind performs better than AF2 and AF3 for all tasks, *i.e.,* contact map and interface residue predictions across different coarse-grained resolutions (Table 1, ipTM = 0.75). Disordered residues, however, are known to impact the pTM and ipTM metrics negatively [18,47]. Considering this, we relaxed the ipTM cutoff to 0.4 to assess the performance of AF2 and AF3, while keeping the per-residue metric, *i.e.* pLDDT and PAE cutoffs as is, guided by recent benchmarks and the AF3 guide [18,30]. Here also, Disobind performs better than both AF2 and AF3 for both contact map and interface residue prediction (Table 1, ipTM = 0.4). AF3 performs worse than both Disobind and AF2. Furthermore, this trend remains more or less the same even when we do not apply an ipTM cutoff (Table 1, ipTM = 0.0).

Further, AF2 and AF3 are trained to output a distogram [18,20,27]. The auxiliary distogram head is used to ensure that the learned internal pair representation accurately reflects the distribution of distances between the residues (See section Distogram head in AF2 and AF3 supplementary text). AF3 provides the contact probabilities derived from the distogram head in the output JSON file. We compared the performance of AF3 on the contact map and interface residue prediction task on the OOD test set using residue-level contact maps derived in two ways: 1) from the predicted AF3 structure, and 2) from the contact probabilities from the distogram head. The results from both methods are in good agreement, indicating that either of them can be used to obtain contact maps from AF3 predictions (Table S7).

AF2 performs better than AF3

Similar to Disobind, coarse-graining improves the performance of both AF2 and AF3 for both contact map and interface predictions (Table 1). Across all tasks, AF2 performs better than AF3. Moreover, AF3 predictions have lower confidence as measured by their ipTM scores (Fig S3). For 35 of 52 OOD test set entries, the ipTM score of the best AF2 model was greater than that of the best AF3 model. Overall, only 8 of 52 predictions from AF2 and 2 of 52 predictions from AF3 had a high confidence prediction, *i.e.*, ipTM score higher than 0.75.

Combining Disobind and AlphaFold2 predictions

As AF2 performed better than AF3, we asked if combining the predictions from Disobind and AF2 could further improve over either method. We combine the predictions from Disobind and AF2 using a logical OR operation and evaluate the performance on the OOD test set. The combined model, “Disobind+AF2”, performs better than either of the methods (Table 2, see also Supplementary Text).

Performance by residue types

Next, we assess the performance of Disobind, AF2, and Disobind+AF2 by residue type to examine which regions each method tends to perform better on (See Methods). Disobind outperforms AF2 on both the contact map and interface residue

prediction in disordered regions, whereas AF2 performs slightly better than Disobind for ordered regions (Table S8). Importantly, the combined model Disobind+AF2 performs better for both disordered and ordered regions, likely because the two models learn complementary information by focusing on different regions.

Further, Disobind predictions are more accurate than AF2 for disorder-promoting and hydrophobic, whereas AF2 and Disobind perform similarly for aromatic and polar residues (Table S9). Disobind is more accurate than AF2 in predicting contacts and interface residues involving LIPs (linear interacting peptides) (Table S9). The combined model is better across all residue types (Table S9).

Lastly, we assess the performance of Disobind, AF2, and Disobind+AF2 on the 17 OOD complexes containing long IDRs, *i.e.,* complexes where 100% of the residues in the IDR fragment are annotated as disordered (Table S10). Disobind and AF2 perform comparably at contact map prediction, while Disobind performs slightly better at interface prediction. Disobind+AF2 achieves the best results overall for both contact map and interface residue prediction (Table S10).

Comparison with interface predictors for IDRs

We compared Disobind with partner-independent interface predictors for IDRs, including AIUPred, MORFchibi, and DeepDISOBind, which are comparable to the state-of-the-art in the Critical Assessments of protein Intrinsic Disorder prediction (CAID2) [25,48–50]. For a fair comparison, we evaluated interface residue predictions for only the IDR protein in the OOD set (See Methods). Disobind outperforms AIUPred, MORFchibi, and DeepDISOBind at predicting interface residues for the IDR across all coarse-grained resolutions (Table 3).

Evaluating Disobind for protein-protein interaction prediction

We evaluated Disobind for predicting protein-protein interactions (PPI), *i.e.* to predict if two input protein sequences will bind or not. We used the five PPI test datasets for IDRs provided by [51]. The dataset comprises UniProt accessions for interacting and non-interacting protein pairs. This includes a total of 9993 protein pairs. The dataset was pre-processed by removing protein pairs for which the UniProt accessions were part of the Disobind training set, the UniProt sequence could not be downloaded, and the sequence length was greater than 10000 residues. This resulted in a total of 2355 protein pairs. Disobind interface residue predictions at coarse-grained resolution 1 were obtained for all the protein pairs. To obtain the PPI predictions from interface residue predictions, we aggregate the residue-level predictions using two ways: (i) maximum sigmoid score across both proteins, and (ii) average sigmoid score across both proteins, and evaluate the performance using the Area Under the Receiver Operator Curve (AUROC). Disobind achieves an AUROC of 0.48 and 0.5, for the two ways, respectively, indicating a random prediction.

Input Embeddings

Next, we investigate the effect of using embeddings from various pLMs and the different embedding types on the performance of Disobind.

Protein language models allow for a shallow architecture

Protein language models (pLMs) provide context-aware representations of protein sequences that facilitate training models for downstream tasks like contact map prediction *via* transfer learning. They allow using a shallow downstream architecture when training with limited data [22,52]. Several pLMs have been developed including the ESM series of models [53], ProstT5 [54], ProtT5 [44], ProtBERT [44], and ProSE [55].

The pLMs differ in the model architecture, the training datasets, and the training regime. These differences can affect their ability to learn contextual representations for protein sequences. ProSE uses a bidirectional LSTM trained on UniRef protein

sequences and structures with a multitask loss combining similarity and contact prediction objectives [55]. The other pLMs use variants of the transformer architecture. ProtBERT uses the Bidirectional Encoder Representations from Transformers (BERT) architecture, an encoder-only model, trained on the UniRef100 and BFD100 datasets. It is trained to reconstruct corrupted tokens given the context of the neighbouring tokens [44]. ProtT5 uses the Text-to-Text Transfer Transformer (T5) architecture, an encoder-decoder model, trained on the UniRef50 and BFD100 datasets. In contrast to the BERT-style training objective, ProtT5 is trained to reconstruct spans of corrupted tokens [44,56]. ProstT5 is a newer version of ProtT5, fine-tuned with AlphaFold2 predicted protein structures [57].

Disobind uses embeddings from ProtT5, one of the best performing pLMs in our comparison (Table S11). The T5-based models, ProtT5 and ProstT5, perform similarly, but outperform the BERT-styled ProtBERT and LSTM-based ProSE models. Due to limitations in memory, we could not generate ESM2-650M-global embeddings for the entire training set. Hence, we do not include ESM in our comparison.

Global embeddings perform better than local embeddings

We then evaluate whether models trained with global sequence embeddings perform better than those with local embeddings. Local embeddings are obtained by providing the sequence of the fragment as input to the pLM. Global embeddings, in contrast, are obtained by providing the entire protein sequence corresponding to the fragment as input to the pLM, followed by extracting the embeddings corresponding to the fragment (Fig S4). Models with global embeddings perform better than those with local embeddings across all the pLMs we tested (Table S11). Using the global embeddings provides the context of the complete protein sequence corresponding to the fragment, plausibly resulting in better performance. This context may contain information relevant to binding: for example, the binding of IDRs is known to be affected by the flanking regions [8].

Case studies

IDRs enable dynamic interactions, molecular recognition, and functional adaptability, making them a crucial component of various signaling and regulatory pathways. We used Disobind+AF2 for predicting interface residues for several IDR complexes, as it performs better than the other models (Fig 4, Fig S5). Here, we discuss in detail two cases in which the IDR becomes ordered upon binding (DOR) and two cases where the IDR remains disordered (DDR). A larger set of NMR examples of mostly DDR cases from MobiDB, PED, BMRB, and LLPSdb are also shown separately (Fig S5) [35,58–60].

The cellular prion protein (PrP^{C}) plays an important role in several cellular processes, including synapse formation, regulating circadian rhythm, and maintaining ion homeostasis (Kraus et al., 2021; Sigurdson et al., 2019). It is known to convert to an infectious PrP^{Sc} form, which forms amyloid fibrils associated with several diseases, including Creutzfeldt-Jakob disease and bovine spongiform encephalopathy (Kraus et al., 2021; Sigurdson et al., 2019). PrP^{C} has an N-terminal disordered region and a primarily α-helical C-terminal region, whereas the conversion to PrP^{Sc} involves significant structural changes, resulting in a β-sheet-rich structure. Disobind+AF2 could predict the interface residues in the β-sheet-rich PrP^{Sc} dimer with an F1-score of 0.71 (Fig 4a).

Chemokines are small secretary molecules that play several important roles, including immune cell trafficking, wound healing, and lymphoid cell development (Rossi & Zlotnik, 2000; Sepuru et al., 2020). Dysregulation of chemokine release is associated with several diseases, including chronic pancreatitis, inflammatory bowel disease, and psoriasis (Rossi & Zlotnik, 2000; Sepuru et al., 2020). Chemokines bind to chemokine receptors that belong to the G-protein coupled receptor (GPCR) family. Tha latter contains an N-terminal disordered region that becomes structured and adopts an extended conformation upon binding. Disobind + AF2 reliably predicts the interface residues for the chemokine Interleukin-8 binding to the N-terminus of its cognate receptor, CXCR1 with an F1 score of 0.73 (Fig 4b).

The death domain-associated protein 6 (Daxx) is a transcriptional coregulator that plays a key role in apoptosis [61,62]. It contains a C-terminal disordered region that attains an extended conformation upon binding to its partner SUMO-1 [61]. This interaction is crucial for its localization to nuclear PML bodies, the absence of which leads to disruption of PML nuclear bodies as seen in acute myelocytic leukaemia [62]. Disobind+AF2 could reliably capture interface residues in this region, achieving an F1-score of 0.81 (Fig 4c).

UV-stimulated scaffold protein A (UVSSA) is a crucial component for transcription-coupled nucleotide excision repair (TC-NER) pathway [63,64]. Disruption of the TC-NER pathway is associated with various diseases, including xeroderma pigmentosa and UV-sensitive syndrome [63]. In complex with USP7, UVSSA is involved in processing the stalled RNA pol II and recruiting TFIIH. A short acidic region in UVSSA is an IDR and interacts with the N-terminal PH domain in the p62 subunit of the TFIIH complex, which is crucial for the recruitment of TFIIH to the TCR site [63]. Disobind+AF2 was able to capture the interface residues in this case with an F1-score of 0.71 (Fig 4d). In summary, across all the cases, our combined model achieves high accuracy in predicting interface residues for IDR complexes.

Discussion

Here, we discuss about the uses and limitations of Disobind, challenges involved, the performance of AlphaFold, and future directions. Our combined model, Disobind+AF2, can be readily extended to any current or future structure prediction method that provides confidence metrics such as pLDDT and PAE for the predicted structures. Predictions from Disobind+AF2 may be used to improve the localization of IDRs in integrative models of large assemblies, our primary motivation for developing the method. Macromolecular assemblies contain significant portions of IDRs, for example, the Fg Nups in the nuclear pore complex, the MBD3-IDR in the nucleosome remodeling and deacetylase (NuRD) complex, and the N-terminus of Plakophilin1 in the desmosome [65–67]. These regions typically lack data for structural modeling, resulting in integrative models of poor precision [66,67]. The predictions from Disobind+AF2 can be used in integrative modeling methods such as IMP,

HADDOCK, and Assembline as inter-protein distance restraints [68–70]. Further, the predicted contacts can be combined with molecular dynamics (MD) simulations to provide mechanistic insights into their dynamic behaviour. Additionally, our method can be used to characterize IDR-mediated interactions across proteomes, helping identify new binding motifs linked to subcellular localization or function. It may also guide modulation of IDR interactions, for instance, by suggesting plausible mutations.

However, our method also has several limitations. First, it is limited to binary IDR-partner complexes. For complexes formed by three or more subunits, binary predictions from Disobind would need to be combined. Second, it assumes that the IDR and its partner are known to bind. Disobind cannot reliably distinguish binders from non-binders. Third, the accuracy of the predictions depends on the ability of the pLM to provide accurate representations of the IDR and its partner. Finally, our method cannot be used to assess the effects of post-translational modifications as the pLM used does not distinguish post-translationally modified amino acids.

The aforementioned limitations in part arise due to the challenges in predicting contact maps and interface residues for IDRs in complex with a partner. First, a limited number of experimental structures are available for IDRs in complexes [22,38]. Second, IDRs adopt an ensemble of conformations, and the available structures may only partially capture this conformational diversity. Third, inter-protein contact maps are typically sparse, with only a few residues forming contacts. Although not sufficient, we gather all available structures for IDRs in complexes. With these structures, we create a dataset of merged binary complexes for training Disobind. Using merged binary complexes helps overcome the issue of sparsity associated with training on an ensemble of contact maps. Predicting interface residues instead of contacts and using coarse-graining helps overcome the challenges associated with the multivalency of IDRs and the sparsity of inter-protein contact maps (Fig 3a).

Several studies indicate that the low-confidence regions in AF2 predictions overlap with the presence of disordered regions, although these low-confidence regions cannot be considered as a representative conformation for IDRs [28,29,71]. Despite the success of AF2 and AF3 for ordered proteins, their performance on IDR complexes

remains limited, as shown in benchmarks [19,28,30,72]. This could potentially be due to several reasons. First, IDRs show low sequence conservation, resulting in poor quality MSAs [7]. Second, IDRs are best described by an ensemble of conformations which capture their dynamic behaviour. AF2 and AF3, however, predict a single best structure, which may not be representative of IDRs [29]. Further, the diversity of binding modes of IDRs in complexes, *e.g.,* DOR and DDR, poses a significant challenge. In particular, AlphaFold predictions are less accurate for fuzzy complexes or DDR complexes, compared to DOR complexes [28,30]. Third, the confidence metrics of AF2 and AF3 may not reliably reflect structural accuracy [73,74]. Particularly, the presence of IDRs is known to negatively impact the confidence metrics (Dunbrack, 2025; Magana & Kovalevskiy, 2024; Omidi et al., 2024; Varga et al., 2024).

On our OOD test set, most of the predictions from AF2 and AF3 had low confidence, with an ipTM score lower than 0.75. Notably, compared to AF2, AF3 predictions were less accurate and had lower confidence. It is possible that the confidence cutoffs used for ordered proteins do not apply to IDRs and predictions involving the latter might require different cutoffs. A recent benchmark suggested a lower ipTM score of 0.4 for assessing interfaces involving IDRs [30]. Further, the per-residue metrics such as pLDDT and PAE may be more relevant for assessing AF2 and AF3 predictions than the global metrics such as the ipTM and pTM, as is also suggested by a recent benchmark and the AF3 guide [18,30]. Thus, AlphaFold confidence metrics might need to be interpreted differently for IDRs, and this requires rigorous benchmarking [76–78].

AF2 has been shown to successfully predict the structures of protein-peptide complexes and domain-motif complexes, although the results could be sensitive to the sequence fragment size, fragment delimitation, and the MSA alignment mode [19,21]. Some studies show that MSA subsampling [79], clustering the sequences in the MSA [80], and using sliding fragments as input to AF2 [19] result in better predictions and can be used to predict multiple conformations. In our comparison, we did not explore these strategies, though they may improve the model predictions.

We conclude by highlighting potential future directions. One of the major roadblocks in training methods such as Disobind is the lack of data. More experimental

structures of IDRs in complexes would be valuable [22,38]. Whereas IDR ensembles derived from MD can be used, generating MD ensembles for IDR complexes is computationally expensive and challenging, and the existing databases like PED contain very few such ensembles for IDR complexes [58,81]. Alternatively, deep generative models can be used to generate ensembles for IDRs in complexes [81–83]. However, the current methods are limited to generating ensembles for monomers.

Compared to contact maps, distograms model the probability distribution across multiple distance ranges for each residue pair and thus provide a richer representation of structural heterogeneity for IDRs. However, predicting distograms for IDRs presents some non-trivial challenges. The limited structural data available for IDRs in complexes introduces strong biases toward a few observed conformations, making it difficult to learn representative distance distributions that capture the dynamic nature of IDRs. Furthermore, a distogram representation exacerbates the data sparsity problem due to the increased dimensionality of the predicted output from $\mathbb{R}^{N\times N}$ for a contact map to $\mathbb{R}^{N\times N\times d}$ in a distogram, where N is the protein length and d is the number of distance bins. Thus, developing methods to predict distograms using sparse data for IDR complexes presents a promising avenue for future research.

Disobind and similar methods can be further enhanced by improving the existing pLMs to provide better representations for IDRs plausibly by incorporating physical priors and/or structural information [81,84,85]. Additionally, these methods could be extended to predict protein-protein interactions (PPIs) involving IDRs, and aid in the design of IDR binders. The behaviour of IDRs within cells remains largely unexplored. Methods like Disobind are expected to facilitate an improved understanding of the interactions, function, and modulation of IDRs.

Resource Availability

Lead Contact: Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Shruthi Viswanath (shruthiv@ncbs.res.in).

Materials Availability: This study did not generate new materials.

Data and Code Availability:

- Source data statement: The datasets created in this manuscript have been deposited to Zenodo at https://doi.org/10.5281/zenodo.14504762. The deposition contains all datasets (train, dev, ID, OOD) along with the input sequences and target contact maps, AF2 and AF3 predictions for the OOD test set, and the PDB structures, SIFTS mappings, PDB API files, and Uniprot Sequences used to create the datasets.
- Code statement: The Disobind model and scripts to create datasets, train the model, use Disobind/Disobind+AF2, and perform analysis are available at GitHub https://github.com/isblab/disobind.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contacts upon request.

Acknowledgement

We thank ISB Lab members Shreyas Arvindekar, Muskaan Jindal, Omkar Golatkar, and Mubashira KP for their useful comments on the manuscript. We also thank Aditi S. Pathak, Revathy Menon, Sarthak Joshi, Jagannath Mondal, Anand Srivastava, Vinothkumar KR, and Shachi Gosavi for valuable suggestions on the manuscript. We thank Omkar Golatkar for help in preparing Fig. 1. Molecular graphics images were produced using the UCSF Chimera and UCSF ChimeraX packages from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR001081, NIH R01-GM129325, and National Institute of Allergy and Infectious Diseases).

Author contribution

Conceptualization: K.M., S.V.

Reading and synthesis: K.M., S.V.

Data curation: K.M.
Methodology: K.M., V.U., S.V.
Writing – original draft: K.M.
Writing – review & editing: K.M., V.U., S.V.
Resources, Supervision, Funding acquisition: S.V.

Funding

This work has been supported by the following grants: Department of Atomic Energy (DAE) TIFR grant RTI 4006, Department of Science and Technology (DST) SERB grant SPG/2020/000475, and Department of Biotechnology (DBT) grant BT/PR40323/BTIS/137/78/2023 from the Government of India to S.V.

Declaration of interests

None declared.

Diversity and inclusion statement

One or more of the authors in this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors in this paper self-identifies as a gender minority in their field of research.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used ChatGPT to refine the wording. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Supplemental information titles and legends

Supplementary material PDF – Figures S1-S5 and Tables S1-S11.

Table S1: Hyperparams.xlsx – Excel file containing hyperparameters for model training.

Main figures and legends

Figure 1: Disobind dataset creation pipeline. We compiled our dataset by gathering structures of IDR-containing complexes from DIBS, MFIB, FuzDB, PDBtot, PDBcdr, DisProt, IDEAL, and MobiDB. For each PDB chain, missing residues in the structures are excluded, resulting in sequence fragments. Further, we restrict the maximum length of a sequence fragment to 100 residues. We use annotations from DisProt, IDEAL, and MobiDB to classify a sequence fragment as an IDR and non-IDR. Next, we obtain inter-chain binary complexes comprising at least one IDR fragment (IDR with an ordered protein or both IDRs). Contact maps are obtained for all binary complexes which are subsequently grouped based on the UniProt accessions of the sequence fragment pair. The contact maps of all binary complexes in a group are merged, creating merged contact maps. A sequence fragment pair and its corresponding merged contact map comprise a merged binary complex. The merged binary complexes are redundancy reduced at a 20% sequence identity cutoff to create an out-of-distribution (OOD) test set. The remaining complexes comprise the Disobind training set which is split into train, dev, and in-distribution (ID) test sets.

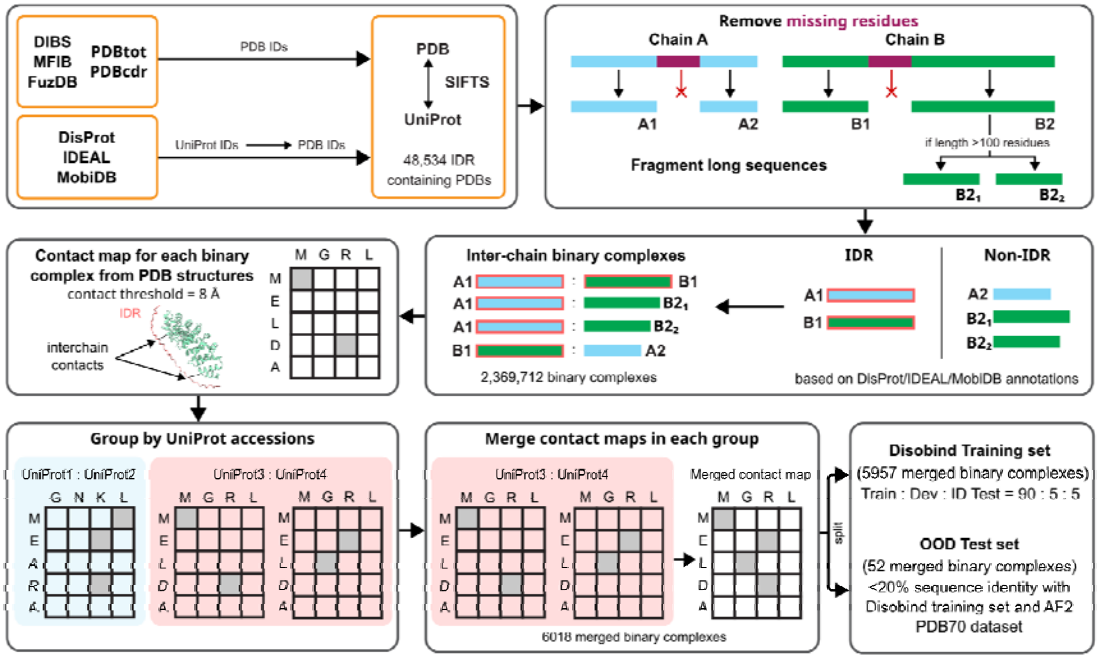


Figure 2: Disobind architecture. The inputs to Disobind are the sequence embeddings for the sequence pair obtained using a pLM. The sequence embeddings are projected to a lower dimension using a projection block. Next, in the interaction block, the projected embeddings are used to compute the outer product and outer difference, which is subsequently concatenated along the feature dimension, resulting in an interaction tensor. For contact map prediction, this interaction tensor is processed by a multi-layer perceptron (MLP). For interface residue prediction, the interaction tensor is reduced by computing row-wise and column-wise averages and further processed by an MLP. Finally, an output block provides element-wise sigmoid scores, with a score greater than 0.5 representing a contact or interface residue. Here, N, L, C represent the batch size, sequence length, and feature dimension respectively.

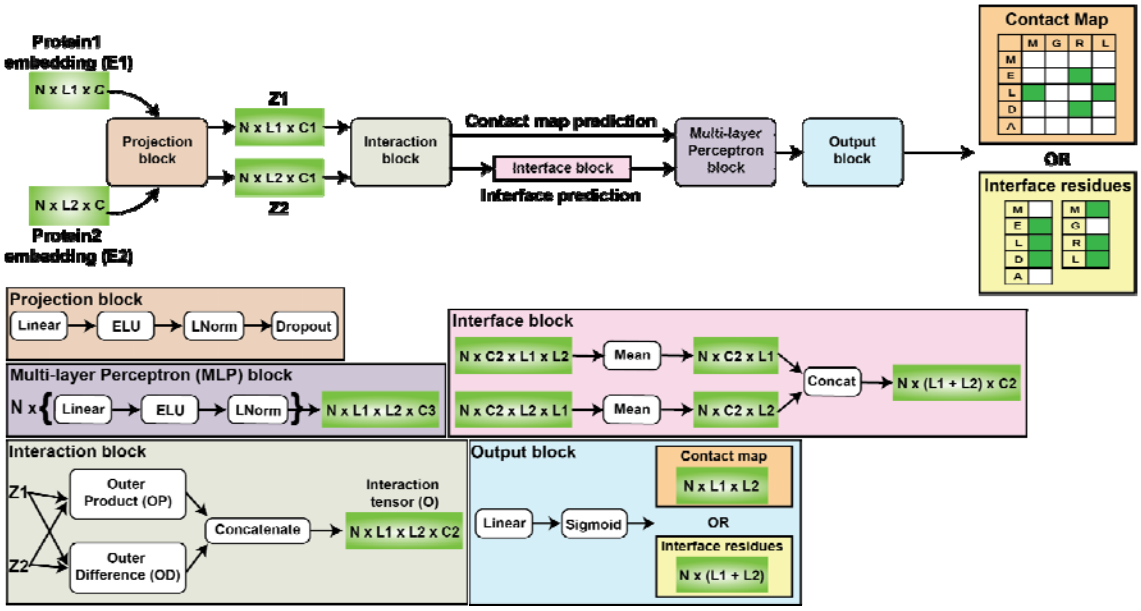


Figure 3: Predicting coarse-grained contact maps and interface residues. (a) We use coarse-graining and predict interface residues as an alternative to predicting residue-wise inter-protein contact maps. This addresses the problems associated with the multivalency of IDRs and the sparsity of inter-protein contact maps. **(b)** Disobind can predict inter-protein contact maps and interface residues at multiple coarse-grained resolutions (1, 5, 10).

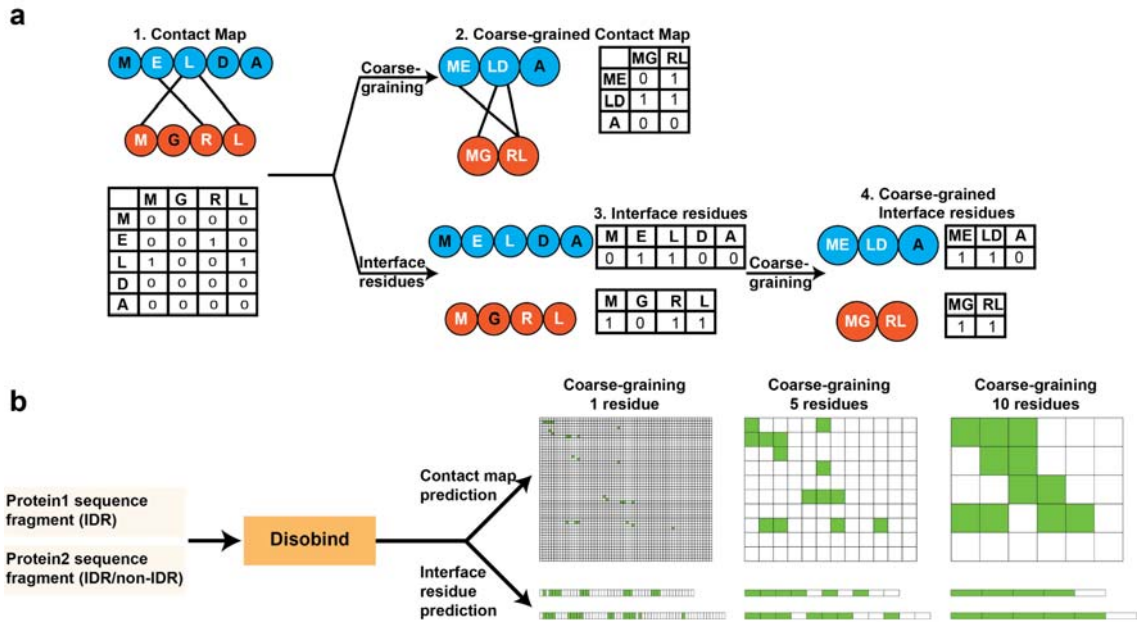
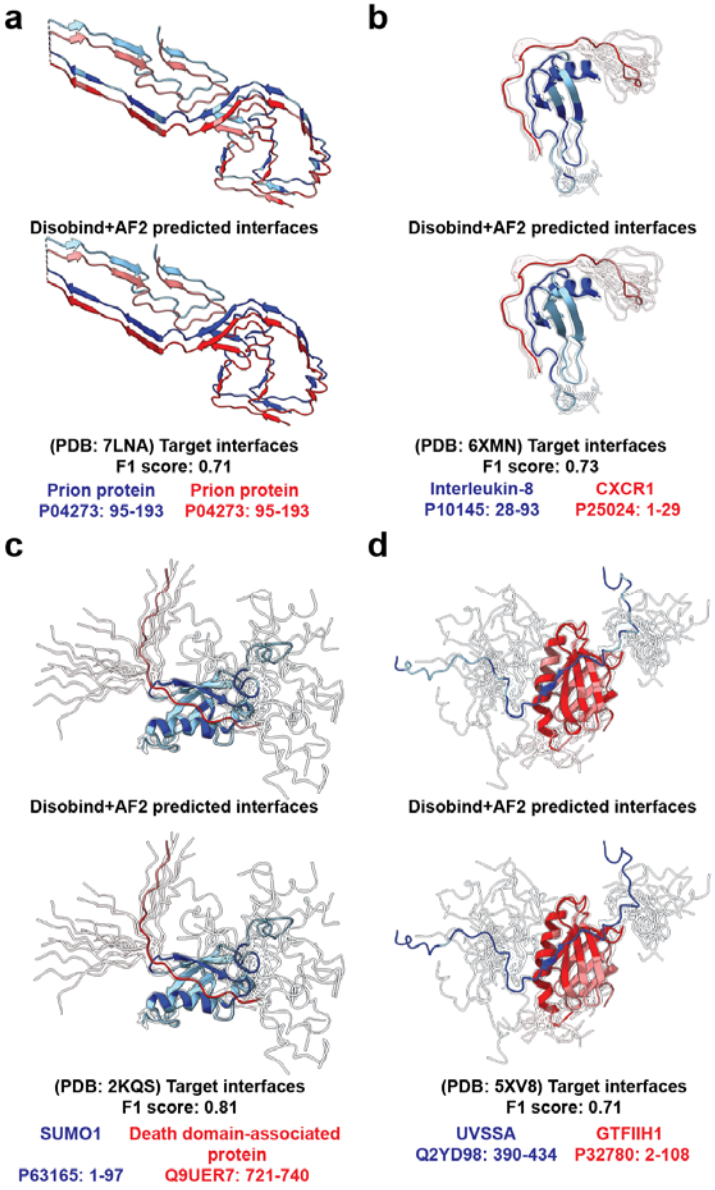


Figure 4: Case studies visualizing Disobind+AF2 predictions. The figures visualize the Disobind+AF2 predicted and target interface residues for both proteins in complexes where IDRs undergo a DOR (panels a and b) or DDR (panels c and d) transition. The protein chains are coloured in blue and red, where the lighter shade represents the chain and the darker shade represents the interface residues for the respective chains. For all NMR structures, the ensemble is represented with reduced transparency in the background, whereas the first model is shown in the foreground. The PDB IDs, protein names, and corresponding UniProt accession, along with the residue numbers for which the predictions were made, are shown for all complexes.



Tables

Table 1: Comparing Disobind, AF2, and AF3 performance. We compare the performances of Disobind, AF2, and AF3 on the OOD set for contact map and interface residue prediction for different coarse-grained (CG) resolutions. We also include a random baseline (RB) for comparison. For AF2 and AF3 we only consider the confident predictions (pLDDT >= 70 and PAE <= 5) at multiple ipTM cutoffs (0, 0.4, 0.75) for evaluation. The number after the '_' represents the ipTM cutoff used.

Model	Recall	Precision	F1 score	Recall	Precision	F1 score
	Contact map			**Interface**		
	Coarse Graining 1					
Disobind	0.22	0.63	0.33	0.5	0.46	0.48
RB	0.51	0.01	0.02	0.5	0.26	0.34
AF2_0.75	0.12	0.58	0.2	0.18	0.76	0.29
AF3_0.75	0.02	0.41	0.04	0.02	0.43	0.04
AF2_0.4	0.2	0.45	0.28	0.35	0.7	0.47
AF3_0.4	0.08	0.33	0.13	0.18	0.6	0.27
AF2_0.0	0.2	0.45	0.28	0.35	0.7	0.47
AF3_0.0	0.08	0.31	0.13	0.18	0.57	0.27
	Coarse Graining 5					
Disobind	0.38	0.61	0.47	0.73	0.56	0.63
RB	0.52	0.08	0.13	0.53	0.45	0.49
AF2_0.75	0.15	0.58	0.24	0.2	0.79	0.32
AF3_0.75	0.02	0.34	0.03	0.02	0.5	0.04
AF2_0.4	0.29	0.52	0.37	0.43	0.76	0.54
AF3_0.4	0.14	0.47	0.22	0.22	0.72	0.34
AF2_0.0	0.29	0.52	0.37	0.43	0.76	0.54

AF3_0.0	0.14	0.45	0.22	0.23	0.7	0.34
	Coarse Graining 10					
Disobind	0.42	0.49	0.45	0.8	0.61	0.69
RB	0.54	0.17	0.26	0.48	0.52	0.5
AF2_0.75	0.16	0.62	0.25	0.2	0.83	0.32
AF3_0.75	0.02	0.4	0.04	0.03	0.59	0.05
AF2_0.4	0.32	0.57	0.41	0.44	0.8	0.57
AF3_0.4	0.16	0.52	0.25	0.24	0.78	0.36
AF2_0.0	0.32	0.57	0.41	0.44	0.8	0.57
AF3_0.0	0.16	0.51	0.25	0.24	0.75	0.37

Table 2: Evaluating the performance of Disobind+AF2 on the OOD test set. We combine Disobind and AF2 predictions using a logical OR operation and evaluate the performance on the OOD test set for contact map and interface residue prediction. For AF2 we consider the confident predictions (pLDDT >= 70 and PAE <= 5) for evaluation. No ipTM cutoff is applied for AF2 and Disobind+AF2 predictions.

Model	Recall	Precision	F1 score	Recall	Precision	F1 score
	Contact map			Interface		
	Coarse Graining 1					
Disobind	0.22	0.63	0.33	0.5	0.46	0.48
AF2_0.0	0.2	0.45	0.28	0.35	0.7	0.47
Disobind+AF2	0.33	0.49	0.4	0.6	0.46	0.52
	Coarse Graining 5					
Disobind	0.38	0.61	0.47	0.73	0.56	0.63
AF2_0.0	0.29	0.52	0.37	0.43	0.76	0.54
Disobind+AF2	0.49	0.5	0.5	0.78	0.54	0.64
	Coarse Graining 10					
Disobind	0.42	0.49	0.45	0.8	0.61	0.69
AF2_0.0	0.32	0.57	0.41	0.44	0.8	0.57
Disobind+AF2	0.55	0.46	0.5	0.86	0.61	0.71

Table 3: Comparison with interface predictors for IDRs. We compare the performance of Disobind with interface predictors for IDRs, including AIUPred, MORFchibi, and DeepDISOBind, on the IDR sequences from the OOD test set.

Model	Recall	Precision	F1 score
	Interface residue prediction, CG1		
Disobind	0.5	0.49	0.5
Aiupred	0.36	0.38	0.37
Morfchibi	0.97	0.24	0.39
Deepdisobind	0.22	0.45	0.29
	Interface residue prediction, CG5		
Disobind	0.73	0.56	0.63
Aiupred	0.43	0.58	0.49
Morfchibi	0.96	0.41	0.58
Deepdisobind	0.21	0.73	0.33
	Interface residue prediction, CG10		
Disobind	0.8	0.61	0.69
Aiupred	0.43	0.61	0.51
Morfchibi	0.95	0.5	0.66
Deepdisobind	0.2	0.79	0.31

STAR Methods

Resource Availability

Lead Contact: Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Shruthi Viswanath (shruthiv@ncbs.res.in).

Materials Availability: This study did not generate new materials.

Data and Code Availability:

- Source data statement: The datasets created in this manuscript have been deposited to Zenodo at https://doi.org/10.5281/zenodo.14504762. The deposition contains all datasets (train, dev, ID, OOD) along with the input sequences and target contact maps, AF2 and AF3 predictions for the OOD test set, and the PDB structures, SIFTS mappings, PDB API files, and Uniprot Sequences used to create the datasets.
- Code statement: The Disobind model and scripts to create datasets, train the model, use Disobind/Disobind+AF2, and perform analysis are available at GitHub https://github.com/isblab/disobind.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contacts upon request.

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
NA		
NA		
NA		
NA		
Bacterial and virus strains		
NA		
NA		
NA		
NA		
Biological samples		
NA		
NA		
NA		
NA		
Chemicals, peptides, and recombinant proteins		
NA		
NA		
NA		
NA		
Critical commercial assays		
NA		
NA		
NA		
NA		
Deposited data		
DIBS	Schad E. et. al. (2018) Bioinform.	https://dibs.enzim.ttk.mta.hu/
MFIB	Fichó E. et. al. (2017) Bioinform.	https://mfib.pbrg.hu/
FuzDB	Miskei M. et. al. (2017) NAR	https://fuzdb.org/
PDBtot, PDBcdr	Miskei M. et. al. (2020) JMB	https://doi.org/10.1016/j.jmb.2020.02.017
MobiDB	Piovesan D. et. al. (2024) NAR	https://mobidb.org/

DisProt	Aspromonte M.C. et. al. (2024) NAR	https://disprot.org/
IDEAL	Fukuchi S. et. al. (2014) NAR	https://www.ideal-db.org/
PDB	Berman H.M. et. al. (2000) NAR	https://www.rcsb.org/
UniProt	The UniProt Consortium (2025) NAR	https://www.uniprot.org/
Experimental models: Cell lines		
NA		
NA		
NA		
NA		
Experimental models: Organisms/strains		
NA		
NA		
NA		
NA		
NA		
Oligonucleotides		
NA		
NA		
NA		
NA		
Recombinant DNA		
NA		
NA		
NA		
NA		
Software and algorithms		
SIFTS	Velanker S. et. al. (2013) NAR	10.1093/nar/gks1258
AlphaFold2	Jumper J. et. al. (2021) Nature	https://github.com/google-deepmind/alphafold
AlphaFold3 server	Abramson J. et. al. (2024) Nature	https://alphafoldserver.com/
MMSeqs2	Steinegger M. et. al. (2017) Nat. Biotechnol.	https://github.com/soedinglab/MMseqs2.git
Pytorch v 2.0.1		

CUDA v 11.8		
AIUPred	Erdős G et. al. (2025) JMB	https://doi.org/10.1093/nar/gkae385
MORFchibi 2.0	Malhis N and Gsponer J (2025) bioRxiv	https://doi.org/10.1101/2025.01.31.635962
DeepDISOBind	Zhang F et al (2021) Brief. in Bioinform.	https://doi.org/10.1093/bib/bbab521
ProtT5, ProtBERT	Elnaggar A. et. al. (2023) IEEE PAMI	https://github.com/agemagician/ProtTrans.git
ProstT5	Heinzinger M. et. al. (2024) NAR genom. bioinform.	https://github.com/mheinzinger/ProstT5.git
ProSE	Bepler and Berger (2019) ICLR	https://github.com/tbepler/prose.git
Disobind	This study	https://doi.org/10.5281/zenodo.14504762
Other		
Disobind GitHub	This study	https://github.com/isblab/disobind
Disobind Zenodo	This study	https://doi.org/10.5281/zenodo.14504762

Experimental model and study participant details

Not applicable

Method details

Creating the dataset for Disobind

Gathering PDB structures of IDRs in complexes

First, we gathered the available PDB entries for protein complexes containing IDRs from databases on structures of disordered proteins: DIBS [36], MFIB [33], and FuzDB [9] (Fig 1). We also included PDB entries from the PDBtot and PDBcdr datasets [10]. PDBtot consists of IDRs that either undergo a DOR or DDR transition, whereas PDBcdr consists of IDRs that undergo both DOR and DDR transitions depending upon the partner. We considered all PDB entries associated with each database. This resulted in a set of 4252 PDB entries. PED is another source of conformational ensembles of IDRs [58]. However, PED entries were not used: several PED entries were already incorporated into our pipeline and the remaining entries correspond to monomers.

We then supplemented this set by querying the PDB for additional complexes containing IDRs. For this, we first obtained Uniprot identifiers (IDs) of proteins containing IDRs from DisProt [32], IDEAL [34], and MobiDB [35] (Fig 1). Specifically, from Disprot, we obtained all UniProt IDs corresponding to each DisProt entry. From IDEAL, we obtained UniProt IDs for entries annotated as "verified ProS". These sequences are experimentally verified to be disordered in isolation and ordered upon binding. From MobiDB, we obtained UniProt IDs for entries annotated "curated-disorder-priority", "homology-disorder-priority", "curated-lip-priority", and "homology-lip-priority". These annotations correspond to sequences with curated experimental evidence for disorder ("curated-disorder"), their homologs ("homology-disorder"), sequences with experimental evidence for disorder with binding motifs such as SLIMs ("curated-lip"), and their homologs ("homology-lip"), respectively. Homology-based annotations were incorporated due to the lack of sufficient experimentally

curated data available for IDRs. Importantly, the annotations were based on the source with the highest confidence ("priority"), and annotations based on indirect sources of evidence or predictions were not considered. Choosing the highest confidence annotation may reduce false positive "disorder" or "lip" annotations that may arise from homology-based evidence.

A total of 357734 unique Uniprot IDs across MobiDB, Disprot, and IDEAL were obtained. The PDB REST API was queried to obtain all the PDB entries associated with a given Uniprot ID, resulting in 48534 PDB entries (Table S1) [86]. Combining these with the PDB entries obtained in the previous step, we obtained 50294 unique PDB entries. We note that MobiDB now provides the PDB IDs for IDR-containing complexes, however, this functionality was not available when this project started.

Defining binary complexes containing IDRs

The above PDB entries were downloaded using the Python requests library and mapped to the sequences in UniProt using the SIFTS (Structure Integration with Function, Taxonomy, and Sequence) tool [87]. Entries with an obsolete PDB ID and those lacking a SIFTS mapping were removed, and deprecated PDB IDs were replaced with superseding PDB IDs. Further, entities corresponding to chimeric proteins or non-protein molecules, those having an obsolete UniProt ID, and those containing more than 10000 residues were removed.

Next, for each PDB entry, we obtained sequence fragments by removing missing residues from the sequence of each chain. The length of these fragments was kept to between 20 and 100 residues. Fragments longer than 100 residues were further divided due to constraints on GPU memory, that limited the scalability of the model architecture. We then identified the disordered residues in these fragments (Fig 1). This was achieved by cross-referencing the UniProt mapping of the fragment with the disorder annotations from DisProt, IDEAL, and MobiDB as obtained above; a residue was considered disordered if it was annotated as such in any of these three databases. Fragments comprising at least 20% disordered residues were considered as IDR fragments whereas the others were considered as non-IDR fragments. This cutoff balances an over-representation of ordered residues while ensuring enough

training data. This is also consistent with proteome-wide analyses which indicate that proteins with IDRs have 13–28% disordered residues. Specifically, annotations based on experiments indicate a 17% disorder fraction [35].

Subsequently, a set of binary complexes was constructed from each PDB entry. Each binary complex consisted of an IDR fragment paired with another IDR or non-IDR fragment. Binary complexes comprising solely of intra-chain fragments or non-IDR fragments were excluded (Fig 1). This resulted in a total of 2369712 binary complexes.

Creating merged binary complexes

We then created contact maps from these binary complexes. A contact map is a binary matrix of zeros (non-contacts) and ones (contacts); two residues are in contact if the distance between their $C\alpha$ atoms is less than 8 Å. We merged the contact maps across different binary complexes of a sequence fragment pair, to create “merged contact maps” (Fig 1). This allows us to account for contacts from all available complexes. Further, a merged contact map is less sparse and therefore easier to learn compared to an ensemble of contact maps. It also mitigates the bias caused by having different numbers of structures for different sequence fragment pairs.

We first grouped the binary complexes across all the PDB entries by the UniProt ID of the constituent proteins in the sequence fragment pair, resulting in 14599 UniProt ID pairs/groups. For computational efficiency, all binary complexes with no contacts were eliminated. The remaining binary complexes in a group were sorted based on the UniProt start residue positions of the protein fragments in the pair and further partitioned into sets. Each set comprised binary complexes containing overlapping sequences (more than one residue overlap) for both fragments such that the sequence length for each fragment, merged across the overlapping sequences, does not exceed the maximum fragment length (100 residues). Contact maps for all binary complexes in a set were aggregated using a logical OR operation to form “merged contact maps”. A sequence fragment pair with the corresponding merged contact map comprises a merged binary complex. We obtained 24442 merged binary

complexes. Finally, we filtered merged binary complexes for which the contact density was very high (>5%) or very low (<0.5%), resulting in 6018 merged binary complexes (Fig 1).

Creating the training set and OOD test set

We created an out-of-distribution (OOD) test set that is sequence non-redundant with our training dataset and the PDB70 dataset used for training AlphaFold. We first removed merged binary complexes whose sequence fragments were derived from a PDB chain in PDB70. We then obtained a non-redundant set of sequences by clustering our dataset of sequence fragments and the PDB70 cluster representatives using MMSeqs2 [88]. MMSeqs2 was used at a 20% sequence identity threshold and cluster mode 1. Since the PDB70 dataset is already clustered at 40% sequence similarity, we consider only the representative cluster members for sequence identity comparisons.

For the OOD test set, we selected clusters without any chain from PDB70 that are singleton (only one cluster member *i.e.* self) or doublet (two member sequences, both belonging to the same merged binary complex) clusters. This resulted in 53 merged binary complexes for the OOD test set. We removed an OOD test set entry for which AF2 predictions could not be obtained, resulting in 52 OOD test set complexes. The remaining dataset was split into a training (train), development (dev), and an in-distribution (ID) test set in the ratio 0.9:0.05:0.05, resulting in 5362 train, 298 dev, and 297 ID test set merged complexes.

Disobind Model Architecture and Training

Notations

Here, we define the notation used in this paper.

$\mathbb{R}^{A \times B \times C \times D}$: 4D tensor with the axes i, j, k, l corresponding to the four dimensions of size A, B, C, D respectively

N: batch size used for training.

L_1, L_2: lengths of protein sequence fragments 1 and 2 indicating the number of residues in each protein.

C, C_1, C_2, C_3: the sizes of embedding or feature dimensions.

E_1, E_2: embeddings for protein sequence fragments 1 and 2.

O: interaction tensor obtained post-concatenation.

O_{red}: output from the multi-layer perceptron block, reduced from the interaction tensor.

I_1, I_2 interface tensors for proteins 1 and 2.

I: concatenated interface tensors.

$Linear$: applies a linear transformation, $Linear(X) = XW^T + b$.

ELU: exponential linear unit, $ELU(x) = max(x, \alpha(e^x - 1))$.

$Sigmoid(x) : \frac{1}{1+e^{-x}}$.

$Dropout$: regularization technique that randomly sets weights to 0.

$LNorm$: layer normalization.

$Concat([A, B]_i)$: concatenate tensors A, B along the axis i.

$A_{ijkl \to ikjl}$: permute tensor A along the specified axes.

Inputs and outputs for training

We train six separate models, corresponding to six prediction tasks. These tasks correspond to two types of predictions: a) inter-protein contact map and b) interface residue predictions at three coarse-grained (CG) resolutions (1, 5, and 10 sequence-contiguous residues). Coarse-graining at resolution 1 is equivalent to no coarse-graining, *i.e.*, residue-level predictions. The input to the model is the embedding for the two sequence fragments from the ProtT5 pLM $E_1 \in \mathbb{R}^{L_1 \times C}, E_2 \in \mathbb{R}^{L_2 \times C}$ [44]. The outputs for the two types of predictions respectively are a) a binary contact map, $ContactMap \in \mathbb{R}^{L_1/k \times L_2/k}$ and b) a binary interface residue vector, $Interface \in \mathbb{R}^{L_1/k + L_2/k}$ at a particular CG resolution, where $k \in [1, 5, 10]$ represents the CG resolution.

For coarse-grained predictions, the input embeddings were coarse-grained using an AvgPool1d whereas the corresponding contact maps were coarse-grained using a MaxPool2d; the kernel size and stride were set to the CG resolution in both cases.

The target vectors for interface residue predictions were obtained by reducing the rows and columns of the corresponding contact maps by an OR operation. The input

embeddings, output contact maps, and interface residue vectors were zero-padded to the maximum length (100 residues).

Model architecture

Disobind comprises a projection block, an interface block, a multi-layer perceptron (MLP) block, and an output block (Fig 2). The models for contact map and interface residue prediction employ a contact map block and interface block respectively.

Projection block

The input embeddings E_1, E_2 are projected to a lower dimension, C_1 using a linear layer.

$Z_1 = Dropout(\, ELU(\, Linear(\, E_1 \,)\,)\,)\,;\; Z_1 \in \mathbb{R}^{N \times L_1 \times C_1}$ – (1)

$Z_2 = Dropout(\, ELU(\, Linear(\, E_2 \,)\,)\,)\,;\; Z_2 \in \mathbb{R}^{N \times L_2 \times C_1}$ – (2)

Interaction block

The model captures complementarity between the projected embeddings Z_1, Z_2 by concatenating the outer product and the absolute value of the outer difference between the projected embeddings along the feature dimension.

$O = Concat(\, [OuterProduct(\, Z_1, Z_2 \,), |OuterDifference(\, Z_1, Z_2 \,)|]_l);\; O \in \mathbb{R}^{N \times L_1 \times L_2 \times C_2}$ – (3)

Interface block

To capture the interface residues for both proteins I, we compute row-wise and column-wise averages followed by concatenating along the length.

$I_1 = Mean(O_{ijkl \to iljk});\; I_1 \in \mathbb{R}^{N \times C_2 \times L_1}$ – (4)

$I_2 = Mean(O_{ijkl \to ilkj});\; I_2 \in \mathbb{R}^{N \times C_2 \times L_2}$ – (5)

$I = Concat([I_1, I_2]_k);\; I_2 \in \mathbb{R}^{N \times C_2 \times (L_1 + L_2)}$ – (6)

$I = I_{ijk \to ikj};\; I \in \mathbb{R}^{N \times (L_1 + L_2) \times C_2}$ – (7)

Multi-layer perceptron (MLP) block

The feature dimension of the interaction tensor O is progressively decreased using a multilayer perceptron (MLP) to obtain O_{red}.

$O_{red} = m \times \{ LNorm(ELU(Linear(x)))\}$; $O_{red} \in \mathbb{R}^{N \times L_1 \times L_2 \times C_3}$ – (8)

We used a 3-layer ($m = 3$) MLP for contact map prediction whereas it was not used for interface residue prediction ($m = 0$).

Output block

It comprises a $Linear$ layer followed by a $Sigmoid$ activation.

$ContactMap = Sigmoid(Linear(O_{red}))$; $ContactMap \in \mathbb{R}^{N \times L_1 \times L_2}$ – (9)

$Interface = Sigmoid(Linear(I))$; $Interface \in \mathbb{R}^{N \times (L_1 + L_2)}$ – (10)

Training

The models were trained using PyTorch version 2.0.1 and CUDA version 11.8, using an NVIDIA A6000 GPU. The model was trained using the Singularity Enhanced (SE) loss function [37], a modified form of the binary cross-entropy (BCE) loss used for class-imbalanced training, with $\alpha = 0.9, \beta = 3$ and AdamW (amsgrad variant) optimizer with a weight decay of 0.05. We used an exponential learning rate scheduler for slowly decaying the base learning rate.

$$SELoss = \sum \quad (\alpha \cdot y \, log(\hat{y}) + (1 - \alpha) \cdot (1 - y) \, log((1 - \hat{y})) \cdot (1 + \hat{y})^{\beta}) \text{ – (10)}$$

A binary mask was used to exclude the padding from the loss calculation. The train and dev set loss were monitored to check for over/under-fitting (Fig S1).

Evaluating predictions

The Disobind predicted score was converted to a binary output using a threshold of 0.5. The model performance was evaluated on the train, dev, and ID test set using recall, precision, and F1-score. All metrics were calculated using the torchmetrics library.

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP} \qquad F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (11)$$

Hyperparameter tuning and ablations

Here, we explore the effect of select hyperparameters on the model performances in the dev and ID test sets. Hyperparameters were tuned for the contact map and interface prediction at CG 1 and minimally modified for the CG 5 and CG 10 models. The final set of hyperparameters is provided (Table S1).

Projection dimension

We tuned the projection dimension C_1 for the projection block (Table S2).

Number of layers in the MLP

$Linear$ layers in the MLP were used to either upsample (upsampling layers, US) or downsample (downsampling layers, DS) the feature dimension by a factor of 2. We tuned the number of US (DS) layers. Additionally, we removed the MLP for interface residue prediction (Table S3).

α, β for SE loss

We tuned the α and β parameters for SE loss; these parameters weigh the contribution of the contact and non-contact elements to the loss (Table S4).

pLM embeddings

We compared the model performance using embeddings from ProtT5 $(C = 1024)$, ProstT5 $(C = 1024)$, ProSE $(C = 6165)$, and ProtBERT [44,54,55] $(C = 1024)$.

Further, we compared two types of embeddings: global and local embeddings (Fig S4, Table S11). Local embeddings were obtained using the sequence of the fragment as input to the pLM. In contrast, global embeddings were obtained by using the complete UniProt sequence as input to the pLM and extracting fragment embedding from it.

Assessment

Random baseline

To compute random baseline predictions, we predicted random contacts and interface residues based on their frequencies in the training set. For the prediction tasks involving coarse-graining, these frequencies were determined from the coarse-grained contact maps and coarse-grained interface residues in the training set.

AlphaFold predictions on the OOD test set

We used a local AF2 installation and the AF3 webserver for obtaining OOD test set predictions [18,27]. AF2 was run with the default arguments as specified in their GitHub repository. Due to the failure of AF2 to relax several of the OOD test set entries we switch off the Amber relaxation step and consider the unrelaxed models for further evaluation. Additionally, we removed an OOD test set entry for which we were unable to obtain the AF2 prediction. We considered the best models from AF2 (based on the ipTM+pTM score) and AF3 (based on the ranking score) for comparison with Disobind. The ipTM score was used to assess the overall model confidence and the pLDDT and PAE metrics assessed the per-residue confidence. pLDDT scores higher than 70 and PAE values lower than 5 are considered as confident [89]. An ipTM score higher than 0.8 represents a high-confidence prediction whereas an ipTM score lower than 0.6 represents a likely incorrect prediction; predictions with ipTM scores between 0.6-0.8 require further validation [18,27,45,46]. Given that only five AF2 predictions and none of the AF3 predictions in the OOD test set have an ipTM score greater than 0.8, we initially applied an ipTM cutoff of 0.75 as recommended in benchmarks on AlphaFold2 [45,46]. Later we relaxed the ipTM cutoffs to 0.4 and 0 (*i.e.*, no ipTM cutoff).

To compare the AlphaFold outputs to those from our models, we first derive the binary contact maps from the best-predicted model. We then apply binary masks on this output to ignore low-confidence interactions (pLDDT<70 and PAE>5). Both the pLDDT and PAE masks are applied on the CG 1 contact map obtained from AF2 and AF3 before proceeding further. To apply an ipTM cutoff, we zero the contact maps for AlphaFold predictions with an ipTM lower than the cutoff. The binary contact map was zero-padded to the maximum length (100 residues). The interface

residues and coarse-grained predictions were derived from the CG 1 contact map as mentioned previously (See Inputs and Outputs for training).

Disobind+AF2 predictions

The Disobind+AF2 predictions were obtained by combining the corresponding predictions from Disobind and AF2 using a logical OR operation. We ignore low-confidence AF2 predictions having a pLDDT less than 70 and a PAE greater than 5. No ipTM cutoff was used.

Performance by residue type

Given a protein sequence, disordered regions were identified by combining DisProt, IDEAL, and MobiDB annotations as described in Creating dataset for Disobind section. Residues in these regions were annotated as “disordered residues”, the remaining residues in the sequence were considered “ordered”. Linear interacting peptide (LIP) annotations were obtained from MobiDB as described earlier. We categorized residues into disorder-promoting amino acids (R, P, Q, E, G, S, A, K), aromatic amino acids (F, Y, W), hydrophobic amino acids (A, V, L, I, P, M, F, W), and polar amino acids (S, T, C, N, Q, Y, D, E, K, R, H) [90,91]. For LIPs, we evaluated contacts between residues in LIPs and any residue on the partner protein. For all other residue types, we evaluated contacts between residues of the same type. To assess the predictions on specific residue types, we applied a binary mask on the predictions and the target to ignore interface residues (contacts) not involving the relevant residues (residue pairs). We only considered confident predictions from AF2 having pLDDT scores higher than 70 and PAE values lower than 5. No ipTM cutoff was used.

Comparison with interface predictors for IDRs

We selected state-of-the-art interface predictors including AIUPred, MORFchibi, and DeepDISOBind for comparison to Disobind on the OOD test set [25,49,50]. These methods have been shown to be comparable to the state-of-the-art in CAID2 [48]. They provide partner-independent interface residue predictions. For comparison to Disobind, we assess their interface residue predictions on the IDR sequence of the

binary complexes in the OOD set. AIUPred predictions were obtained using programmatic access as mentioned in their GitHub repository. For MORFchibi and DeepDISOBind, predictions were obtained using their respective webservers, using defaults. DeepDISOBind provides binary interface predictions. AIUPred predictions are converted to binary interfaces using a 0.5 threshold as specified in the paper [49]. For MORFchibi, we use a 0.77 threshold for interface residues as recommended; interface residue regions shorter than four residues are eliminated [50].

Quantification and statistical analysis

Not applicable

Additional resources

Not applicable

References

1. Oldfield CJ, Dunker AK. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. Annu Rev Biochem [Internet]. 2014 Jun 2 [cited 2024 May 26];83(Volume 83, 2014):553–84. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev-biochem-072711-164947

2. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999 Oct 22;293(2):321–31.

3. Lindorff-Larsen K, Kragelund BB. On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins. J Mol Biol. 2021 Oct 1;433(20):167196.

4. Uversky VN. The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. FEBS Lett. 2013 Jun 27;587(13):1891–901.

5. Fonin AV, Darling AL, Kuznetsova IM, Turoverov KK, Uversky VN. Intrinsically disordered proteins in crowded milieu: when chaos prevails within the cellular gumbo. Cell Mol Life Sci [Internet]. 2018 Nov 1 [cited 2023 Nov 1];75(21):3907–29. Available from: https://doi.org/10.1007/s00018-018-2894-9

6. Tompa P, Davey NE, Gibson TJ, Babu MM. A million peptide motifs for the molecular biologist. Mol Cell. 2014 Jul 17;55(2):161–9.

7. Holehouse AS, Kragelund BB. The molecular basis for cellular function of intrinsically disordered protein regions. Nat Rev Mol Cell Biol [Internet]. 2024 Mar [cited 2025 Jun 20];25(3):187–211. Available from: https://www.nature.com/articles/s41580-023-00673-0

8. Olsen JG, Teilum K, Kragelund BB. Behaviour of intrinsically disordered proteins in protein–protein complexes with an emphasis on fuzziness. Cell Mol Life Sci [Internet]. 2017 Sep 1 [cited 2024 May 26];74(17):3175–83. Available from: https://doi.org/10.1007/s00018-017-2560-7

9. Miskei M, Antal C, Fuxreiter M. FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. Nucleic Acids Res [Internet]. 2017 Jan 4 [cited 2023 Feb 14];45(D1):D228–35. Available from: https://doi.org/10.1093/nar/gkw1019

10. Miskei M, Horvath A, Vendruscolo M, Fuxreiter M. Sequence-Based Prediction of Fuzzy Protein Interactions. J Mol Biol [Internet]. 2020 Mar 27 [cited 2023 Feb 14];432(7):2289–303. Available from: https://www.sciencedirect.com/science/article/pii/S002228362030190X

11. Orand T, Jensen MR. Binding mechanisms of intrinsically disordered proteins: Insights from experimental studies and structural predictions. Curr Opin Struct Biol [Internet]. 2025 Feb 1 [cited 2025 Jan 17];90:102958. Available from: https://www.sciencedirect.com/science/article/pii/S0959440X24001854

12. Christoffer C, Kihara D. IDP-LZerD: Software for Modeling Disordered Protein Interactions. In: Kihara D, editor. Protein Structure Prediction [Internet]. New York, NY: Springer US; 2020 [cited 2024 Oct 15]. p. 231–44. Available from: https://doi.org/10.1007/978-1-0716-0708-4_13

13. Dai B, Bailey-Kellogg C. Protein interaction interface region prediction by geometric deep learning. Bioinformatics [Internet]. 2021 Sep 1 [cited 2023 Feb 14];37(17):2580–8. Available from: https://doi.org/10.1093/bioinformatics/btab154

14. Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods [Internet]. 2020 Feb [cited 2023 May 7];17(2):184–92. Available from: https://www.nature.com/articles/s41592-019-0666-6

15. Guo Z, Liu J, Skolnick J, Cheng J. Prediction of inter-chain distance maps of protein complexes with 2D attention-based deep neural networks. Nat Commun [Internet]. 2022 Nov 15 [cited 2024 Oct 15];13(1):6963. Available from: https://www.nature.com/articles/s41467-022-34600-2

16. Lin P, Tao H, Li H, Huang SY. Protein–protein contact prediction by geometric triangle-aware protein language models. Nat Mach Intell [Internet]. 2023 Nov [cited 2024 Oct 15];5(11):1275–84. Available from: https://www.nature.com/articles/s42256-023-00741-2

17. Pittala S, Bailey-Kellogg C. Learning context-aware structural representations to predict antigen and antibody binding interfaces. Bioinforma Oxf Engl. 2020 Jul 1;36(13):3996–4003.

18. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature [Internet]. 2024 Jun [cited 2024 Jun 26];630(8016):493–500. Available from: https://www.nature.com/articles/s41586-024-07487-w

19. Bret H, Gao J, Zea DJ, Andreani J, Guerois R. From interaction networks to interfaces, scanning intrinsically disordered regions using AlphaFold2. Nat Commun [Internet]. 2024 Jan 18 [cited 2024 Oct 15];15(1):597. Available from: https://www.nature.com/articles/s41467-023-44288-7

20. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature [Internet]. 2021 Aug 26 [cited 2023 Jun 29];596(7873):583–9. Available from: https://www.nature.com/articles/s41586-021-03819-2

21. Lee BG, Merkel F, Allegretti M, Hassler M, Cawood C, Lecomte L, et al. Cryo-EM structures of holo condensin reveal a subunit flip-flop mechanism. Nat Struct Mol Biol [Internet]. 2020 Aug [cited 2024 Jan 20];27(8):743–51. Available from: https://www.nature.com/articles/s41594-020-0457-x

22. Jahn LR, Marquet C, Heinzinger M, Rost B. Protein Embeddings Predict Binding Residues in Disordered Regions [Internet]. bioRxiv; 2024 [cited 2024 May 24]. p. 2024.03.05.583540. Available from: https://www.biorxiv.org/content/10.1101/2024.03.05.583540v2

23. Mészáros B, Simon I, Dosztányi Z. Prediction of Protein Binding Regions in Disordered Proteins. PLOS Comput Biol [Internet]. 2009 May 1 [cited 2023 Jun

16];5(5):e1000376. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000376

24. Singh J, Litfin T, Singh J, Paliwal K, Zhou Y. SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model. Bioinformatics [Internet]. 2022 Mar 28 [cited 2024 Oct 15];38(7):1888–94. Available from: https://doi.org/10.1093/bioinformatics/btac053

25. Zhang F, Zhao B, Shi W, Li M, Kurgan L. DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning. Brief Bioinform [Internet]. 2022 Jan 1 [cited 2024 Oct 15];23(1):bbab521. Available from: https://doi.org/10.1093/bib/bbab521

26. Ginell GM, Emenecker RJ, Lotthammer JM, Keeley AT, Plassmeyer SP, Razo N, et al. Sequence-based prediction of intermolecular interactions driven by disordered regions. Science [Internet]. 2025 May 22 [cited 2025 Jun 11];388(6749):eadq8381. Available from: https://www.science.org/doi/10.1126/science.adq8381

27. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer [Internet]. bioRxiv; 2022 [cited 2023 Oct 29]. p. 2021.10.04.463034. Available from: https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2

28. Alderson TR, Pritišanac I, Moses AM, Forman-Kay JD. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2 [Internet]. bioRxiv; 2022 [cited 2023 May 7]. p. 2022.02.18.481080. Available from: https://www.biorxiv.org/content/10.1101/2022.02.18.481080v1

29. Ruff KM, Pappu RV. AlphaFold and Implications for Intrinsically Disordered Proteins. J Mol Biol [Internet]. 2021 Oct 1 [cited 2024 May 29];433(20):167208. Available from: https://www.sciencedirect.com/science/article/pii/S0022283621004411

30. Omidi A, Møller MH, Malhis N, Bui JM, Gsponer J. AlphaFold-Multimer

accurately captures interactions and dynamics of intrinsically disordered protein regions. Proc Natl Acad Sci [Internet]. 2024 Oct 29 [cited 2024 Nov 17];121(44):e2406407121. Available from: https://www.pnas.org/doi/10.1073/pnas.2406407121

31. Verburgt J, Zhang Z, Kihara D. Multi-level analysis of intrinsically disordered protein docking methods. Methods [Internet]. 2022 Aug 1 [cited 2024 Oct 15];204:55–63. Available from: https://www.sciencedirect.com/science/article/pii/S1046202322001207

32. Aspromonte MC, Nugnes MV, Quaglia F, Bouharoua A, DisProt Consortium, Tosatto SCE, et al. DisProt in 2024: improving function annotation of intrinsically disordered proteins. Nucleic Acids Res [Internet]. 2024 Jan 5 [cited 2024 May 16];52(D1):D434–41. Available from: https://doi.org/10.1093/nar/gkad928

33. Fichó E, Reményi I, Simon I, Mészáros B. MFIB: a repository of protein complexes with mutual folding induced by binding. Bioinformatics [Internet]. 2017 Nov 15 [cited 2023 Feb 14];33(22):3682–4. Available from: https://doi.org/10.1093/bioinformatics/btx486

34. Fukuchi S, Amemiya T, Sakamoto S, Nobe Y, Hosoda K, Kado Y, et al. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. Nucleic Acids Res [Internet]. 2014 Jan 1 [cited 2023 Jun 5];42(D1):D320–5. Available from: https://doi.org/10.1093/nar/gkt1010

35. Piovesan D, Del Conte A, Mehdiabadi M, Aspromonte MC, Blum M, Tesei G, et al. MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins. Nucleic Acids Res [Internet]. 2024 Oct 29 [cited 2024 Dec 11];gkae969. Available from: https://doi.org/10.1093/nar/gkae969

36. Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mészáros B. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. Bioinforma Oxf Engl. 2018 Feb 1;34(3):535–7.

37. Si Y, Yan C. Improved protein contact prediction using dimensional hybrid residual networks and singularity enhanced loss function. Brief Bioinform. 2021 Nov 5;22(6):bbab341.

38. Zhang J, Durham J, Qian Cong. Revolutionizing protein–protein interaction prediction with deep learning. Curr Opin Struct Biol [Internet]. 2024 Apr 1 [cited 2024 Nov 17];85:102775. Available from: https://www.sciencedirect.com/science/article/pii/S0959440X24000022

39. Arvindekar S, Pathak AS, Majila K, Viswanath S. Optimizing representations for integrative structural modeling using Bayesian model selection. Martelli PL, editor. Bioinformatics [Internet]. 2024 Mar 4 [cited 2024 May 29];40(3):btae106. Available from: https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btae106/7613065

40. Noid WG. Perspective: Coarse-grained models for biomolecular systems. J Chem Phys. 2013 Sep 7;139(9):090901.

41. Christensen LF, Staby L, Bugge K, O'Shea C, Kragelund BB, Skriver K. Evolutionary conservation of the intrinsic disorder-based Radical-Induced Cell Death1 hub interactome. Sci Rep [Internet]. 2019 Dec 12 [cited 2025 Jun 29];9(1):18927. Available from: https://www.nature.com/articles/s41598-019-55385-3

42. Cumberworth A, Lamour G, Babu MM, Gsponer J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochem J [Internet]. 2013 Aug 29 [cited 2025 Jun 29];454(3):361–9. Available from: https://doi.org/10.1042/BJ20130545

43. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, et al. Analysis of Molecular Recognition Features (MoRFs). J Mol Biol [Internet]. 2006 Oct 6 [cited 2025 Jun 29];362(5):1043–59. Available from: https://www.sciencedirect.com/science/article/pii/S0022283606009831

44. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al.

ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell. 2022 Oct;44(10):7112–27.

45. O'Reilly FJ, Graziadei A, Forbrig C, Bremenkamp R, Charles K, Lenz S, et al. Protein complexes in cells by AI-assisted structural proteomics. Mol Syst Biol [Internet]. 2023 Apr 12 [cited 2023 Oct 29];19(4):e11544. Available from: https://www.embopress.org/doi/full/10.15252/msb.202311544

46. Yin R, Feng BY, Varshney A, Pierce BG. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. Protein Sci Publ Protein Soc. 2022 Aug;31(8):e4379.

47. Magana PG, Kovalevskiy GO. AlphaFold - A practical guide [Internet]. 2024 [cited 2024 Nov 17]. Available from: https://www.ebi.ac.uk/training/online/courses/alphafold

48. Conte AD, Mehdiabadi M, Bouhraoua A, Miguel Monzon A, Tosatto SCE, Piovesan D. Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. Proteins Struct Funct Bioinforma [Internet]. 2023 [cited 2025 Jun 11];91(12):1925–34. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26582

49. Erdős G, Deutsch N, Dosztányi Z. AIUPred – Binding: Energy Embedding to Identify Disordered Binding Regions. J Mol Biol [Internet]. 2025 Aug 1 [cited 2025 Jun 11];437(15):169071. Available from: https://www.sciencedirect.com/science/article/pii/S0022283625001378

50. Malhis N, Gsponer J. Predicting molecular recognition features in protein sequences with MoRFchibi 2.0 [Internet]. bioRxiv; 2025 [cited 2025 Jun 11]. p. 2025.01.31.635962. Available from: https://www.biorxiv.org/content/10.1101/2025.01.31.635962v1

51. Perovic V, Sumonja N, Marsh LA, Radovanovic S, Vukicevic M, Roberts SGE, et al. IDPpi: Protein-Protein Interaction Analyses of Human Intrinsically Disordered Proteins. Sci Rep [Internet]. 2018 Jul 12 [cited 2025 Jun

20];8(1):10563. Available from: https://www.nature.com/articles/s41598-018-28815-x

52. Bepler T, Berger B. Learning the protein language: Evolution, structure, and function. Cell Syst [Internet]. 2021 Jun 16 [cited 2023 Jun 7];12(6):654-669.e3. Available from: https://www.sciencedirect.com/science/article/pii/S2405471221002039

53. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci [Internet]. 2021 Apr 13 [cited 2024 May 24];118(15):e2016239118. Available from: https://www.pnas.org/doi/10.1073/pnas.2016239118

54. Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Mirdita M, Steinegger M, et al. Bilingual language model for protein sequence and structure. NAR Genomics Bioinforma [Internet]. 2024 Dec 1 [cited 2024 Dec 11];6(4):lqae150. Available from: https://doi.org/10.1093/nargab/lqae150

55. Bepler T, Berger B. Learning protein sequence embeddings using information from structure [Internet]. arXiv; 2019 [cited 2024 May 24]. Available from: http://arxiv.org/abs/1902.08661

56. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [Internet]. arXiv; 2023 [cited 2025 Jun 11]. Available from: http://arxiv.org/abs/1910.10683

57. Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Steinegger M, Rost B. ProstT5: Bilingual Language Model for Protein Sequence and Structure [Internet]. bioRxiv; 2023 [cited 2024 Dec 11]. p. 2023.07.23.550085. Available from: https://www.biorxiv.org/content/10.1101/2023.07.23.550085v1

58. Ghafouri H, Lazar T, Del Conte A, Tenorio Ku LG, PED Consortium, Tompa P, et al. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. Nucleic Acids Res. 2024 Jan

5;52(D1):D536–44.

59. Hoch JC, Baskaran K, Burr H, Chin J, Eghbalnia HR, Fujiwara T, et al. Biological Magnetic Resonance Data Bank. Nucleic Acids Res [Internet]. 2023 Jan 6 [cited 2025 Jun 14];51(D1):D368–76. Available from: https://doi.org/10.1093/nar/gkac1050

60. Li Q, Peng X, Li Y, Tang W, Zhu J, Huang J, et al. LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro. Nucleic Acids Res [Internet]. 2020 Jan 8 [cited 2025 Jun 14];48(D1):D320–7. Available from: https://doi.org/10.1093/nar/gkz778

61. Chang CC, Naik MT, Huang YS, Jeng JC, Liao PH, Kuo HY, et al. Structural and Functional Roles of Daxx SIM Phosphorylation in SUMO Paralog-Selective Binding and Apoptosis Modulation. Mol Cell [Internet]. 2011 Apr 8 [cited 2025 Jun 11];42(1):62–74. Available from: https://www.sciencedirect.com/science/article/pii/S1097276511001638

62. Salomoni P, Khelifi AF. Daxx: death or survival protein? Trends Cell Biol [Internet]. 2006 Feb 1 [cited 2025 Jun 11];16(2):97–104. Available from: https://www.sciencedirect.com/science/article/pii/S0962892405003107

63. Okuda M, Nakazawa Y, Guo C, Ogi T, Nishimura Y. Common TFIIH recruitment mechanism in global genome and transcription-coupled repair subpathways. Nucleic Acids Res [Internet]. 2017 Dec 15 [cited 2025 Jun 11];45(22):13043–55. Available from: https://doi.org/10.1093/nar/gkx970

64. Schwertman P, Lagarou A, Dekkers DHW, Raams A, van der Hoek AC, Laffeber C, et al. UV-sensitive syndrome protein UVSSA recruits USP7 to regulate transcription-coupled repair. Nat Genet [Internet]. 2012 May [cited 2025 Jun 11];44(5):598–602. Available from: https://www.nature.com/articles/ng.2230

65. Akey CW, Singh D, Ouch C, Echeverria I, Nudelman I, Varberg JM, et al. Comprehensive structure and functional adaptations of the yeast nuclear pore complex. Cell [Internet]. 2022 Jan [cited 2024 Jan 30];185(2):361-378.e25. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867421014537

66. Arvindekar S, Jackman MJ, Low JKK, Landsberg MJ, Mackay JP, Viswanath S. Molecular architecture of nucleosome remodeling and deacetylase sub complexes by integrative structure determination. Protein Sci [Internet]. 2022 Sep [cited 2024 Jan 30];31(9):e4387. Available from: https://onlinelibrary.wiley.com/doi/10.1002/pro.4387

67. Pasani S, Menon KS, Viswanath S. The molecular architecture of the desmosomal outer dense plaque by integrative structural modeling. Protein Sci [Internet]. 2024 [cited 2024 Nov 17];33(12):e5217. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.5217

68. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc. 2003 Feb 19;125(7):1731–7.

69. Rantos V, Karius K, Kosinski J. Integrative structural modeling of macromolecular complexes using Assembline. Nat Protoc [Internet]. 2022 Jan [cited 2023 Oct 31];17(1):152–76. Available from: https://www.nature.com/articles/s41596-021-00640-z

70. Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, et al. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. PLoS Biol [Internet]. 2012 Jan 17 [cited 2023 Jun 22];10(1):e1001244. Available from: https://dx.plos.org/10.1371/journal.pbio.1001244

71. Escobedo N, Saldaño T, Donagh JM, Sawicki LR, Palopoli N, Alberti SF, et al. Ghost interactions: revealing missing protein-ligand interactions using AlphaFold predictions [Internet]. bioRxiv; 2023 [cited 2024 May 29]. p. 2023.10.18.561916. Available from: https://www.biorxiv.org/content/10.1101/2023.10.18.561916v1

72. Lee CY, Hubrich D, Varga JK, Schäfer C, Welzel M, Schumbera E, et al. Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation. Mol Syst Biol [Internet]. 2024 Feb 2 [cited 2024 Oct 15];20(2):75–97. Available from: https://www.embopress.org/doi/full/10.1038/s44320-023-00005-6

73. Elofsson A. AlphaFold3 at CASP16 [Internet]. bioRxiv; 2025 [cited 2025 Jun 10]. p. 2025.04.10.648174. Available from: https://www.biorxiv.org/content/10.1101/2025.04.10.648174v1

74. Guan L, Keating AE. How AlphaFold and related models predict protein-peptide complex structures [Internet]. bioRxiv; 2025 [cited 2025 Jun 26]. p. 2025.06.18.660495. Available from: https://www.biorxiv.org/content/10.1101/2025.06.18.660495v1

75. Dunbrack RL. Rēs ipSAE loquunt: What's wrong with AlphaFold's ipTM score and how to fix it [Internet]. bioRxiv; 2025 [cited 2025 Jun 22]. p. 2025.02.10.637595. Available from: https://www.biorxiv.org/content/10.1101/2025.02.10.637595v1

76. Varga JK, Ovchinnikov S, Schueler-Furman O. actifpTM: a refined confidence metric of AlphaFold2 predictions involving flexible regions [Internet]. arXiv; 2024 [cited 2025 Jan 18]. Available from: http://arxiv.org/abs/2412.15970

77. Chakravarty D, Lee M, Porter LL. Proteins with alternative folds reveal blind spots in AlphaFold-based protein structure prediction. Curr Opin Struct Biol [Internet]. 2025 Feb 1 [cited 2025 Jan 10];90:102973. Available from: https://www.sciencedirect.com/science/article/pii/S0959440X24002008

78. Kim AR, Hu Y, Comjean A, Rodiger J, Mohr SE, Perrimon N. Enhanced Protein-Protein Interaction Discovery via AlphaFold-Multimer [Internet]. bioRxiv; 2024 [cited 2025 Jan 18]. p. 2024.02.19.580970. Available from: https://www.biorxiv.org/content/10.1101/2024.02.19.580970v1

79. del Alamo D, Sala D, Mchaourab HS, Meiler J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. Robertson JL, Swartz KJ, Robertson JL, editors. eLife [Internet]. 2022 Mar 3 [cited 2024 Nov 17];11:e75751. Available from: https://doi.org/10.7554/eLife.75751

80. Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature [Internet]. 2024 Jan [cited 2024 Nov 17];625(7996):832–9.

Available from: https://www.nature.com/articles/s41586-023-06832-9

81. Majila K, Arvindekar S, Jindal M, Viswanath S. Frontiers in integrative structural biology: modeling disordered proteins and utilizing in situ data [Internet]. arXiv; 2024 [cited 2024 Oct 13]. Available from: https://arxiv.org/abs/2407.00566

82. Janson G, Feig M. Transferable deep generative modeling of intrinsically disordered protein conformations [Internet]. 2024 [cited 2024 May 21]. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.02.08.579522

83. Mansoor S, Baek M, Park H, Lee GR, Baker D. Protein Ensemble Generation Through Variational Autoencoder Latent Space Sampling. J Chem Theory Comput [Internet]. 2024 Apr 9 [cited 2024 May 24];20(7):2689–95. Available from: https://doi.org/10.1021/acs.jctc.3c01057

84. Rogers JR, Nikolényi G, AlQuraishi M. Growing ecosystem of deep learning methods for modeling protein–protein interactions. Protein Eng Des Sel [Internet]. 2023 Jan 1 [cited 2024 Nov 17];36:gzad023. Available from: https://doi.org/10.1093/protein/gzad023

85. Wang D, Pourmirzaei M, Abbas UL, Zeng S, Manshour N, Esmaili F, et al. S-PLM: Structure-aware Protein Language Model via Contrastive Learning between Sequence and Structure. BioRxiv Prepr Serv Biol. 2024 May 13;2023.08.06.552203.

86. Rose Y, Duarte JM, Lowe R, Segura J, Bi C, Bhikadiya C, et al. RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. J Mol Biol. 2021 May 28;433(11):166704.

87. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res [Internet]. 2013 Jan 1 [cited 2023 Jun 12];41(D1):D483–9. Available from: https://doi.org/10.1093/nar/gks1258

88. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol [Internet]. 2017

Nov [cited 2024 Dec 11];35(11):1026–8. Available from: https://www.nature.com/articles/nbt.3988

89. Edmunds NS, Genc AG, McGuffin LJ. Benchmarking of AlphaFold2 accuracy self-estimates as indicators of empirical model quality and ranking: a comparison with independent model quality assessment programmes. Bioinformatics [Internet]. 2024 Aug 1 [cited 2024 Dec 11];40(8):btae491. Available from: https://doi.org/10.1093/bioinformatics/btae491

90. Uversky VN. The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins. Intrinsically Disord Proteins [Internet]. 2013 Jan 1 [cited 2025 Jun 14];1(1):e24684. Available from: https://doi.org/10.4161/idp.24684

91. Uversky VN. Intrinsically Disordered Proteins and Their “Mysterious” (Meta)Physics. Front Phys [Internet]. 2019 Feb 7 [cited 2024 Jan 30];7:10. Available from: https://www.frontiersin.org/article/10.3389/fphy.2019.00010/full