Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning

Alex X Lu[1], Amy X Lu[2], Iva Pritišanac[3,4], Taraneh Zarin[5],
Julie D Forman-Kay[4,6]. Alan M Moses[1,3]

1 Department of Computer Science, University of Toronto, Toronto, Canada
2 Department of Computer Science, University of Toronto, Toronto, Canada
(Present Address: Department of Electrical Engineering and Computer Sciences, Berkeley)
3 Department of Cell and Systems Biology, University of Toronto, Canada
4 Program in Molecular Medicine, Hospital for Sick Children, Toronto, Canada
5 Department of Cell and Systems Biology, University of Toronto, Canada
(Present Address: Systems Biology Program, Center for Genomic Regulation, Barcelona, Spain)
6 Department of Biochemistry, University of Toronto, Canada

Abstract

A major challenge to the characterization of intrinsically disordered regions (IDRs), which are widespread in the proteome, but relatively poorly understood, is the identification of molecular features, such as short motifs, amino acid repeats and physicochemical properties that mediate the functions of these regions. Here, we introduce a proteome-scale feature discovery method for IDRs. Our method, which we call "reverse homology", exploits the principle that important functional features are conserved over evolution as a contrastive learning signal for deep learning: given a set of homologous IDRs, the neural network has to correctly choose a randomly held-out homologue from another set of IDRs sampled randomly from the proteome. We pair reverse homology with a simple architecture and interpretation techniques, and show that the network learns conserved features of IDRs that can be interpreted as motifs, repeats, and other features. We also show that our model can be used to produce specific predictions of what residues and regions are most important to the function, providing a computational strategy for designing mutagenesis experiments in uncharacterized IDRs. Our results suggest that feature discovery using neural networks is a promising avenue to gain systematic insight into poorly understood protein sequences.

Introduction

Despite their critical role in protein function, the systematic characterization of intrinsically disordered regions (IDRs) remains elusive (Van Der Lee *et al.*, 2014; Kulkarni and Uversky, 2018; Lindorff-Larsen and Kragelund, 2021). IDRs comprise of about 40% of the residues in eukaryotic proteomes (Davey, 2019). Unlike structured domains, IDRs do not fold into a stable secondary or tertiary structure, and this lack of structure helps facilitate many key functions. For example, some IDRs mediate protein-protein interactions or signaling, because their lack of structure allows them to adapt their conformation to different interaction partners (Wright and Dyson, 2015; Davey, 2019). A general property of IDRs is that they diverge rapidly at the primary amino acid sequence level (Pritišanac *et al.*, 2019). This rapid evolution means that IDRs challenge classic bioinformatics techniques that depend upon positional sequence alignments, including BLAST and Pfam (Van Der Lee *et al.*, 2014; Lindorff-Larsen and Kragelund, 2021).

Instead of relying on detection of sequence homology by alignment, a recently proposed strategy has been to identify functional features of IDRs that can be computed from sequence (Zarin *et al.*, 2019). Although the sequence may diverge, higher-order features that are critical to the proper function of the IDRs are typically conserved (Moses *et al.*, 2007; Beh, Colwell and Francis, 2012; Zarin *et al.*, 2017). These features are highly diverse. The best understood features are “short linear motifs”, peptides of 4-12 residues that fall into sequence families (Kumar *et al.*, 2020). In some cases, multiple copies or local clustering of motifs is necessary (Moses, Hériché and Durbin, 2007). Other IDRs depend upon global “bulk” features that are distributed through the entire sequence. For example, mitochondrial import IDRs require the sequence to be positively charged and hydrophobic (Bauer, Doetsch and Corbett, 2015), certain phase-separating proteins require IDRs with many R/G repeats that facilitate condensate-forming behavior (Chong, Vernon and Forman-Kay, 2018), and alternating positive and negative charged regions in the IDR of a cell-cycle regulating protein mediates the strength of phosphorylation of key regulatory sites (Das *et al.*, 2016). Combining features like these, curated from nearly three decades of IDR research, can characterize IDR function on a proteome-wide level, by describing IDR properties as a pattern across many distinct features (Zarin *et al.*, 2019). Indeed, conserved molecular functions were used as input for general predictions of IDR functions (Zarin *et al.*, 2021).

These literature-curated features are likely biased by researchers’ interests because they stem from small-scale experiments. Our knowledge of features important to IDRs is therefore not likely to be comprehensive. Indeed, features are continuously being discovered as research on IDRs develops: recently characterized features include aromatic amino acid patterning for prion-like domains (Martin *et al.*, 2020) or hydrophobic residues for activation domains (Erijman *et al.*, 2020). Features important for less characterized IDRs are less likely to be represented.

Here, we set out to design a systematic computational method for discovering features in IDRs, that is unbiased by prior knowledge or interest. The problem of feature discovery runs closely parallel to the concept of motif discovery (Das and Dai, 2007; Mohamed, Elloumi and Thompson, 2016): given a set of functionally related sequences, motif discovery methods attempt to find overrepresented subsequences with the idea that these motifs may represent conserved binding, interaction, or regulatory sites informative of the function of proteins. Motif discovery approaches range from fully unsupervised to regression approaches where function is predicted from sequence. Among the most successful strategies for motif discovery are those that exploit the principle that important functional motifs are conserved over evolution (Hardison, 2003; Budovskaya *et al.*, 2005; Xie *et al.*, 2005). Because comparative sequence data is available at genomic and proteomic scales, and is unbiased by a particular experimental condition or research question, comparative genomic and proteomic approaches have the potential to discover large numbers of functional motifs. However, alignment-based approaches to find conserved motifs in IDRs identify only a small minority (~5%) of the residues in IDRs (Nguyen Ba *et al.*, 2012); short motifs of about 2-10 residues often occur as small islands of conservation in IDRs that have no detectable sequence homology otherwise (Davey *et al.*, 2012). These short motifs are not expected to describe the “bulk” molecular properties such as charge or hydrophobicity, that are expected to be important for IDR function and appear to be conserved during evolution (Zarin *et al.*, 2019).

In order to develop a proteome-scale feature discovery approach capable of using evolution to learn more expressive features, we applied neural networks. To learn biologically relevant features, neural networks must be asked to solve a training task (i.e. a pre-specified loss function) (LeCun, Bengio and Hinton, 2015). Current approaches to infer sequence function using neural networks employ regression tasks, where models learn to predict expert annotations or large-scale measurements (Alipanahi *et al.*, 2015; Avsec *et al.*, 2021; Dhaval Vaishnav *et al.*, 2021). For example, training genomic sequence models on labels representing the presence or absence of transcription factor binding leads to the model learning features that directly correspond with the consensus motifs for these transcription factors (Koo and Eddy, 2019). Similarly, training neural networks to predict high-throughput measurements of activation domain function led to the discovery of clusters of hydrophobic residues within acidic regions as a key sequence feature (Erijman *et al.*, 2020; Sanborn *et al.*, 2021). While these supervised approaches discover important features, we reasoned that they would only learn features relevant for the specific training task.

Instead, we sought to use evolutionary conservation as a learning signal. Since orthologous sequences can be automatically obtained using sequence comparison and gene order (Altenhoff *et al.*, 2021; Howe *et al.*, 2021), labels about homology can be automatically obtained for IDRs. We therefore investigated self-supervised learning. Self-supervised learning trains models on “proxy” tasks resembling play and exploration (Jing and Tian, 2019), for which the labels can be automatically generated from data. These tasks are not directly useful, but are intended to teach the model transferable skills and representations, and are designed so the models learn autonomously without expert labels. Several self-supervised learning methods have been applied to protein sequences, and have been effective in teaching the models features that are useful for downstream analyses (Alley *et al.*, 2019; Heinzinger *et al.*, 2019; Rao *et al.*, 2019, 2021; Lu *et al.*, 2020; Rives *et al.*, 2021). However, the majority of these tasks directly repurpose methods from natural language processing (Alley *et al.*, 2019; Heinzinger *et al.*, 2019; Rao *et al.*, 2019; Rives *et al.*, 2021), and it is unclear what kinds of features the tasks induce the models to learn in the context of protein sequences.

We designed a new self-supervised method that purposes principles in comparative proteomics as a learning signal for our models. While IDRs generally cannot be aligned over long evolutionary distances (Riback *et al.*, 2017), they can still be considered homologous if they occur at similar positions in homologous proteins (Zarin *et al.*, 2019). Given a subset of homologous IDRs, our model is asked to pick out a held-out homologue from the same family, from a large set of non-homologous sequences. This task, which we call reverse homology, requires our model to learn conserved features of IDRs, in order to distinguish them from non-homologous background sequences. Our method is a contrastive learning method, a strategy that is now frequently employed in self-supervised learning (Oord, Li and Vinyals, 2018; Chen *et al.*, 2020; Liu *et al.*, 2020; Lu *et al.*, 2020; Lu, Lu and Moses, 2020). We show that reverse homology can be applied on a proteome-wide scale to learn a large set of diverse features. While these “reverse homology features” are learned by the neural network to solve the reverse homology proxy task, we show that they can be visualized and interpreted, are enriched for biological function, and can be purposed for bioinformatics analyses that yield hypotheses connecting specific features to function.

We combine reverse homology with a biologically principled architecture and interpretation techniques. We show how these interpretations can be paired with functional enrichments on a protein set level, or used to understand features that contribute towards function for any specific given IDR. Taken together, our results demonstrate that unbiased feature discovery is an unexplored application of self-supervised learning for protein sequences.

Results

Reverse Homology

To learn functional features of IDRs unbiased by prior knowledge, we propose a novel self-supervised proxy task that uses evolutionary homology between protein sequences to pose a contrastive learning problem. Homologous proteins derive from a shared evolutionary ancestor, and will frequently share similar functions (Pearson, 2013). For full proteins and structured domains, homology can be reliably identified based on sequence similarity (Pearson, 2013). Because they evolve rapidly and can have no statistically detectable similarity, it is difficult to assign homology using sequence similarity alone. However, IDRs are usually flanked by structured regions that are more conserved (Zarin *et al.*, 2017). Since these structured regions will usually align well, and since the order of domains is usually strongly conserved in proteins (Kummerfeld and Teichmann, 2009), IDRs that occur at the same position across homologous proteins in a multiple-sequence alignment can be considered to be homologous even when they share little sequence similarity (Figure 1A) (Chen *et al.*, 2006a, 2006b; Bellay *et al.*, 2011; Colak *et al.*, 2013). As bioinformatics tools can accurately annotate what parts of a protein are IDRs (Jones and Cozzetto, 2015; Hanson *et al.*, 2017), defining homologous groups of IDRs using multiple-sequence alignments across the entire proteome can be defined as a fully automated operation (Zarin *et al.*, 2019).

We will use these sets of homologous IDRs as the basis for our proxy task (see Methods for a more detailed definition.) Given a family of homologous IDRs, a neural network (Figure 1D) is asked to determine which sequence is a held-out homologue homologues from a set of IDRs where the other sequences are randomly drawn non-homologous sequences (Figure 1E). We call this task "reverse homology", because it "reverses" the typical sequence homology search process, where we have a target sequence of unknown family, and we search across many query families or sequences to assign homology (Pearson, 2013). In our task, we give the model a known query family, and ask it to determine if target sequences are homologous or not. We show a schematic description of reverse homology for IDRs as Figure 1.

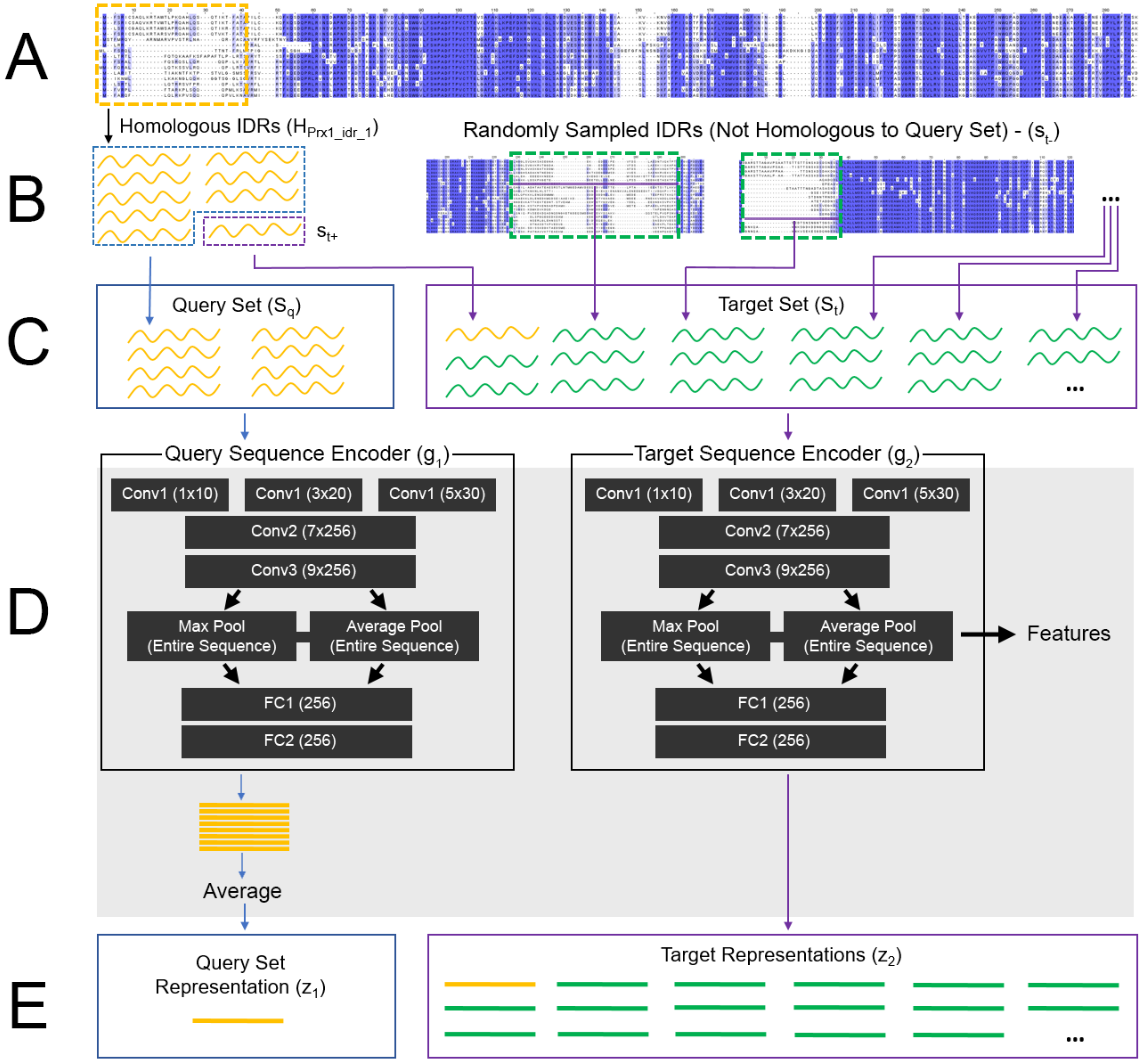

A
Homologous IDRs ($H_{Prx1\_idr\_1}$)
Randomly Sampled IDRs (Not Homologous to Query Set) - (s_{t-})
B
s_{t+}
...
C
Query Set (S_q)
Target Set (S_t)
...
D
Query Sequence Encoder (g_1)
Conv1 (1x10)
Conv1 (3x20)
Conv1 (5x30)
Conv2 (7x256)
Conv3 (9x256)
Max Pool (Entire Sequence)
Average Pool (Entire Sequence)
FC1 (256)
FC2 (256)
Target Sequence Encoder (g_2)
Conv1 (1x10)
Conv1 (3x20)
Conv1 (5x30)
Conv2 (7x256)
Conv3 (9x256)
Max Pool (Entire Sequence)
Average Pool (Entire Sequence)
FC1 (256)
FC2 (256)
Features
Average
E
Query Set Representation (z_1)
Target Representations (z_2)
...

Figure 1. A schematic description of the reverse homology method. Details that are more specific to our implementation, as opposed to being part of the general method, are highlighted in grey. A) At the top, we show the multiple sequence alignment for the yeast protein Prx1 (as an example - during training, we iterate over all IDRs in all proteins) across 15 yeast species. Conserved residues are highlighted in blue. The yellow dotted line box shows the boundaries of an IDR in Prx1. B) By taking IDRs from different species in the yellow dotted box in A, we construct a set of homologous IDRs, H (shown as yellow lines) C) We sample a subset of IDRs (blue dotted box) from H and use this to construct the query set (blue box). We also sample a single IDR (purple dotted box) from H not used in the query set and add this to the target set (purple box). Finally, we populate the target set with non-homologous IDRs (green), sampled at random from other IDRs from other proteins in the proteome. D) This panel includes detail s that are more specific to our implementation (highlighted in grey). The query set is encoded by the query set encoder g_1. The target set is encoded by the target set encoder g_2. In our implementation, we use a five-layer convolutional neural network architecture. We label convolutional layers with the number of kernels x the number of filters in each layer. Fully connected layers are labeled with the number of filters. E) The output of g_1 is a single representation for the entire query set. In our implementation, we pool the sequences in the query set using a simple average of their representations. The output of g_2 is a

representation for each sequence in the target set. The training goal of reverse homology is to learn encoders g_1 and g_2 that produce a large score between the query set representation and the homologous target representation, but not non-homologous targets. In our implementation, this is the dot product: $g_1(S_q) \cdot\ g_2(s_{t+}) > g_1(S_q) \cdot\ g_2(s_{t-})$. After training, we extract features using the target sequence encoder. For this work, we extract the pooled features of the final convolutional layer, as shown by the arrow in D.

In previous work, we explained the theoretical principles behind using evolutionary homology as a basis for contrastive learning (Lu, Lu and Moses, 2020): our method is expected to learn conserved features of protein sequences, which we argue are likely important for the conserved function of rapidly-diverging IDRs (Supplementary Methods).

We implemented our method with a lightweight convolutional neural network architecture (grey box in Figure 1). We use both max and average pooling to reflect different ways local features can contribute to function in IDRs. Some functions may only require a feature to be present or absent; for example, a single SH3 binding motif (Stollar *et al.*, 2009) may be sufficient for recognition and function. We reasoned max pooling, which identifies a single window that maximally activates (i.e. creates the highest feature value for) the feature, would capture these kinds of features. Other functions require multiple copies of a feature (Moses, Hériché and Durbin, 2007), a certain proportion of the sequence to have a feature (Chong, Vernon and Forman-Kay, 2018), or scale as more of the feature is present (Zarin *et al.*, 2017). We reasoned average pooling, which produces the average activation value across all windows, would capture these kinds of features. This architecture facilitates interpretation (see Methods): we pair our trained models with several neural network interpretation methods to understand the features learned. In principle, this architecture can be replaced with many others, but we leave exploration of possible architectures to future work (see Discussion.)

Reverse homology learns a diverse range of features for yeast intrinsically disordered regions

We trained a reverse homology model using 5,306 yeast IDR families containing a total of 94,106 sequences. To qualitatively understand the features that this model learns, we produced a UMAP scatterplot (McInnes, Healy and Melville, 2018), where each point represents a feature, using the correlation distance between activation values across IDR sequences for each feature. We paired this with an interpretation method that generates sequence logos for each feature, (adapted from (Koo and Eddy, 2019); see Methods). Figure 2 shows the features from the final convolutional layer of our target sequence encoder, chosen because the sequence logo interpretation method we use is designed for convolutional layers, and because the target sequence encoder is trained to encode single sequences.

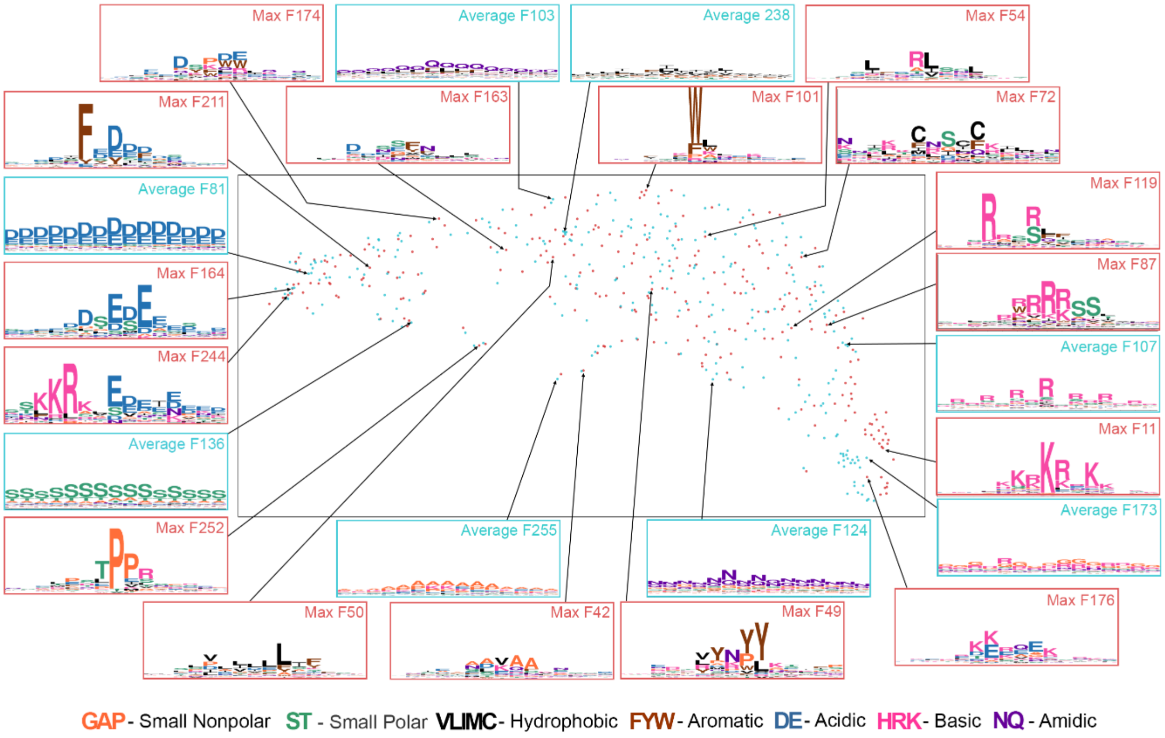


Figure 2. UMAP scatterplot of reverse homology features for our yeast model. Reverse homology features are extracted using the final convolutional layer of the target encoder: max-pooled features are shown in red, while average-pooled features are shown in blue. We show the sequence logo corresponding to select features, named using the index at which they occur in our architecture (see Methods for how these are generated.) Amino acids are colored according to their property, as shown by the legend at the bottom. All sequence logos range from 0 to 4.0 bits on the y-axis.

Overall, we observed four major axes of features. To the left of the scatterplot, we observed negatively charged features (e.g. Average F81 and Max F164). Positively charged features were concentrated in the bottom right (e.g. Max F11 and Average F173). Features containing hydrophobic amino acids are at the top of the distribution (e.g. Average F238, Max F54, and Max F72). Finally, we observed features rich in uncharged polar amino acids (e.g. Average F136, Max F252 and Average F124) or alanine (Average F255 and Max F42) scattered along the bottom of our UMAP.

Features in between these poles often exhibited a mixture of properties. For example, Max F211, Max F174, and Max F163 all contain both negative and hydrophobic residues, Max F54 contains both positive and hydrophobic residues, and Max F49 contains both aromatic and polar amino acids.

Specific features captured both motifs and bulk properties known to be important for IDR function. As examples of motifs, Max F252 is consistent with the TPP phosphorylation motif (Schwartz and Gygi, 2005), while Max F87 is similar to the PKA phosphorylation motif RRxS (Smith, Samelson and Scott, 2011). As examples of bulk properties, Average F173 captures RG repeats important for phase separation (Chong, Vernon and Forman-Kay, 2018), while other average features look for combinations of amino acids with similar biochemical properties

(Average F136 measures S/T content, Average F124 measures N/Q content, and Average F81 measures acidic amino acids D/E). Finally, other features captured patterns that we were not able to associate with previously known IDR properties: for example, Average F107 captures spaced out arginine repeats (e.g. RxRx), and Max F244 captures a window of positive to negative charge transition. We hypothesize these features could be capturing charge patterning in IDRs (Das and Pappu, 2013; Sawle and Ghosh, 2015).

Overall, we were able to identify 70 of 512 features as complete or partial matches to motifs or bulk features previously considered to be important to IDRs (Supplementary File 1). We consider this number a lower bound as there are features that we were uncertain about. Together, our global analysis demonstrates that reverse homology induces our model to learn a wide diversity of biochemically sensible features.

Reverse homology features are predictive of yeast IDR function and correlated with previous literature-curated features

Having qualitatively confirmed that our model learns diverse features, our next goal was to more systematically evaluate if these features are biologically meaningful. To do this, we first benchmarked vector representations of IDRs extracted using our model on a series of classification problems predicting various aspects of IDR function (Supplementary Tables 2.1, 2.2, and 2.3 in Supplementary File 2; details on classification datasets, classifiers, and baselines are also in Supplementary File 2). We reasoned that performance on these classification problems would indicate if our model was learning features relevant to these IDR functions.

Overall, we observe across most problems, self-supervised protein representation learning methods (Alley *et al.*, 2019; Heinzinger *et al.*, 2019; Rao *et al.*, 2019) outperform expert-designed knowledge-based features curated from literature (Zarin *et al.*, 2019) (Supplementary Tables 2.1-2.3). For our reverse homology method, we observed a trade-off depending on layer between the performance of the representation on these prediction tasks and interpretability. Representations from the final fully connected layer of our target encoder perform comparably to other self-supervised protein representation learning methods, suggesting that our model can represent IDRs at a similar level of performance but with a more constrained low-parameter architecture and substantially less training data. However, the features in this layer are less interpretable with our interpretation methods than the convolutional layers; we find that representations from the final convolutional layer perform worse on our classification tasks. Despite this, representations from our final convolutional layer still outperform the literature-curated features at most problems, suggesting that these features may still encode more functional information than expert-curated features, and features from an untrained randomly-initialized model, confirming that this performance is dependent on training with our self-supervised proxy task. For the remainder of this paper, we will focus on the features from the final convolutional layer of the target encoder, as our goal in this paper is to interpret the features learned to form hypotheses about function.

To test if our reverse homology features were sensitive to different kinds of biological functions than literature-curated features, plotted (Figure 3A) the enrichments for GO terms

(Supplementary Table 2.3 in Supplementary File 2). In this benchmark, we counted proteins that had a protein with the same GO term as their nearest neighbor in the reverse homology feature space, and compared the fraction to the background. Importantly, this analysis includes all GO Slim terms with at least 50 proteins in our set of proteins with IDRs (for a total of 92 terms), so it is not biased towards functions previously known to be associated with IDRs.

Overall, we find that while some GO Slim categories are highly enriched with both feature sets (e.g. “Cell wall organization”, “Translation” or “Inner Mitochondrial Membrane”, highlighted in Figure 3A), other categories are much more enriched with our reverse homology features than literature-curated features (e.g. “Oxidoreductase Activity”, “Intracellular Protein Transport”, or “Golgi Membrane”). Conversely, some categories are more enriched using literature-curated features (e.g. “Meiosis”, “Intracellular Signal Transduction”, or “DNA replication”). These results suggest that the neural network is learning features relevant to biological processes that are different from the ones associated with literature-curated features.

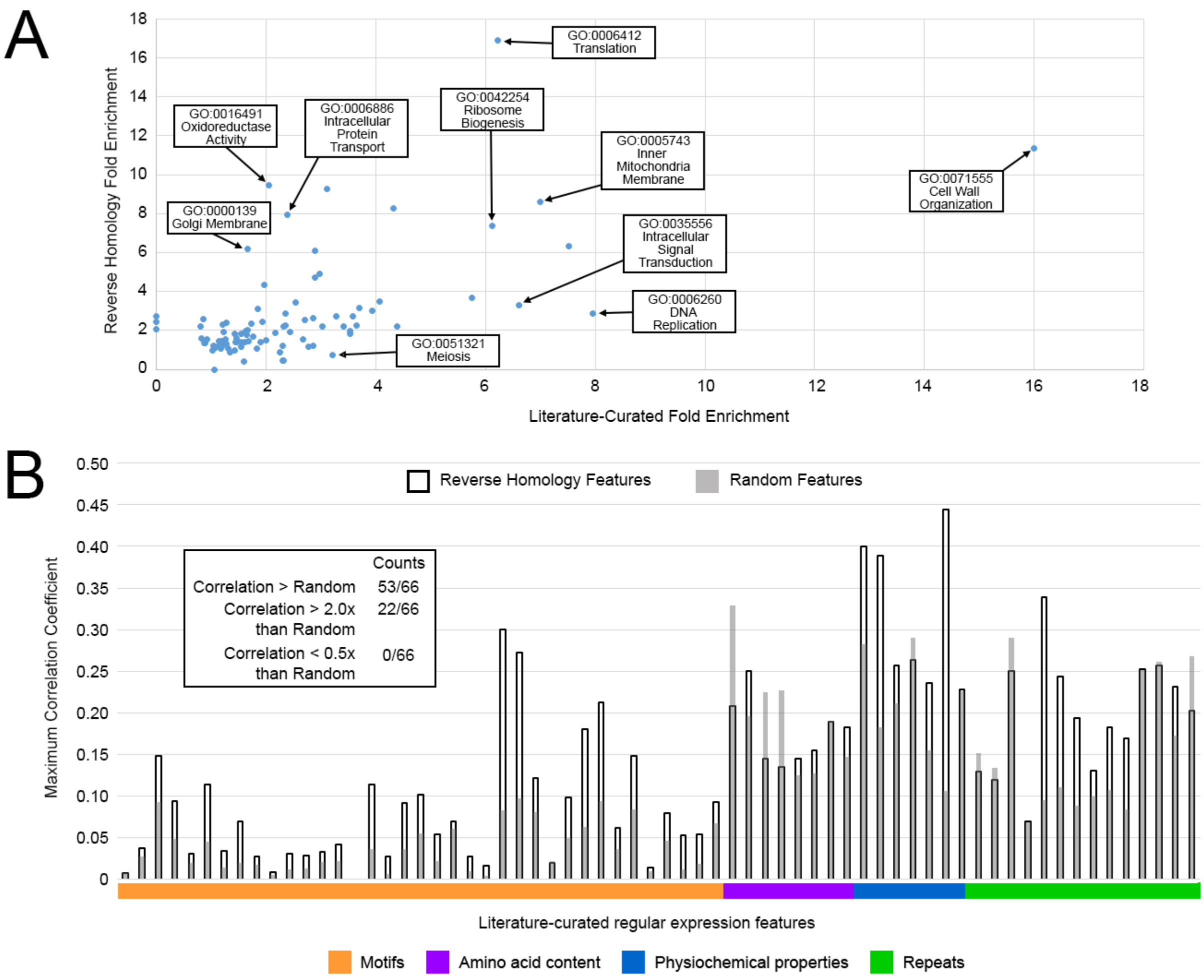


Figure 3. A) Scatterplot of the fold enrichment for the set of nearest neighbors using feature representations from the final convolutional layer of the target encoder of our reverse homology model, versus literature-curated feature representations, for 92 GO Slim terms. We show the names of some GO

terms in text boxes. B) Bar plots of the maximum correlation between all neural network features in the final convolutional layer of our models, versus each of the 66 literature-curated features. Literature-curated features are grouped by their category, as shown in the bottom legend. Correlations from our trained reverse homology model are shown as black outlined boxes, while correlations from a randomly initialized untrained model are shown in grey. In the text box, we show the number of features where the reverse homology features are more correlated, more than 2.0x correlated, and less than 0.5x times correlated than the untrained random features.

Finally, we compared reverse homology features to literature-curated features by Zarin *et al.* (Zarin *et al.*, 2019). For each of these features (66 out of 82 features) that can be expressed as regular expressions we compared the correlation of our trained reverse homology model to a randomly initialized model, as random untrained models have shown to be a strong baseline for protein representation learning problems (Shanehsazzadeh, Belanger and Dohan, 2020; Lu, Lu and Moses, 2021). (Figure 3B). We reasoned that a good measure of consistency between features detected by our neural network and literature-curated features would be if the neural network feature is activated at the same positions in amino acid sequences as the literature-curated feature.

Figure 3B shows the correlation of the maximally correlated neural network feature from the final convolutional layer of our target encoder, for each of the 66 literature-correlated features, using our trained model versus a randomly initialized untrained model. (Supplementary File 3 contains a table of all features, their regular expressions, and the maximal correlations with our trained and random models.) Due to the large number of parameters in neural network models, random untrained models have shown to be a strong baseline for protein representation learning problems (Shanehsazzadeh, Belanger and Dohan, 2020; Lu, Lu and Moses, 2021). In our case, many of the literature-curated features are relatively simple and reflect only single amino acid repeats, or relationships between subsets of 2-3 amino acids, and may be easy to be captured by chance given the large number of parameters. We reasoned that features from a random untrained model would therefore be a strong baseline that would demonstrate where correlations with prior biological features must be learned.

In general, we observe higher correlations with 55 of 63 literature-curated features with our reverse homology model than an untrained random model. Features corresponding to motifs are learned significantly better by our reverse homology model than the random model (n=37, paired t-test p-value 4.63E-06; mean 0.042 and standard deviation 0.0475). Features corresponding to physicochemical properties (n=7; p-value 0.067; mean 0.1091 and standard deviation 0.1266) and repeats (n=14; p-value 0.077; mean 0.042 and standard deviation 0.082) are more correlated, but with less confidence than motifs. In contrast, features learned by our model are not more correlated with features capturing single amino acid content compared to a randomly initialized model, and are in fact, slightly less correlated on average (n=8; p-value 0.4409; mean -0.02 and standard deviation 0.067). These results are consistent with the intuition that shorter repeats of one or two amino acids are more likely to be present in features by random chance, but longer and more specific combinations like motifs must be learned.

Features that recognize bulk properties associated with cell wall maintenance and phase separation

Having confirmed the global biological relevance of our features, we next sought to test whether we could associate individual reverse homology features with previously known functional features and use them to understand functions of uncharacterized IDRs. First, we considered features recognizing S/T repeats (Average F136 – Figure 4A) and RG repeats (Average F65 – Figure 4C).

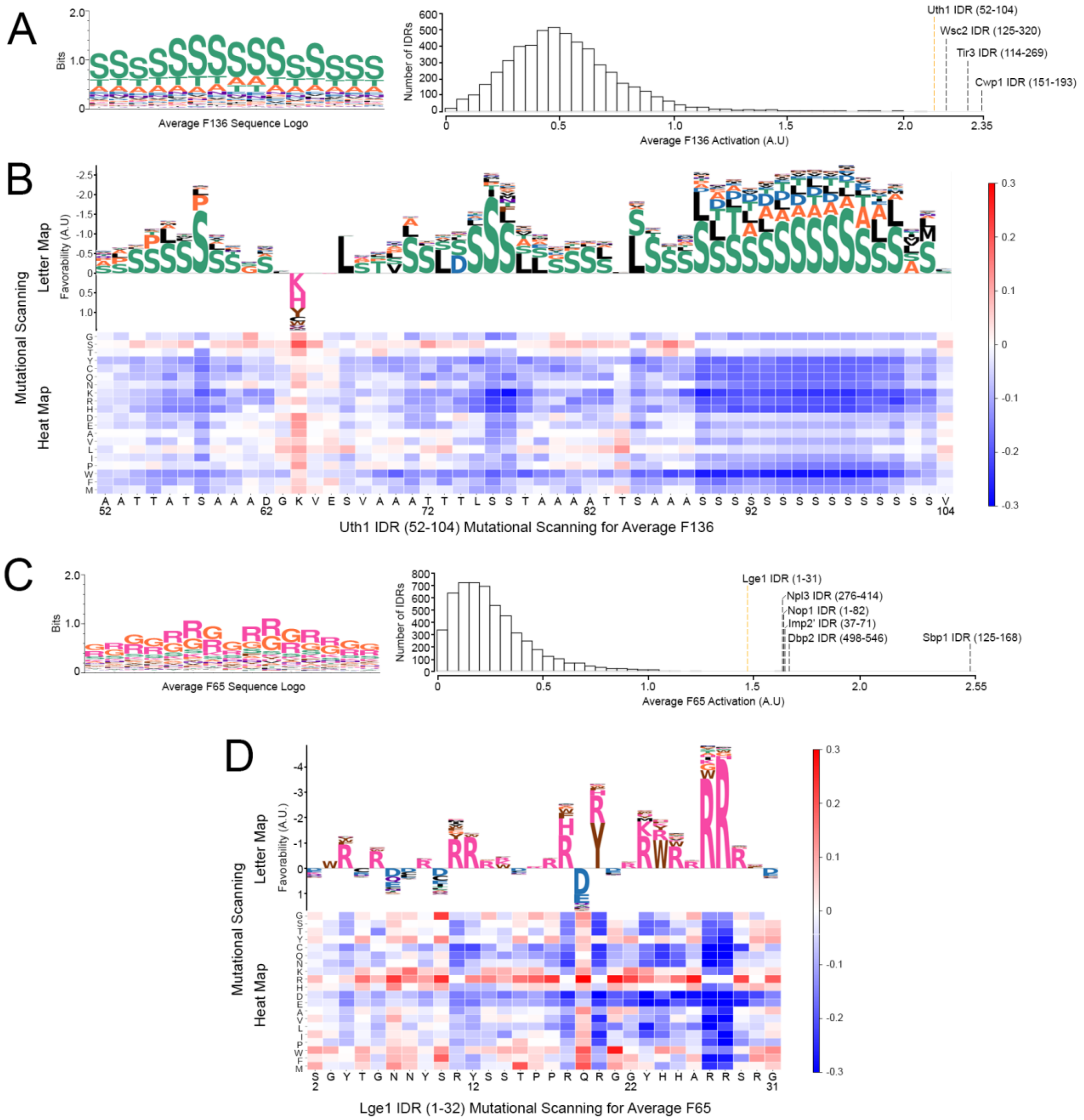


Figure 4. Sequence logos, feature distributions, and one example *in-silico* mutational scanning map for each average feature. (A,C) Sequence logos and a histogram of the value of the feature across all IDRs is shown for Average F136 (A) and Average F65 (C). We annotate the histograms with the top activating

sequences. (B,D) We show mutational scanning maps for F136 for an IDR in Uth1 in B and for F65 for an IDR in Lge1 (D), which are the 4th and 6th most activating sequences for their respective features. Mutational scanning maps are visualized as heat maps and letter logos. For the heat maps, each cell corresponds to the change in value for the feature if that position in the sequence (wild-type sequence is shown on the x-axis, the numbers correspond to the amino acid coordinates in the protein) is mutated to the amino acid shown on the y-axis. A shared color map for all heat maps is shown in the top right. For the mutational scanning letter maps, positions above the axis are positions where retaining the original amino acid is generally preferable, while positions below the axis are positions where the feature could generally be improved by mutating to another amino acid. The height of the combined letters corresponds to the total magnitude of the change in the feature for all possible mutations (which we define as the favourability). For positions above the axis, we show amino acids that result in the highest value for the feature (i.e. the most favored amino acids at that position.) For positions below the axis, we show amino acids that result in the lowest value for the feature (i.e. the most disfavored amino acids at that position.)

Long regions of S/T-rich segments are often sites of *O*-glycosylation in yeast proteins (González, Brito and González, 2012). A previous computational analysis revealed that fungal proteins with an extremely high proportion of S/T-rich regions in their sequence are often cell wall proteins involved in maintenance of the cell wall (González, Brito and González, 2012). Consistent with this, we find an enrichment (using the GOrilla tool (Eden *et al.*, 2009)) for cell wall proteins (15/31, q-value 3.16E-16), cell wall organization or biogenesis proteins (20/31, q-value 1.69E-15), and extracellular region proteins (17/31, q-value 4.15E-20) in the proteins with IDRs that highly activate Average F136. For our S/T repeat feature Average F136, we observed that the top 3 IDRs are all cell wall proteins: Cwp1 (Van der Vaart *et al.*, 1995) and Tir3 (Abramova *et al.*, 2001) are cell wall mannoproteins, while Wsc2 is involved in maintenance of the cell wall under heat shock (Verna *et al.*, 1997). Our 4th ranked IDR is in Uth1, which is predominantly known as a mitochondrial inner membrane protein (Welter *et al.*, 2013). However, deletion of Uth1 alters the polysaccharide composition of the cell wall, with mutants being more robust to lysis conditions, leading to the argument that Uth1's role at the cell wall, not the mitochondria, better explains its functions in cell death (Ritch *et al.*, 2010).

To analyze the IDR in Uth1 in closer detail, we produced *in-silico* mutational scanning maps (Figure 4B). We systematically mutated each amino acid position in the IDR to every other amino acid, and measured the change the mutation induces in the value of Average F136. We visualize these mutational scanning maps two ways. First, we visualize them as heat maps, as shown in the bottom of Figure 4B: mutations that would result in a drop in the value of the feature are shown in blue, while mutations that increase the feature are shown in red. Second, as shown in the top of Figure 4B, we visualize them as sequence logos, which we term "letter maps" to distinguish them from the sequence logos shown in Figure 3 (see Methods for details). In these letter maps, residues that the feature favors (i.e. would generally result in a drop in the feature if was mutated) are shown above the axis, while residues the feature disfavors are shown below the axis. For favored positions, we show the amino acids that are most favored; note that this may not always be the wild-type amino acid, as the feature can still generally favor the wild-type, but favor other amino acids more or equally. For disfavored positions, we show the amino acids that are more disfavored. More details on how these in-silico mutational scanning maps are produced can be found in the Methods. Overall, analyzing the IDR at 52-104 in Uth1 reveals long tracts of S, T, and A-rich regions that are favored by our features (Figure 4B).

Similarly, RG repeats are found in RNA-binding proteins that form membraneless organelles (Chong, Vernon and Forman-Kay, 2018). Consistent with this, we found an enrichment for RNA-binding (13/20, FDR q-value 6.37E-5) proteins and for proteins localizing to the ribonucleoprotein complex (9/20, q-value 1.47E-1) in the IDRs that most strongly activate Average F65.

Indeed, for our RG-repeat feature Average F65, 4 out of 6 of the top IDRs are proteins with known stretches of RG-repeats (Dbp2 - (Kharel *et al.*, 2020)), and 3 are phase-separating proteins mediated by interactions between RG-rich regions (Sbp1 - (Poornima *et al.*, 2019), Npl3 and Nop1 - (Chong, Vernon and Forman-Kay, 2018)). Interestingly, while Lge1 is also known to phase-separate through its N-terminal IDR, also identified in our analysis, this IDR is not canonically considered an RG-rich IDR and instead has been described as an R/Y-rich region (Gallego *et al.*, 2020). Closer analysis of Average F65 applied on the Lge1 N-terminal IDR (Figure 4D) indicates that the feature also prefers Ys and other aromatic acids in addition to R; although replacing G with most other amino acids reduces the value of the feature, replacing G with R, Y, or W improves the value of the feature in most spots. The preference for aromatic amino acids in addition to RG-repeats is consistent with emerging observations that aromatic amino acids can mediate similar pi-pi interactions as RG-repeats (Chong, Vernon and Forman-Kay, 2018; Kharel *et al.*, 2020). We hypothesize that our feature may reflect this relationship and may be subsuming two types of features previously thought of as distinct (RG-repeats and R/Y-rich regions) into a single logic.

Discovery of subclasses of PKA phosphorylation consensus sites

As examples of features that recognize motifs, we identified at least two features that recognize variations of the PKA phosphorylation consensus motif RRxS (Smith, Samelson and Scott, 2011). Max F231 (shown on the x-axis in Figure 5) recognizes the canonical consensus motif, while Max F87 (y-axis in Figure 5) recognizes a more stringent variation RRRSS.

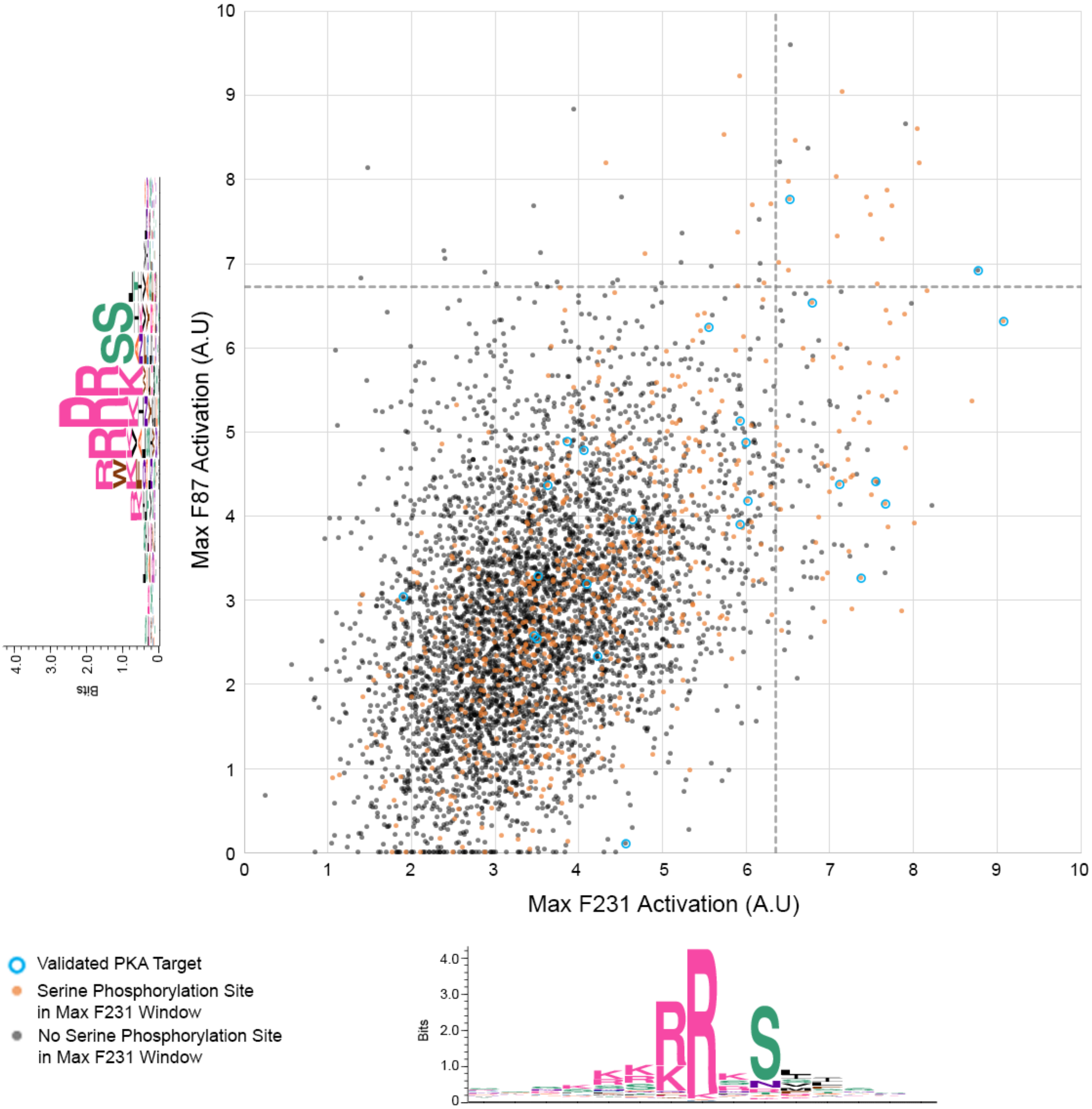


Figure 5. Scatterplot of the activation for Max F231 versus Max F87 across all yeast IDRs. We show the sequence logo for each respective feature on their axis. Cut-offs for highly activating IDRs, defined as at least 70% of the maximum activating IDR, are shown as the grey dotted lines. IDRs with validated PKA (Tpk1, Tpk2, or Tpk3) phosphorylation sites are circled in blue. IDRs where an 8 amino acid window around the maximum activating position for Max F231 overlaps a serine phosphorylation site in Biogrid (Oughtred *et al.*, 2021) are in orange. All other IDRs are shown in grey.

We found that both features were predictive of PKA targets. In total, the Biogrid database (Oughtred *et al.*, 2021) had validated PKA phosphorylation targets in 24 yeast IDRs, shown as the blue outlines in the scatterplot in Figure 5. Of these, 8 of 24 were in the top activating IDRs for Max F231 (defined as more than 70% of the value of the maximally activating IDR), reflecting an enrichment of 18.7 times compared to lesser-activating IDRs (8/139; Fisher exact test p-value 8.8E-08). 2 were contained in the top activating IDRs for Max F87, reflecting a lesser but still significant enrichment of 9.1 times (2/52; p-value 0.022).

We next verified that for the validated PKA targets identified by feature, the feature was correctly identifying the actual PKA phosphorylation site. We found that for both features, a yeast PKA homologue (Toda *et al.*, 1987) phosphorylation site was present in an 8 amino acid window around the maximally activating position in the sequence, with the sole exception of the IDR spanning positions 86-506 in Rgt1. While this IDR meets the threshold for both features, both features identify an "RRKS" subsequence at position R477 as the maximally activating position; however, the actual PKA sites in this IDR are at S283 and S284, and S480 is not recorded as a known serine phosphorylation site. We hypothesize that S480 could represent an additional PKA site for this protein.

Since the exact phosphorylation site of many PKA targets (Pautasso *et al.*, 2016) is not known or at least recorded in the Biogrid database, we also tested if sites identified by our features were enriched in serine phosphorylation sites in general. Overall, Biogrid contains 14994 unique serine phosphorylation sites that occur in yeast IDRs. For the top activating IDRs for each feature, we counted the number of IDRs that had a serine phosphorylation site in an 8 amino acid window around the maximally activated position in the sequence. We compared this to an expectation of 1000 samples of an equivalent number of IDRs chosen at uniform probability, with 8 amino acid windows also sampled randomly from these IDRs. For Max F231, 68/139 IDRs contained at least one serine phosphorylation site, a 4.5 times enrichment over random expectation (z-score 63.8, expectation mean 15.3 and standard deviation 3.7). For Max F87, 25/52 IDRs contained at least one serine phosphorylation site, a 4.4 times enrichment over random expectation (z-score 22.4, expectation mean 5.6 and standard deviation 2.2). Taken together, these results confirm that the neural network has identified two features that represent the phosphorylation consensus site for PKA, and that many of the predicted sites may represent uncharacterized sites that are phosphorylated *in vivo*.

Qualitatively, we observed that the sequence logo of Max F87 resembles two overlapping PKA motifs. A recent study indicates that PKA is capable of phosphorylating two serines in a row (Dengler *et al.*, 2021). The authors observe in these cases, one serine is constitutively phosphorylated, and the other serine requires increased PKA activity, and hypothesize this multi-site phosphorylation might be a regulatory mechanism. Based on this hypothesis, we wondered if the distinction between Max F87 and Max F231 is that the former is more specific to sites of double phosphorylation. We analyzed the frequency of two adjacent phosphorylated serines in the 8 amino acid window around the maximally activated position of the top activating sequences for both features: 13 of 52 IDRs are doubly phosphorylated for windows identified by Max F87, while 15 of 139 IDRs are for windows identified by Max F231, suggesting that Max F87 identifies 2.31 times more doubly phosphorylated regions (Fisher exact test p-value 0.020). Our results suggest that the doubly phosphorylated PKA consensus is a more widespread mechanism than currently appreciated, and illustrate the power of our unsupervised approach to discover unexpected and subtle biological patterns.

A positive-to-negative charge transition window is associated with nucleolar function

Finally, we identified a feature that recognizes a site of positive-to-negative charge transition, Max F244. The sequence logo for this feature (Figure 6A) indicates that sequence segments activating this feature typically have basic amino acids at the start of the sequence, and acidic amino acids at the end of the sequence. We note that this feature is similar to previous expert-defined features that measure the separation between positive and negative charged residues (Das and Pappu, 2013; Sawle and Ghosh, 2015), which is generally known to be important to IDRs; our feature is likely related, but identifies a single local window instead of measuring these charge transitions as a property across the entire sequence.

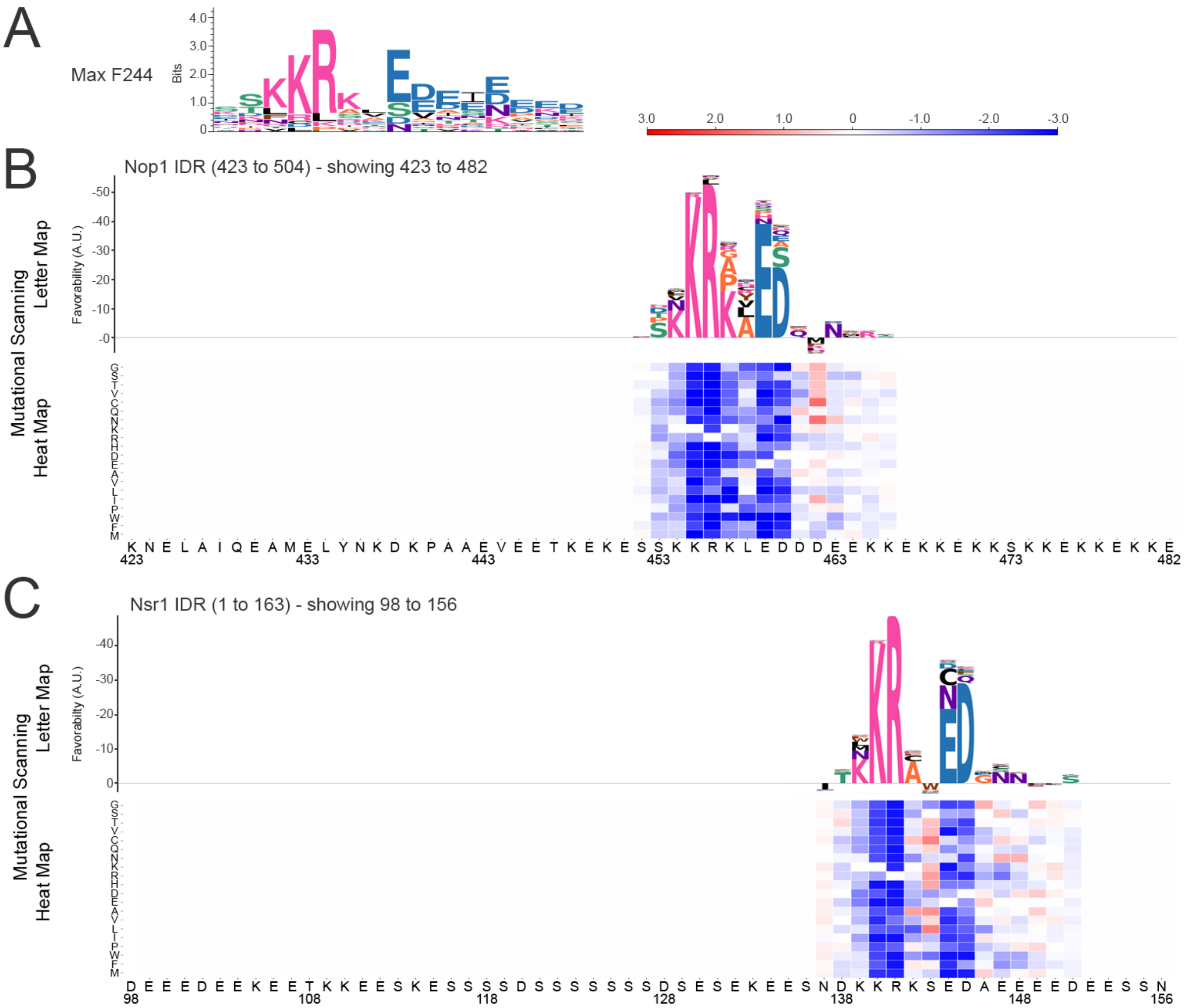


Figure 6. Sequence logos and *in-silico* mutational scanning maps for charge transition feature. (A) The sequence logo for Max F244. (B,C) Mutational scanning maps for Max F244 for an IDR in Nop56 (B) and an IDR in Nsr1 (C). Note that we only show a segment of the mutational scanning maps in this figure due to the long length of these IDRs. Mutational scanning maps are visualized as heat maps and letter maps. For the heat maps, each cell corresponds to the change in value for the feature if that position in the sequence (wild-type sequence is shown on the x-axis, the numbers correspond to the amino acid coordinates in the protein is mutated to the amino acid shown on the y-axis. A shared color map for all

heat maps is shown in the top right. For the mutational scanning letter maps, positions above the axis are positions where retaining the original amino acid is generally preferable, while positions below the axis are positions where the feature could generally be improved by mutating to another amino acid. The height of the combined letters corresponds to the total magnitude of the change in the feature for all possible mutations (which we define as the favourability).. For positions above the axis, we show amino acids that result in the highest value for the feature (i.e. the most favored amino acids at that position.) For positions below the axis, we show amino acids that result in the lowest value for the feature (i.e. the most disfavored amino acids at that position.)

We found an enrichment for nucleus (21/23, q-value 1.22E-2) and nucleolus-localized (8/23, q-value 2.37E-2) proteins in the proteins with IDRs that strongly activate Max F244. Many of these proteins are well-characterized nucleolar proteins (including Nsr1, Nop7, Nop56, Rix1, Rix7, Dbp7 and Erb1).

Due to the nature of the max pooled features, which recognize a single maximally activating window, the *in-silico* mutational scanning maps focus on one specific segment of the sequence, as opposed to average features associated with cell wall maintenance and phase separation discussed above, which generally distribute across the entire sequence. We observed that for several of the nucleolar proteins, the region of the sequence activating Max F244 was often preceded or succeeded by a highly charged region on one end, and a more neutral region on another end. We show two examples in Figure 6. In Nop56 (Figure 6B), the C-terminal succeeding the activating region is characterized by lysine repeats (with some acidic amino acids interspersed.) In Nsr1 (Figure 6C), the N-terminal preceding the activating region is characterized by repetitive negative charges and serines.

While we were not able to associate these activating regions with any previous literature, we did find that they are often adjacent to charged regions that are critical to protein-protein interactions. In Nop56, the C-terminal lysine-rich region from E464 to D504 has previously been described as the "K-tail" (Oruganti *et al.*, 2007) and is important interaction with fibrillarin (Gagnon *et al.*, 2012); our *in-silico* mutational scanning map indicates that K464 to D460 is the most important region for Max F244 in the C-terminal IDR of Nop56. Similarly, in Nsr1, a deletion study of the acidic N-terminal region spanning M1 to S125 indicates that the region is important for interactions with Top1 (Edwards *et al.*, 2000); our mutational scanning map indicates that K140 to D145 is the most important region for Max F244 in the N-terminal IDR of Nsr1. Overall, these findings suggest that charged interaction domains in nucleolar proteins are often connected to the rest of the protein by a short window of positive-to-negative charge transition. While we leave investigation of the possible function of this feature to future work, this example illustrates how features learned by our unsupervised feature discovery approach can implicate specific regions of sequences to provide hypotheses for follow-up experimental work.

Reverse homology trained on human IDRs also yields diverse features correlated with literature-curated features

Having confirmed that our reverse homology task was effective at learning features that could be interpreted to drive hypotheses about yeast biology, our next goal was to train a model for human

IDRs. We trained a reverse homology model using 16,328 human IDR homology families for a total of 1,604,052 sequences (see Methods).

We qualitatively confirmed that this model was learning a similar diversity of features as our yeast model (Supplementary File 4). We were able to identify features for both short linear motifs, such as consensus motifs for phosphorylation sites or metal ion binding motifs, and bulk features like repeats or charge. As with our yeast model, we found that our features were significantly more correlated with literature-curated features than a random model (paired t-test p-value 1.589E-10, average 0.084, standard deviation 0.090). Supplementary File 5 contains the maximum correlations with the 66 features we previously tested for the yeast reverse homology model (for human IDRs, no comprehensive set of features has yet been published.)

Analysis of diverse human IDRs reveals consistency between reverse homology predictions and known features in literature

We decided to test if our human reverse homology model was capable of predicting specific residues and regions of functional significance for specific IDRs of interest. While our model is not directly trained to predict function, our reverse homology task requires the model to identify conserved features between homologues. We reasoned that features conserved by evolution are also likely to be important to function, so analyzing what features our model was detecting in a given IDR to determine homology would likely yield insights into function. Current computational methods to predict IDR function are either not designed to pinpoint specific residues, or yield lists of matches to pre-defined motifs. For example, the ANCHOR2 method, which predicts binding regions in IDRs (Mészáros, Erdős and Dosztányi, 2018), predicts almost the entire IDR of the human cell cycle regulator protein p27 (Supplementary Figure 1A). The ELM prediction tool scans for matches to a database of pre-defined short linear motifs (Kumar *et al.*, 2020), and predicts 63 matches for the IDR of p27 (Supplementary Figure 1B).

To test if our model was capable of retrieving functionally relevant features for specific human IDRs, we examined three cases. First, we focused on p27 (also known as CDKN1B), because it has a well-studied C-terminal IDR spanning positions 83 to 198 in the sequence, known as the kinase inhibitory domain (p27-KID). This region mediates promiscuous interactions with cyclin-dependent kinase (Cdk)/cyclin complexes through a disorder-to-order transition (Yoon *et al.*, 2012). We reasoned that given the abundance of literature allowing us to assess the relevance of our predictions, p27-KID would be a good case to test if our model, which has never been trained on any of this prior knowledge in literature, was retrieving functionally relevant features.

We show a summary of known post-transcriptional modification sites and localization signals (Vervoorts and Lüscher, 2008; Abbastabar *et al.*, 2018) in purple in Figure 7A. The top five features predicted by our reverse homology model, which are all max pooled features for this IDR, all overlap these sites (red in Figure 7A). (Note that there can be a mixture of both max and average pooled features in the top ranked features, but in our examples, the top ranked features all happen to be either all max or all average.)

We observed through the mutational scanning letter maps that the important residues for our reverse homology features were consistent with these known sites. Max F49 (Figure 7A.1) overlaps Y88 and Y89, which are modification sites for SRC family tyrosine kinases (Vervoorts and Lüscher, 2008): the letter map indicates that Y88 is the most important residue for this feature. Max F54 (Figure 7A.2) overlaps the nuclear localization sequence (NLS): previous work established four basic residues K153, R154, K165, and R166 are especially critical to the nuclear localization in site-directed mutation experiments, and consistent with this, K153 and R154 are predicted as the most important residues for this feature. Finally, Max F11 (Figure 7A.3) overlaps a key phosphorylation site T198 (Vervoorts and Lüscher, 2008): we observe that the feature strongly favors the arginine residues preceding the phosphorylation site (especially R195 and R196), and while T198 is generally favorable for this feature, S or H is also permissible in this position.

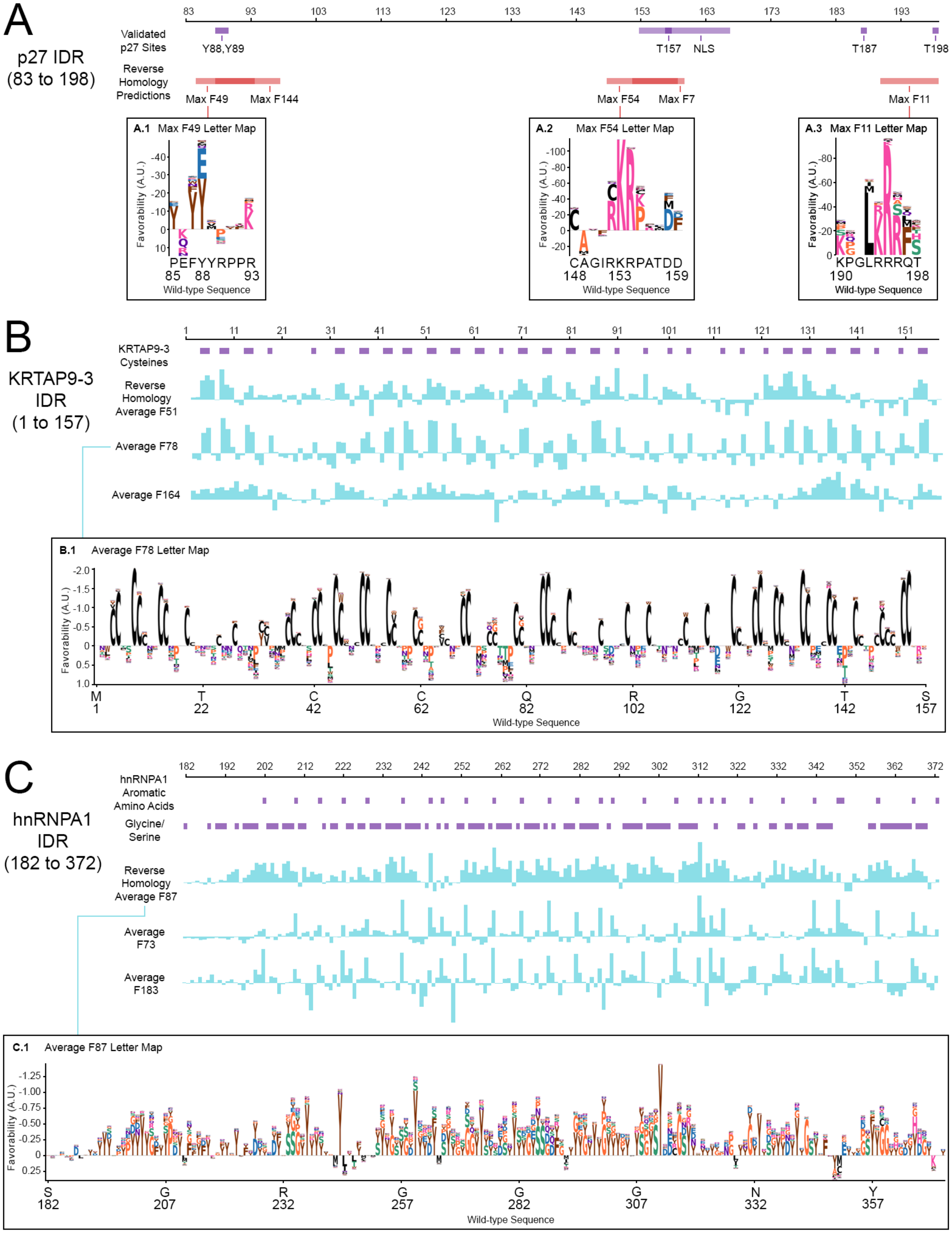


Figure 7. Summaries of known features (purple) compared to the top ranked reverse homology features (red and blue) for three IDRs, plus letter maps for selected features. We show the position of max pooled features in red (boundaries set using a cut-off of -10 or lower in magnitude), and the values of average

features in blue. Average features are sorted in descending order (i.e. the top ranked feature is at the top.) For the mutational scanning letter maps, positions above the axis are positions where retaining the original amino acid is generally preferable, while positions below the axis are positions where the feature could generally be improved by mutating to another amino acid. The height of the combined letters corresponds to the total magnitude of the change in the feature for all possible mutations (which we define as the favourability). For positions above the axis, we show amino acids that result in the highest value for the feature (i.e. the most favored amino acids at that position.) For positions below the axis, we show amino acids that result in the lowest value for the feature (i.e. the most disfavored amino acids at that position.)

Second, as a case of an IDR that is more mediated by bulk functions as opposed to motifs, we looked at hair keratin-associated proteins, many of which are characterized by abundant cysteine repeats that form disulfide cross-links (Wu, Irwin and Zhang, 2008). We chose KRTAP9-3 as an example, which has an N-terminal IDR spanning positions 1 to 157 in the sequence.

Figure 7B shows the cysteine residues in purple, and the top three features from our reverse homology model, which are all average features, in blue. We observed that the top three features all generally preferred the cysteine residues or their surrounding content. We show the mutational scanning letter map for Average F78 as in Figure B.1. While most of the cysteine residues are generally favored by this feature, the repeats of two cysteine residues are more generally favorable over single residues, as evident from their greater height in the letter map.

Third, as an example of a less characterized IDR, we examined hnRNPA1, which has an N-terminal IDR from positions 183 to 372. This IDR is known as a prion-like domain that facilitates liquid-liquid phase separation (Wang *et al.*, 2018). A recent study showed that the uniform patterning of aromatic residues in this IDR is critical to phase-separation, while also inhibiting aggregation (Martin *et al.*, 2020).

Consistent with these findings, we find that the top three features for our reverse homology model (all average features, shown in blue in Figure 7C) are all sensitive to the aromatic amino acids in the sequence of hnRNPA1. We show the *in-silico* mutational scanning map for Average F87 in Figure 7C.1, which shows that the feature is sensitive to YG or YS repeats: in many of the positions flanking tyrosine residues, glycine or serine is most favorable to the feature. While previous work did not analyze what amino acids are permissible as spacers between aromatic amino acids (Martin *et al.*, 2020), we observe that in the wild-type sequence, the tyrosine residues are often flanked by serine or glycine (purple in Figure 7C).

Overall, these cases demonstrate that our reverse homology model learns versatile features capable of identifying important features for both IDRs where function is mediated by short linear motifs like in p27, and where function is mediated by low-complexity repeats and patterning like in KRTAP9-3. Moreover, our method is effective for both well-characterized IDRs, and less or recently characterized IDRs like in hnRNPA1, suggesting that it can be used to generate hypotheses for IDRs that are currently poorly understood.

Discussion

We present, to our knowledge, the first proteome-wide evolutionary approach for feature discovery using neural networks. Compared to other systematic homology-based approaches for intrinsically disordered regions with motif finding methods (Davey *et al.*, 2012; Nguyen Ba *et al.*, 2012), our method discovers more flexible and expressive features than local motifs: we show that our models learn features like repeats or distributed charge properties, in addition to motifs. This expressiveness is important in the context of IDRs, where previous studies have shown that function is often mediated by global "bulk" properties (Zarin *et al.*, 2019). Like previous comparative proteomics methods, our method is systematic, in that it discovers a large set of features informative of many different functions in the proteome, and unbiased by prior knowledge, in that it relies only on automatically-assigned sequence homology to discover features. The latter property sets our method apart from deep learning approaches on protein sequences that use regression problems to train models, such as a recent study that discovered features of disordered activation domains by training deep learning models to predict the results of a transcriptional activation assay (Erijman *et al.*, 2020). We argue that optimizing models to predict prior knowledge of function, or assay measurements that reflect specific aspects of function, will lead to the model learning features for these functions exclusively. In contrast, training a model to predict evolutionary homology yields a potentially more general set of features that are conserved over evolution.

In many cases, our method learns features that are highly consistent with consensus motifs or bulk features previously known: this congruence is exciting because the model learns independent of prior knowledge, so "re-discovering" this biology supports the claim that our models are learning biological signal. At the same time, even when the model learns features that are consistent with prior expert-defined features, there is often additional subtlety or depth. For example, for our yeast RG-repeat feature Average F65, we showed that the feature has an additional preference for aromatic amino acids, consistent with the recent knowledge that these amino acids mediate similar interactions to RG-repeats in these sequences (Chong, Vernon and Forman-Kay, 2018; Kharel *et al.*, 2020). Similarly, we showed that our model develops two subclasses of PKA-like consensus motifs, and one is more sensitive to double phosphorylation sites. These examples demonstrate the power of unsupervised analysis to refine previous knowledge.

From a computational biology perspective, we note that many of the individual feature analyses we presented in this study resemble bioinformatics studies that make functional predictions based on conserved motifs or other features (Beltrao and Serrano, 2005). For example, our analysis of PKA phosphorylation sites parallels a previous evolutionary proteomics study that systematically identified PKA substrates (Budovskaya *et al.*, 2005): of the 25 of 92 conserved PKA motifs identified that are present in IDRs, our automatically learned yeast PKA feature Max F231 overlaps 10 in the 139 most activated IDRs, suggesting that our feature is in good agreement with this previous study and can also be used to identify putative modification sites. Unlike these studies that start with a known feature and search for new predictions, our approach learns many features in parallel, without having to pre-specify motifs/features of interest. In principle, these bioinformatics analyses can be applied to all of the 512 features learned by our model, enabling hypothesis discovery at an unprecedented scale. (We do note that many of the features learned by our model appear to be redundant with each other (see our annotations in Supplementary Files 1 and 4), so this is an upper bound.)

Perhaps more exciting, our analysis of individual regions (such as in p27 shown above) indicate that unsupervised deep learning approaches like reverse homology, paired with appropriate interpretation methods, will lead to highly specific predictions of functional residues and regions within IDRs. This would represent an urgently needed advance, especially for IDRs that are mutated in disease, for which we have few mechanistic hypotheses about function (Vacic *et al.*, 2012; Pritišanac *et al.*, 2019; Tsang *et al.*, 2020; Lindorff-Larsen and Kragelund, 2021).

From a technical perspective, reverse homology employs a self-supervised approach, as many emerging representation learning approaches for protein sequences do (Alley *et al.*, 2019; Heinzinger *et al.*, 2019; Rao *et al.*, 2019; Lu *et al.*, 2020; Rives *et al.*, 2021). Unlike these methods, which are mostly based on methods adapted from natural language processing, we proposed a novel proxy task that purposes principles of evolutionary proteomics as a learning signal instead (Lu, Lu and Moses, 2020). Another distinction in our study is that previous approaches primarily focus on representation learning, with the aim of optimizing the performance of the representation on downstream regression tasks reflecting protein design or classification problems. In contrast, we focus more on feature discovery in this study. We argue that representation learning and feature discovery are distinct aims that require different design philosophies. For example, in this study, we employed a lightweight convolutional architecture, because the interpretation of features is a necessary property. Moreover, we preprocessed the data to remove global information like sequence length or whether the sequence was at the N-terminal: while this information is often useful for downstream tasks, we observed that our models learn fewer “interesting” local features without this preprocessing. In other words, while representation learning does not care about what features are learned as long as they contribute signal to downstream classification or regression problems, we designed our feature discovery approach to learn general, interpretable features that would reflect the biology of IDRs.

However, as deep learning architectures are developed for protein sequences, and as new interpretation methods are designed to complement these architectures, updated implementations of our method with these architectures are also possible. Currently, a major limitation of our convolutional neural network architecture is that it does not capture distal interactions. However, transformer architectures are capable of modeling distal interactions, and there has been progress in making these models tailored to multiple sequence alignments, and interpreting the self-attention modules (Rao *et al.*, 2021). A second limitation is that our convolutional model requires sequences to be standardized in length as input. This preprocessing requirement means that we may lose key elements of longer sequences. Shorter sequences require padding; we used “repeat” padding, since we found with a special padding token the neural network can use cues about length to trivially eliminate many possible proteins from the contrastive task, but this runs the risk of creating new spurious repeats. Recurrent architectures address this limitation and allow for arbitrary-sized inputs (Alley *et al.*, 2019; Heinzinger *et al.*, 2019). Overall, integrating these kinds of advances with our method in future work is expected to make the model more expressive, increasing the scope of the features we can discover.

Methods

Details of Reverse Homology

In this paper, we studied sets of automatically obtained homologous IDRs. We concentrate on IDRs in this work, but our contrastive learning task can easily be extended to other definitions of homologous sequences including full protein sequences or structured domains. We will use these sets of homologous sequences as the basis of our self-supervised task.

Let $H_i = \{s_{i,1}, \dots, s_{i,n}\}$ be a set of homologous sequences. We define a set of query sequences, , S_q such that all sequences in the query set are homologous to each other, so $S_q \subset H_i$. Then, we define a set of target sequences associated with the query set, $S_t = \{s_{t-,1}, \dots, s_{t-,m-1}\} \cup \{s_{t+}\}$ where s_{t+} is a held-out homologue $s_{t+} \in H_i, s_{t+} \notin S_q$ and s_{t-} are not homologous to the query set, $s_{t+} \in H_j, j \neq i$.

Let g_1 be a function that embeds S_q into a latent feature representation, so $g_1(S_q) = z_1$, and g_2 be a function that embeds members of S_t into a latent feature representation, so $g_2(s_t) = z_2$ (in this work, g_1 and g_2 are convolutional neural network encoders.) Our task is to optimize the categorical cross-entropy loss, also commonly called the InfoNCE loss in contrastive learning literature (Oord, Li and Vinyals, 2018), where $f\left(g_1(S_q), g_2(s_t)\right)$ is a score function (in this work, we use the dot product):

$$\mathcal{L}_{NCE} = -\mathbb{E}_{S_q, s_{t+}, s_{t-}}\left[log \frac{\exp(f\left(g_1(S_q), g_2(s_{t+})\right))}{\exp\left(f\left(g_1(S_q), g_2(s_{t+})\right)\right) + \sum_{j=1}^{M-1} \exp(f\left(g_1(S_q), g_2(s_{t-,j})\right))} \right]$$

Implementation of reverse homology

In principle, the sequence encoders g_1 and g_2 are flexibly defined, and many neural network architectures are possible here; previous self-supervised learning methods on protein sequences have generally favored large transformer or LSTM models due to their ability to capture distal interactions in sequences (Alley *et al.*, 2019; Heinzinger *et al.*, 2019; Rao *et al.*, 2019). However, since a priority of this work is interpreting the features learned by our model, not necessarily to learn the most useful or complete representation possible, we decided to implement our encoders as low-parameter convolutional neural networks (CNNs). Many interpretation methods designed for neural networks trained on biological sequences are more specific to CNNs: Koo and Eddy propose a method for generating motif-like visualizations that involves collecting the parts of sequences that maximally activate neurons to calculate position frequency matrices (Koo and Eddy, 2019). Other interpretation methods are less specific to architecture, but benefit from efficient implementations: Alipanahi *et al*. propose systematically introducing every possible point mutation in a sequence (Alipanahi *et al.*, 2015), which requires $20 \times L$ inputs to the model for each protein sequence we want to interpret (where L is the length of a sequence). We reasoned a model with fewer parameters and layers would simply run faster at inference time.

We show a summary of our architecture in Figure 1D. Both encoders g_1 and g_2 begin with three convolutional layers; following Almagro Armenteros *et al*. (Almagro Armenteros *et al.*, 2017), the first layer contains neurons with different kernel sizes (1, 3, and 5). After the convolutional layers, we max and average pool convolutional features over the length of the entire sequence. The max and average-pooled features are concatenated and fed into two fully-connected layers; we scale the average-pooled features by a factor of the post-processed input sequence length divided by the receptive field of the final convolutional layer (17.06 times in this specific architecture) to put the average and max pooled features on the same numerical scale. The output of the final fully connected layer is considered the feature representation. We average the feature representation for all homologues in the query set S_q, and calculate the dot product between this average and the representation for each sequence in the target set s_t. This dot product is considered our score function $f\left(g_1(S_q), g_2(s_t)\right)$ in the InfoNCE loss (i.e. the largest dot product is considered the model's prediction of which sequence in the target set is homologous to the sequences in the query set.)

For our implementation, we use a query set size of 8, and a target set size of 400. The size of the query set should, in principle, control the difficulty of the pretext task, as well as the kinds of features learned: with larger query sets, we expect that information that is incidental to any one homologue and not shared across all homologues will be averaged out. The size of the target set has a similar effect, as the model has to distinguish the homologous sequence from greater or fewer non-homologous sequences depending on the setting of this parameter: Oord *et al*. show that in theory, increasing the size of the target set tightens the lower bound on maximizing mutual information (Oord, Li and Vinyals, 2018). In this study, we set these parameters to reasonable defaults, and leave a full exploration of these parameters to future work.

Training datasets

To train our model, we used sets of homologous yeast IDRs previously defined by Zarin *et al*. (Zarin *et al.*, 2020). Briefly, this dataset was produced by aligning orthologues (homologues from different species) of yeast proteins previously calculated by the Yeast Gene Order Browser (Byrne and Wolfe, 2005). Alignments shown in Figure 1 are visualized using Jalview (Clamp *et al.*, 2004); note that the alignment is only used to identify the boundaries of the IDRs across species, and not supplied to the models during training (i.e. the models are given the unaligned IDR sequence). DISOPRED3 was used to annotate IDRs in *Saccharomyces cerevisiae* sequences, and residues in other species that fell within these regions were considered homologous after some quality control steps (see (Zarin *et al.*, 2019) for details.) We filtered this dataset by removing sequences under 5 amino acids, with undetermined amino acids (“X”) and/or non-standard amino acids, and only kept homologue families with more than 9 sequences represented. Overall, this dataset consists of a total of 94,106 IDR sequences distributed across 5,306 sets of homologues.

In addition to our yeast reverse homology model, we produced a human model trained on sets of homologous vertebrate IDRs. Homologous protein annotations for vertebrates were obtained from the OMA homology database (Altenhoff *et al.*, 2021). UniProt reference human protein sequences were downloaded on September 2019 (UniProt Consortium, 2019), and used for disorder prediction using SPOT-Disorder v1 (Hanson *et al.*, 2017). These sequences were aligned using MAFFT (Katoh *et al.*, 2002), and the disorder boundaries of the human sequence were used as a reference to annotate putative disordered regions in the set of vertebrate homologs. After the same pre-processing operations for length, undetermined/non-standard amino acids, and number of sequences that we used for the yeast model, this dataset consists of 1,702,348 sequences distributed across 17,519 sets of homologues. Finally, to ensure that the sequences did not contain any structured domains, we filtered the sequences based on matches in Prosite (Sigrist *et al.*, 2013). We removed any homology families where the *Homo sapiens* IDR had a match in Prosite above 10 amino acids: this operation led to a total of 1,604,052 sequences across 16,328 human IDR homology families.

Preprocessing and training

We one-hot encoded sequences as input into our models. To standardize the lengths of the sequences, if the sequence was longer than 256 amino acids, we used the first and last 128 amino acids from the sequence. If the sequence was shorter, we "repeat padded" the sequence until it was over 256 amino acids (e.g. in this operation "ACD" becomes "ACDACD" after the first repetition), and clipped off excess length at the end of the padded sequence.

In our preprocessing operations, we sought to reduce the impact of certain global properties, which while may be biologically informative, would allow the model to rule out the majority of non-homologous sequences in the target set on the basis of relatively trivial features for most query sets, reducing the effectiveness of our contrastive task. One is the length of the sequence, motivating our use of the repeat padding operation, which reduces cues about length compared to the use of a special padding token. The other global feature we identified was whether the IDR occurs at the start of a protein or not, as indicated by a methionine (from the start codon) at the beginning of the IDR. We clipped this methionine from the sequence if the IDR was at the N-terminal of a protein.

We trained models for 1,000 epochs, where each epoch iterates over all sets of homologues in the training dataset. For each set of homologues, we randomly drew 8 unique sequences at each epoch to form the query set and 1 non-overlapping sequence for the target set. To save memory and speed up training, we used a shared target set for each batch of query sets: homologous sequences for one query set in the batch would be considered as non-homologous sequences for the other query sets. If the target set size is larger than the batch size (as it was in our experiments), the remainder of non-homologous sequences are sampled at random from homologue sets not used in the batch. We trained models with a batch size of 64, and a learning rate of 1e-4.

Correlation with literature-curated features

To compare our features against literature-curated features, we binarized all amino acid positions in all of the IDR sequences in our yeast data using each of these regular expressions. Amino acids that are contained in a match to the regular expression are assigned a value of 1, while all other amino acids are assigned a value of 0. We calculated the global correlation between these binarized positions and the activation value of neurons in our convolutional neural network at each position. A higher correlation indicates that a neuron outputs high feature values at positions that match the regular expression of a literature-correlated feature and low values at positions that do not match.

Interpretation

To interpret the features learned by our model, we adapted two previous interpretation methods. We also reported the enrichment of features for GO enrichments: note that these are done using a background of all proteins with IDRs (not all proteins), to avoid spurious enrichments.

Sequence logos summarize features

First, to produce a global summary of the kinds of sequences that activate each neuron in a layer of our model, we adapted a method from Koo and Eddy (Koo and Eddy, 2019). This method visualizes neurons by scanning them against every single sequence in a dataset (in our case, the full dataset of yeast IDRs). We collect sequences that reach at least 70% of the maximum activation for that neuron. If there are less than 20 sequences meeting this threshold, we instead collect the 20 highest activating sequences. These sequences are used to produce position frequency matrices (PFMs). For max-pooled features, we collect the maximally activating subsequence, and add all amino acids in this window to the PFM with equal weight. For average-pooled features, we add all windows to the PFM, but weigh all windows in the sequence by the activation for that window divided by the activation of the maximally activating window in that sequence. These PFMs are converted to a position probability matrix and visualized as sequence logos using the Biopython package, modified with a custom color scheme (Cock *et al.*, 2009). Unlike Koo and Eddy (Koo and Eddy, 2019), we do not discard windows that overlap with the start or end of a sequence, to avoid too few inputs due to our larger receptive field and smaller sequence sizes: we simply do not let parts of the sequence overlapping the start and end of the sequence contribute any frequency to their corresponding position in the position frequency matrix.

Overall, this method produces a sequence logo for each neuron, summarizing the kinds of subsequences that activate the neuron. This method can only be applied to convolutional layers before pooling, because it requires us to measure the activation at specific positions in a sequence; in our experiments, we apply it to the final convolutional layer in our models (Conv3 in Figure 3).

In-silico mutational scanning visualizes specific features and sequences

The activation scanning method produces global summaries of the kinds of sequences that activate each neuron but may fail to reveal more nuanced aspects of these neurons. For example, some of our neurons appear to be sensitive to multiple patterns in sequences: the activation scanning logos may only show the most common pattern, with less frequent patterns not as visible. We reasoned that a good way to reveal subtle details in our neurons would be to systematically produce every possible point mutation in a sequence, and measure the effects of the mutation on the output activation of a neuron, inspired by the method previously employed by Alipanahi *et al*. (Alipanahi *et al.*, 2015).

For a given feature and IDR, we systematically mutate each amino acid in the IDR sequence to each other amino acid and measure the change in the value of the feature. We visualize this matrix in two ways. First, we visualize the entire matrix as a heat map. Second, we visualize the mutational scanning map as a sequence logo, which we term a letter map to distinguish them from the per-feature sequence logos. In this letter map, any amino acids that would generally reduce the value of the feature if mutated is shown above the axis, while any amino acids that would generally increase the value of the feature if mutated are shown below the axis. The combined height of the letters corresponds to the overall magnitude of the increase or decrease mutating the position would induce on the feature. For positions above the axis, we show the amino acids that are most permissible in that position. For positions below the axis, we show the amino acids that are least permissible in that position. In summary, letters above the axis are favored residues in favored positions, while letters below the axis are disfavored residues in disfavored positions. We used the Logomaker package, modified with a custom color scheme, to visualize these letter maps (Tareen and Kinney, 2020). More details and formulas for these letter maps are available in Supplementary Methods.

Code and Data Availability

Code for training our models and visualizing/extracting features is available under a CC-BY license at github.com/alexxijielu/reverse_homology/. Pretrained weights for our models, fasta files of IDR sequences used to train both models, and labels for IDRs used in our classification benchmarks are available at zenodo.org/record/5146063.

Acknowledgements

We thank Caroline Uhler, Karren Dai Yang, and Kevin Yang for valuable feedback. This work was funded by an NSERC-CGS grant. This work was partially performed on a GPU generously donated by NVIDIA.

References

Abbastabar, M. *et al.* (2018) 'Multiple functions of p27 in cell cycle, apoptosis, epigenetic modification and transcriptional regulation for the control of cell growth: A double-edged sword protein', *DNA Repair*. DNA Repair (Amst), pp. 63–72. doi: 10.1016/j.dnarep.2018.07.008.
Abramova, N. *et al.* (2001) 'Reciprocal regulation of anaerobic and aerobic cell wall mannoprotein gene expression in Saccharomyces cerevisiae', *Journal of Bacteriology*. J Bacteriol, 183(9), pp. 2881–2887. doi: 10.1128/JB.183.9.2881-2887.2001.
Alipanahi, B. *et al.* (2015) 'Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning', *Nature Biotechnology*. Nature Publishing Group, 33(8), pp. 831–838. doi: 10.1038/nbt.3300.
Alley, E. C. *et al.* (2019) 'Unified rational protein engineering with sequence-based deep representation learning', *Nature Methods*. Nature Research, 16(12), pp. 1315–1322. doi: 10.1038/s41592-019-0598-1.
Almagro Armenteros, J. J. *et al.* (2017) 'DeepLoc: prediction of protein subcellular localization using deep learning', *Bioinformatics (Oxford, England)*. Oxford Academic, 33(21), pp. 3387–3395. doi: 10.1093/bioinformatics/btx431.
Altenhoff, A. M. *et al.* (2021) 'OMA orthology in 2021: Website overhaul, conserved isoforms, ancestral gene order and more', *Nucleic Acids Research*. Oxford University Press, 49(D1), pp. D373–D379. doi: 10.1093/nar/gkaa1007.
Avsec, Ž. *et al.* (2021) 'Base-resolution models of transcription-factor binding reveal soft motif syntax', *Nature*

Genetics. Nature Research, 53(3), pp. 354–366. doi: 10.1038/s41588-021-00782-6.

Bauer, N. C., Doetsch, P. W. and Corbett, A. H. (2015) ‘Mechanisms Regulating Protein Localization’, *Traffic*, 16(10), pp. 1039–1061. doi: 10.1111/tra.12310.

Beh, L. Y., Colwell, L. J. and Francis, N. J. (2012) ‘A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence’, *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 109(18). doi: 10.1073/pnas.1118678109.

Bellay, J. *et al.* (2011) ‘Bringing order to protein disorder through comparative genomics and genetic interactions’, *Genome Biology*. Genome Biol, 12(2). doi: 10.1186/gb-2011-12-2-r14.

Beltrao, P. and Serrano, L. (2005) ‘Comparative Genomics and Disorder Prediction Identify Biologically Relevant SH3 Protein Interactions’, *PLoS Computational Biology*. Public Library of Science (PLoS), 1(3), p. e26. doi: 10.1371/journal.pcbi.0010026.

Budovskaya, Y. V. *et al.* (2005) ‘An evolutionary proteomics approach identifies substrates of the cAMP-dependent protein kinase’, *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 102(39), pp. 13933–13938. doi: 10.1073/pnas.0501046102.

Byrne, K. P. and Wolfe, K. H. (2005) ‘The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species’, *Genome Research*, 15(10), pp. 1456–1461.

Chen, J. W. *et al.* (2006a) ‘Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions’, *Journal of Proteome Research*. American Chemical Society, 5(4), pp. 879–887. doi: 10.1021/pr060048x.

Chen, J. W. *et al.* (2006b) ‘Conservation of intrinsic disorder in protein domains and families: II. Functions of conserved disorder’, *Journal of Proteome Research*. American Chemical Society, 5(4), pp. 888–898. doi: 10.1021/pr060049p.

Chen, T. *et al.* (2020) ‘A Simple Framework for Contrastive Learning of Visual Representations’, *ICLR 2020*. Available at: http://arxiv.org/abs/2002.05709 (Accessed: 28 October 2020).

Chong, P. A., Vernon, R. M. and Forman-Kay, J. D. (2018) ‘RGG/RG Motif Regions in RNA Binding and Phase Separation’, *Journal of Molecular Biology*. Academic Press, pp. 4650–4665. doi: 10.1016/j.jmb.2018.06.014.

Clamp, M. *et al.* (2004) ‘The Jalview Java alignment editor’, *Bioinformatics*. Bioinformatics, 20(3), pp. 426–427. doi: 10.1093/bioinformatics/btg430.

Cock, P. J. A. *et al.* (2009) ‘Biopython: Freely available Python tools for computational molecular biology and bioinformatics’, *Bioinformatics*. Oxford Academic, 25(11), pp. 1422–1423. doi: 10.1093/bioinformatics/btp163.

Colak, R. *et al.* (2013) ‘Distinct Types of Disorder in the Human Proteome: Functional Implications for Alternative Splicing’, *PLoS Computational Biology*. PLoS Comput Biol, 9(4). doi: 10.1371/journal.pcbi.1003030.

Das, M. K. and Dai, H. K. (2007) ‘A survey of DNA motif finding algorithms’, in *BMC Bioinformatics*. BioMed Central, pp. 1–13. doi: 10.1186/1471-2105-8-S7-S21.

Das, R. K. *et al.* (2016) ‘Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 113(20), pp. 5616–5621. doi: 10.1073/pnas.1516277113.

Das, R. K. and Pappu, R. V. (2013) ‘Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues’, *Proceedings of the National Academy of Sciences of the United States of America*. PNAS, 110(33), pp. 13392–13397. doi: 10.1073/pnas.1304749110.

Davey, N. E. *et al.* (2012) ‘SLiMPrints: Conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions’, *Nucleic Acids Research*. Oxford University Press, 40(21), pp. 10628–10641. doi: 10.1093/nar/gks854.

Davey, N. E. (2019) ‘The functional importance of structure in unstructured protein regions’, *Current Opinion in Structural Biology*. Elsevier Ltd, pp. 155–163. doi: 10.1016/j.sbi.2019.03.009.

Dengler, L. *et al.* (2021) ‘Regulation of trehalase activity by multi-site phosphorylation and 14-3-3 interaction’, *Scientific Reports*. Nature Research, 11(1), pp. 1–14. doi: 10.1038/s41598-020-80357-3.

Dhaval Vaishnav, E. *et al.* (2021) ‘A comprehensive fitness landscape model reveals the evolutionary history and future evolvability of eukaryotic cis-regulatory DNA sequences’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 2021.02.17.430503. doi: 10.1101/2021.02.17.430503.

Eden, E. *et al.* (2009) ‘GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists’, *BMC Bioinformatics*. BioMed Central, 10(1), p. 48. doi: 10.1186/1471-2105-10-48.

Edwards, T. K. *et al.* (2000) ‘Role for nucleolin/Nsr1 in the cellular localization of topoisomerase I’, *Journal of Biological Chemistry*. Elsevier, 275(46), pp. 36181–36188. doi: 10.1074/jbc.M006628200.

Erijman, A. *et al.* (2020) ‘A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning’, *Molecular Cell*. Cell Press, 78(5), pp. 890-902.e6. doi:

10.1016/j.molcel.2020.04.020.

Gagnon, K. T. *et al.* (2012) 'Structurally conserved Nop56/58 N-terminal domain facilitates archaeal box C/D ribonucleoprotein-guided methyltransferase activity', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 287(23), pp. 19418–19428. doi: 10.1074/jbc.M111.323253.

Gallego, L. D. *et al.* (2020) 'Phase separation directs ubiquitination of gene-body nucleosomes', *Nature*. Nature Research, 579(7800), pp. 592–597. doi: 10.1038/s41586-020-2097-z.

González, M., Brito, N. and González, C. (2012) 'High abundance of Serine/Threonine-rich regions predicted to be hyper-O-glycosylated in the secretory proteins coded by eight fungal genomes', *BMC Microbiology*. BioMed Central, 12(1), pp. 1–10. doi: 10.1186/1471-2180-12-213.

Hanson, J. *et al.* (2017) 'Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks', *Bioinformatics*. Oxford University Press, 33(5), pp. 685–692. doi: 10.1093/bioinformatics/btw678.

Hardison, R. C. (2003) 'Comparative genomics', *PLoS Biology*. Public Library of Science, p. e58. doi: 10.1371/journal.pbio.0000058.

Heinzinger, M. *et al.* (2019) 'Modeling aspects of the language of life through transfer-learning protein sequences', *BMC Bioinformatics*. BioMed Central, 20(1), p. 723. doi: 10.1186/s12859-019-3220-8.

Howe, K. L. *et al.* (2021) 'Ensembl 2021', *Nucleic Acids Research*. Oxford University Press, 49(D1), pp. D884–D891. doi: 10.1093/nar/gkaa942.

Jing, L. and Tian, Y. (2019) 'Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey'. Available at: http://arxiv.org/abs/1902.06162 (Accessed: 30 September 2019).

Jones, D. T. and Cozzetto, D. (2015) 'DISOPRED3: Precise disordered region predictions with annotated protein-binding activity', *Bioinformatics*. Oxford University Press, 31(6), pp. 857–863. doi: 10.1093/bioinformatics/btu744.

Katoh, K. *et al.* (2002) 'MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic Acids Research*. Oxford University Press, 30(14), pp. 3059–3066. doi: 10.1093/nar/gkf436.

Kharel, P. *et al.* (2020) 'Properties and biological impact of RNA G-quadruplexes: From order to turmoil and back', *Nucleic Acids Research*. Oxford University Press, pp. 12534–12555. doi: 10.1093/nar/gkaa1126.

Koo, P. K. and Eddy, S. R. (2019) 'Representation learning of genomic sequence motifs with convolutional neural networks', *PLOS Computational Biology*. Edited by J. Listgarten. Public Library of Science, 15(12), p. e1007560. doi: 10.1371/journal.pcbi.1007560.

Kulkarni, P. and Uversky, V. N. (2018) 'Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome', *Proteomics*. Wiley-VCH Verlag, 18(21–22). doi: 10.1002/pmic.201800061.

Kumar, M. *et al.* (2020) 'ELM-the eukaryotic linear motif resource in 2020', *Nucleic Acids Research*. Nucleic Acids Res, 48(D1), pp. D296–D306. doi: 10.1093/nar/gkz1030.

Kummerfeld, S. K. and Teichmann, S. A. (2009) 'Protein domain organisation: Adding order', *BMC Bioinformatics*. BioMed Central, 10(1), p. 39. doi: 10.1186/1471-2105-10-39.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*. Nature Publishing Group, 521(7553), pp. 436–444. doi: 10.1038/nature14539.

Van Der Lee, R. *et al.* (2014) 'Classification of intrinsically disordered regions and proteins', *Chemical Reviews*. American Chemical Society, pp. 6589–6631. doi: 10.1021/cr400525m.

Lindorff-Larsen, K. and Kragelund, B. B. (2021) 'On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins'. Available at: http://arxiv.org/abs/2106.00855 (Accessed: 10 June 2021).

Liu, X. *et al.* (2020) 'Self-supervised Learning: Generative or Contrastive', *arXiv*. Available at: http://arxiv.org/abs/2006.08218 (Accessed: 2 October 2020).

Lu, A. X. *et al.* (2020) 'Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2020.09.04.283929. doi: 10.1101/2020.09.04.283929.

Lu, Amy X., Lu, Alex X. and Moses, A. (2020) 'Evolution Is All You Need: Phylogenetic Augmentation for Contrastive Learning'. Available at: http://arxiv.org/abs/2012.13475 (Accessed: 2 June 2021).

Lu, T., Lu, A. X. and Moses, A. M. (2021) 'Random Embeddings and Linear Regression can Predict Protein Function'. Available at: https://arxiv.org/abs/2104.14661v1 (Accessed: 12 July 2021).

Martin, E. W. *et al.* (2020) 'Valence and patterning of aromatic residues determine the phase behavior of prion-like domains', *Science*. Science, 367(6478), pp. 694–699. doi: 10.1126/science.aaw8653.

McInnes, L., Healy, J. and Melville, J. (2018) 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'. Available at: http://arxiv.org/abs/1802.03426 (Accessed: 25 June 2019).

Mészáros, B., Erdős, G. and Dosztányi, Z. (2018) 'IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding', *Nucleic Acids Research*. Oxford Academic, 46(W1), pp. W329–W337.

doi: 10.1093/NAR/GKY384.

Mohamed, S. A. E. H., Elloumi, M. and Thompson, J. D. (2016) 'Motif Discovery in Protein Sequences', in *Pattern Recognition - Analysis and Applications*. InTech. doi: 10.5772/65441.

Moses, A. M. *et al.* (2007) 'Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 104(45), pp. 17713–17718. doi: 10.1073/pnas.0700997104.

Moses, A. M., Hériché, J.-K. and Durbin, R. (2007) 'Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase', *Genome Biology2*, 8, p. R23.

Nguyen Ba, A. N. *et al.* (2012) 'Proteome-wide discovery of evolutionary conserved sequences in disordered regions', *Science Signaling*. Sci Signal, 5(215). doi: 10.1126/scisignal.2002515.

Oord, A. van den, Li, Y. and Vinyals, O. (2018) 'Representation Learning with Contrastive Predictive Coding'. Available at: http://arxiv.org/abs/1807.03748 (Accessed: 27 October 2020).

Oruganti, S. *et al.* (2007) 'Alternative Conformations of the Archaeal Nop56/58-Fibrillarin Complex Imply Flexibility in Box C/D RNPs', *Journal of Molecular Biology*. Academic Press, 371(5), pp. 1141–1150. doi: 10.1016/j.jmb.2007.06.029.

Oughtred, R. *et al.* (2021) 'The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions', *Protein Science*. Blackwell Publishing Ltd, 30(1), pp. 187–200. doi: 10.1002/pro.3978.

Pautasso, C. *et al.* (2016) 'Identification of novel transcriptional regulators of PKA subunits in Saccharomyces cerevisiae by quantitative promoter-reporter screening', *FEMS Yeast Research*. Oxford University Press, 16(5), p. 46. doi: 10.1093/femsyr/fow046.

Pearson, W. R. (2013) 'An introduction to sequence similarity ("homology") searching', *Current Protocols in Bioinformatics*. NIH Public Access, 0 3(SUPPL.42). doi: 10.1002/0471250953.bi0301s42.

Poornima, G. *et al.* (2019) 'RGG-motif self-association regulates eIF4G-binding translation repressor protein Scd6', *RNA Biology*. Taylor and Francis Inc., 16(9), pp. 1215–1227. doi: 10.1080/15476286.2019.1621623.

Pritišanac, I. *et al.* (2019) 'Entropy and information within intrinsically disordered protein regions', *Entropy*. MDPI AG, 21(7), p. 662. doi: 10.3390/e21070662.

Rao, R. *et al.* (2019) 'Evaluating Protein Transfer Learning with TAPE', *NeurIPS 2019*. Available at: http://arxiv.org/abs/1906.08230 (Accessed: 2 October 2020).

Rao, R. *et al.* (2021) 'MSA Transformer', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2021.02.12.430858. doi: 10.1101/2021.02.12.430858.

Riback, J. A. *et al.* (2017) 'Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response.', *Cell*. Elsevier, 168(6), pp. 1028-1040.e19. doi: 10.1016/j.cell.2017.02.027.

Ritch, J. J. *et al.* (2010) 'The Saccharomyces SUN gene, UTH1, is involved in cell wall biogenesis', *FEMS Yeast Research*. FEMS Yeast Res, 10(2), pp. 168–176. doi: 10.1111/j.1567-1364.2009.00601.x.

Rives, A. *et al.* (2021) 'Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 118(15). doi: 10.1073/pnas.2016239118.

Sanborn, A. L. *et al.* (2021) 'Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to mediator', *eLife*. eLife Sciences Publications Ltd, 10. doi: 10.7554/ELIFE.68068.

Sawle, L. and Ghosh, K. (2015) 'A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins', *Journal of Chemical Physics*. American Institute of Physics Inc., 143(8). doi: 10.1063/1.4929391.

Schwartz, D. and Gygi, S. P. (2005) 'An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets', *Nature Biotechnology*. Nat Biotechnol, 23(11), pp. 1391–1398. doi: 10.1038/nbt1146.

Shanehsazzadeh, A., Belanger, D. and Dohan, D. (2020) 'Is Transfer Learning Necessary for Protein Landscape Prediction?' Available at: http://arxiv.org/abs/2011.03443 (Accessed: 10 June 2021).

Sigrist, C. J. A. *et al.* (2013) 'New and continuing developments at PROSITE', *Nucleic Acids Research*. Oxford Academic, 41(D1), pp. D344–D347. doi: 10.1093/NAR/GKS1067.

Smith, F. D., Samelson, B. K. and Scott, J. D. (2011) 'Discovery of cellular substrates for protein kinase A using a peptide array screening protocol', *Biochemical Journal*. NIH Public Access, 438(1), pp. 103–110. doi: 10.1042/BJ20110720.

Stollar, E. J. *et al.* (2009) 'Structural, functional, and bioinformatic studies demonstrate the crucial role of an extended peptide binding site for the SH3 domain of yeast Abp1p', *Journal of Biological Chemistry*. J Biol Chem, 284(39), pp. 26918–26927. doi: 10.1074/jbc.M109.028431.

Tareen, A. and Kinney, J. B. (2020) ‘Logomaker: Beautiful sequence logos in Python’, *Bioinformatics*. Oxford University Press, 36(7), pp. 2272–2274. doi: 10.1093/bioinformatics/btz921.

Toda, T. *et al.* (1987) ‘Three different genes in S. cerevisiae encode the catalytic subunits of the cAMP-dependent protein kinase’, *Cell*. Cell, 50(2), pp. 277–287. doi: 10.1016/0092-8674(87)90223-6.

Tsang, B. *et al.* (2020) ‘Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations’, *Cell*. Cell Press, 183(7), pp. 1742–1756. doi: 10.1016/J.CELL.2020.11.050.

UniProt Consortium (2019) ‘UniProt: A worldwide hub of protein knowledge’, *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D506–D515. doi: 10.1093/nar/gky1049.

Van der Vaart, J. M. *et al.* (1995) ‘Identification of three mannoproteins in the cell wall of Saccharomyces cerevisiae’, *Journal of Bacteriology*. American Society for Microbiology, 177(11), pp. 3104–3110. doi: 10.1128/jb.177.11.3104-3110.1995.

Vacic, V. *et al.* (2012) ‘Disease-Associated Mutations Disrupt Functionally Important Regions of Intrinsic Protein Disorder’, *PLoS Computational Biology*. PLoS Comput Biol, 8(10). doi: 10.1371/journal.pcbi.1002709.

Verna, J. *et al.* (1997) ‘A family of genes required for maintenance of cell wall integrity and for the stress response in Saccharomyces cerevisiae’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 94(25), pp. 13804–13809. doi: 10.1073/pnas.94.25.13804.

Vervoorts, J. and Lüscher, B. (2008) ‘Post-translational regulation of the tumor suppressor p27KIP1’, *Cellular and Molecular Life Sciences*. Cell Mol Life Sci, pp. 3255–3264. doi: 10.1007/s00018-008-8296-7.

Wang, J. *et al.* (2018) ‘A molecular grammar governing the driving forces for phase separationof prion-like RNA binding proteins’, *Cell*. NIH Public Access, 174(3), p. 688. doi: 10.1016/J.CELL.2018.06.006.

Welter, E. *et al.* (2013) ‘Uth1 is a mitochondrial inner membrane protein dispensable for post-log-phase and rapamycin-induced mitophagy’, in *FEBS Journal*. FEBS J, pp. 4970–4982. doi: 10.1111/febs.12468.

Wright, P. E. and Dyson, H. J. (2015) ‘Intrinsically disordered proteins in cellular signalling and regulation’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 16(1), pp. 18–29. doi: 10.1038/nrm3920.

Wu, D.-D., Irwin, D. M. and Zhang, Y.-P. (2008) ‘Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair’, *BMC Evolutionary Biology 2008 8:1*. BioMed Central, 8(1), pp. 1–15. doi: 10.1186/1471-2148-8-241.

Xie, X. *et al.* (2005) ‘Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals’, *Nature*. Nature Publishing Group, 434(7031), pp. 338–345. doi: 10.1038/nature03441.

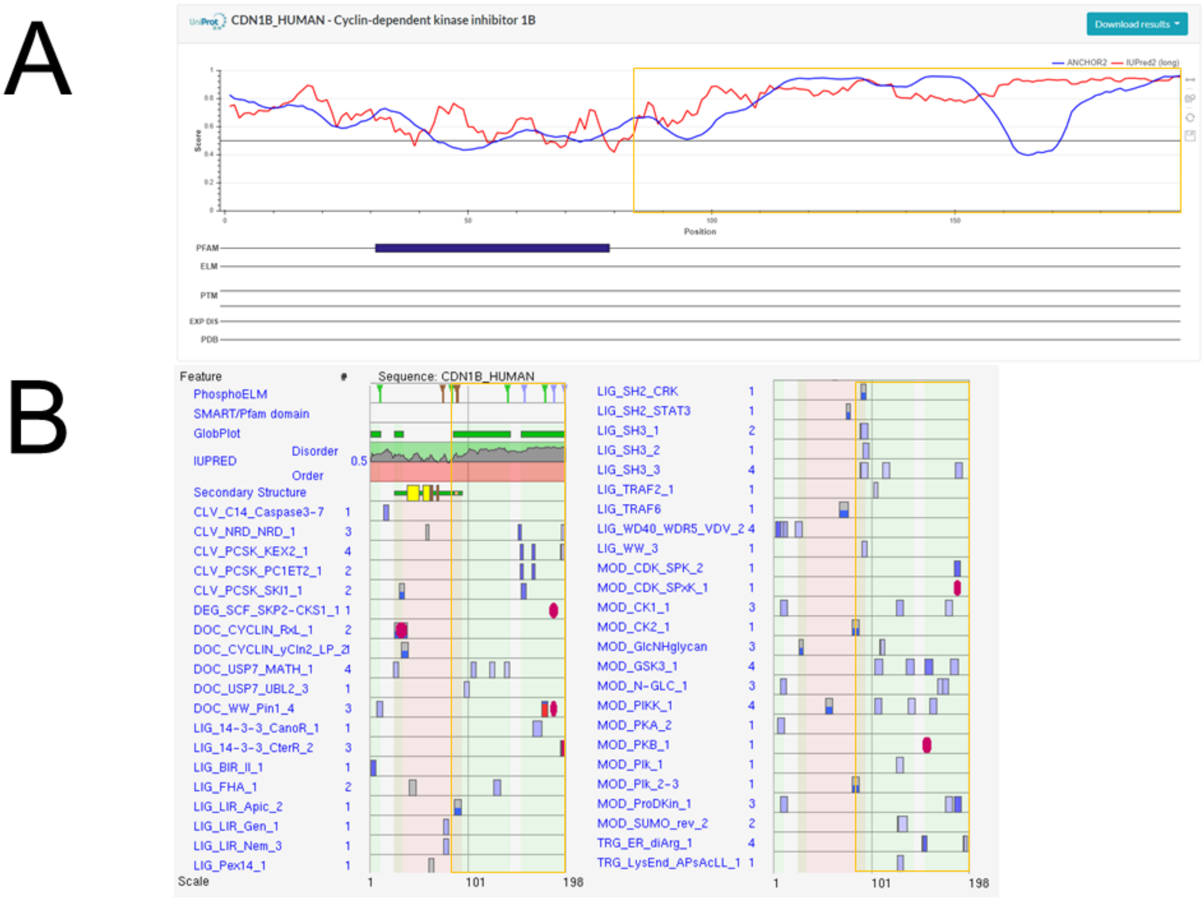
Yoon, M. K. *et al.* (2012) ‘Cell cycle regulation by the intrinsically disordered proteins p21 and p27’, *Biochemical Society Transactions*. NIH Public Access, pp. 981–988. doi: 10.1042/BST20120092.

Zarin, T. *et al.* (2017) ‘Selection maintains signaling function of a highly diverged intrinsically disordered region’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(8), pp. E1450–E1459. doi: 10.1073/pnas.1614787114.

Zarin, T. *et al.* (2019) ‘Proteome-wide signatures of function in highly diverged intrinsically disordered regions’, *eLife*, 8. doi: 10.7554/eLife.46883.

Zarin, T. *et al.* (2020) ‘Identifying molecular features that are associated with biological function of intrinsically disordered protein regions’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 2020.06.23.167361. doi: 10.1101/2020.06.23.167361.

Zarin, T. *et al.* (2021) ‘Identifying molecular features that are associated with biological function of intrinsically disordered protein regions’, *eLife*. eLife Sciences Publications Ltd, 10, pp. 1–36. doi: 10.7554/eLife.60220.



Supplementary Figure 1. Predictions for p27 for ANCHOR2 (A) and ELM (B). For both predictions, we inputted the full protein, so we highlight the C-terminal IDR in gold. A) The blue line shows the ANCHOR2 score predicting disordered binding regions. B) The blue boxes show matches to short linear motifs within the sequence, as labeled on the left. The darker the blue, the more conserved the motif is across orthologues. The red circles indicate known instances annotated from literature.