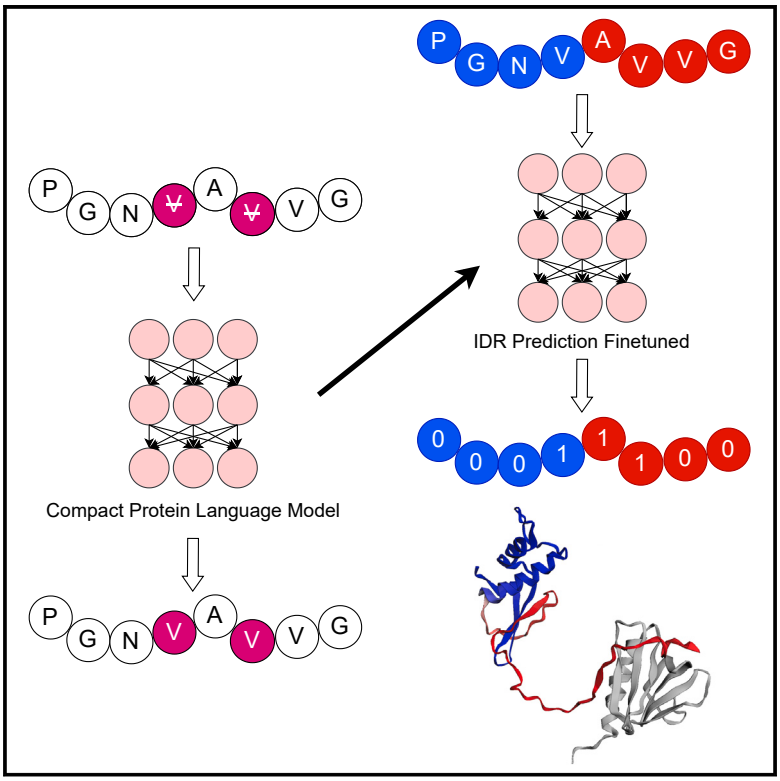
Structure

DR-BERT: A protein language model to annotate disordered regions

Graphical abstract



Authors

Ananthan Nambiar,
John Malcolm Forsyth, Simon Liu,
Sergei Maslov

Correspondence

nambiar4@illinois.edu (A.N.),
maslov@illinois.edu (S.M.)

In brief

Nambiar et al. present DR-BERT, a lightweight protein language model that outperforms many existing methods in predicting intrinsically disordered protein regions. Leveraging contextual information, DR-BERT's pretraining-based approach offers a computationally efficient and accurate means for IDR annotation, allowing easier access to computational annotation of IDR.

Highlights

- DR-BERT, a protein language model, makes accurate predictions of disordered regions
- Performance is due to pretraining and DR-BERT's ability to use contextual information
- DR-BERT does not require sequence alignments or biophysical properties as an input

Nambiar et al., 2024, Structure *32*, 1–9
August 8, 2024
https://doi.org/10.1016/j.str.2024.04.010

CellPress

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010

Structure

CellPress

Resource

DR-BERT: A protein language model to annotate disordered regions

Ananthan Nambiar,[1,2,7,*] John Malcolm Forsyth,[2,3,7] Simon Liu,[2,3,6] and Sergei Maslov[1,2,4,5,8,*]
[1]Department of Bioengineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[2]Carl R. Woese Institute for Genomic Biology, Urbana, IL 61801, USA
[3]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[4]Department of Physics, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[5]Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, IL 60439, USA
[6]Present address: Roblox, 970 Park Pl, San Mateo, CA 94403, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: nambiar4@illinois.edu (A.N.), maslov@illinois.edu (S.M.)
https://doi.org/10.1016/j.str.2024.04.010

SUMMARY

Despite their lack of a rigid structure, intrinsically disordered regions (IDRs) in proteins play important roles in cellular functions, including mediating protein-protein interactions. Therefore, it is important to computationally annotate IDRs with high accuracy. In this study, we present Disordered Region prediction using Bidirectional Encoder Representations from Transformers (DR-BERT), a compact protein language model. Unlike most popular tools, DR-BERT is pretrained on unannotated proteins and trained to predict IDRs without relying on explicit evolutionary or biophysical data. Despite this, DR-BERT demonstrates significant improvement over existing methods on the Critical Assessment of protein Intrinsic Disorder (CAID) evaluation dataset and outperforms competitors on two out of four test cases in the CAID 2 dataset, while maintaining competitiveness in the others. This performance is due to the information learned during pretraining and DR-BERT's ability to use contextual information.

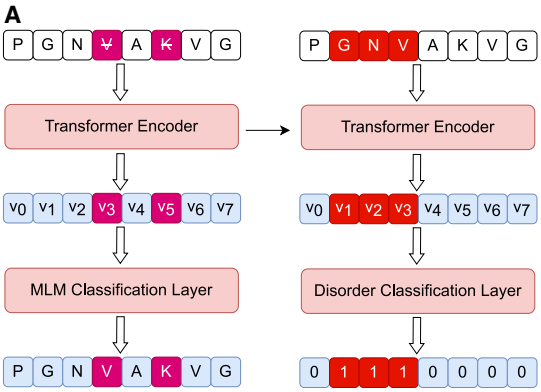
INTRODUCTION

Over a century ago, the chemist Emil Fischer postulated the lock-and-key model for enzymatic reactions, giving rise to the theory that a protein's function depends on its unique and rigid three-dimensional structure.[1] Within this paradigm, two proteins can interact if they have complementary structures. This idea has contributed to several advances in the understanding of protein function, and it is undeniable that the structure of a protein affects its function. However, studies in the late 1990s and early 2000s recognized that a stable structure is often not necessary for functional function.[2,3] Segments that lack a rigid structure, also known as intrinsically disordered regions (IDRs), have been found in many proteins and shown to actively participate in diverse functions.[4] In fact, these disordered regions are critical for some proteins with central roles in cellular signaling and regulatory networks, allowing them to interact with different proteins.[3,5]

Given the functional importance of disordered regions, computational methods for predicting disordered regions have been studied for decades, and over a hundred methods, ranging from biophysical to machine-learning-based models, have been developed.[6] Recently, predictors that use deep learning have gained traction.[6] This was particularly evident in the Critical Assessment of protein Intrinsic Disorder (CAID) competitions, where deep-learning-based models consistently delivered the best performance.[7] Many existing deep learning methods to predict disordered regions utilize recurrent neural networks and convolutional neural networks, sometimes paired with an attention mechanism.[8–10] This success of deep-learning-based methods to predict disordered regions in proteins can be attributed to both the complex and non-linear nature of sequence-structure maps as well as the steady increase of data availability.[11]

Protein language modeling has been a particularly fast-growing area of deep learning research for computational biology. Inspired by natural language processing, the core idea of protein language modeling is that the amino acids (or sometimes small groups of amino acids) that make up a protein are analogous to the words that make up a sentence.[12,13] Like their natural language counterparts, protein language models leverage the large number of unannotated amino acid sequence data to pretrain deep learning models before specializing them on much smaller amounts of annotated data. Usually, this pretraining step consists of training the model either to predict the context surrounding a particular residue[14] or to predict the identity of a hidden residue, given its surrounding context, i.e., the set of its nearby amino acid residues.[13] These models have then been successfully used to perform various downstream tasks including protein family labeling,[12,14] prediction of protein

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010



B

	# sequences	% disordered residues
Uniref90	6,564,742	n/a
DisProt	2,386	27

C

Model	AU-ROC	F1	MCC
DR-BERT	**0.82/83**	**0.55/55**	**0.43/0.43**
flDPnn	0.81/0.82	0.44/0.51	0.35/0.39
RawMSA	0.79/0.80	0.49/0.51	0.35/0.38
SPOT Disorder 2	0.78/0.80	0.5/0.53	0.37/0.42

Figure 1. The DR-BERT model is pretrained on the masked language modeling task and fine-tuned on predicting disordered regions in proteins

(A) A schematic of the DR-BERT model and the pretraining and fine-tuning procedures.

(B) The statistics of data used in this study and (C) the CAID 1 and CAID 2 results of DR-BERT compared to some of the best-performing models from the CAID competitions.[7] Cells are colored based on the performance of each model for a particular metric for CAID 1.

interactions[12] and subcellular localization,[15] and the inference of evolutionary trajectories and phylogenetic relationships of proteins.[16,17] While most protein language models tend to be large and graphics processing unit (GPU) intensive, there have been studies proposing small and computationally inexpensive protein language models.[12]

In this paper, we present Disordered Region prediction using Bidirectional Encoder Representations from Transformers (DR-BERT), a small protein language model that is first pretrained on a large corpus of amino acid sequences and then fine-tuned to predict disordered regions in proteins. We validate our model on both CAID 1 and CAID 2 evaluation data and benchmark it against some of the best-performing models. We then investigate the impact of pretraining on the performance of DR-BERT. Finally, we dive into one particular biological case study involving RPB6, a subunit of RNA polymerase, to illustrate how DR-BERT arrives at its predictions and learns to use contextual information from the amino acid sequence.

RESULTS

While many models for disordered region prediction depend on knowledge of biophysical properties of amino acids used as inputs, previous work has shown that pretraining a protein language model may allow it to learn these biophysical and functional properties in a self-supervised manner.[12,13] Therefore, we chose to build our DR-BERT model using only the amino acid sequence of a protein as the input. This model is first pretrained on the masked language modeling task as shown in Figure 1 before it is fine-tuned to predict IDRs.

The model itself is a neural network with a transformer encoder block composed of six stacked transformer encoder layers (see STAR Methods for details). The purpose of the encoder block is to create contextual latent representations of each residue. That is, each residue is represented by a vector that captures the context of the rest of the sequence. By stacking multiple transformer encoder layers within the encoder block, the final latent representations can capture more complex higher-level information and relationships from the amino acid sequence. These vectors are then passed to a final linear layer that constructs a task-specific output.

In the pretraining task of masked language modeling, the neural network is asked to predict the identities of amino acids that have been masked in the input. In this study, we pretrained our model on 6,564,742 proteins randomly sampled from the UniRef90 dataset.[18] Next, we fine-tuned DR-BERT by tasking it to classify residues in proteins as disordered or ordered using annotated data from the DisProt database. The performance of DR-BERT on this fine-tuning task is shown in Figure 1C alongside previous state-of-the-art methods.

Benchmarking DR-BERT's performance

When fine-tuning DR-BERT on the disordered region classification task, we split the DisProt data into train/validation/test sets with the aim of enabling a systematic and unbiased comparison against existing methods. In particular, proteins from the CAID competitions were reserved as test data and were not available to the model during training. In addition, any proteins that shared more than 25% similarity to proteins from the test set was excluded from the train set. We ran out our benchmarking on both CAID 1 and CAID 2. This left us with 1,408 examples in the train set, 156 sequences in validation, and 652 in the test set for CAID 1 and 2,013 examples in the train set, 216 sequences in validation, and 348 in the test set for CAID 2. By doing so, we were able to reproduce the results of some of the top-performing models from CAID.[7] In particular, we benchmarked DR-BERT against flDPnn,[19] RawMSA,[20] SPOT-Disorder2,[8] DisoMine,[21] Espritz-D,[22] AUCpreD,[23] IUPred2A/3,[24] and Predisorder.[25]

Of these methods, all but IUPred2A/3 are deep-learning-based models based on feedforward, recurrent, and convolutional neural network architectures. flDPnn is a feedforward neural network that uses evolutionary and structural information in addition to disordered region predictions from simpler models; RawMSA uses convolutional neural networks (CNNs) on evolutionary information (in the form of MSAs); SPOT-Disorder2 uses a combination of CNNs and recurrent neural networks (RNNs) on input with evolutionary information; DisoMine uses RNNs on structural information; Espritz uses RNNs on evolutionary information; AUCpreD uses CNNs on sequence information (with optional evolutionary information); and Predisorder uses RNNs with structural, biophysical, and evolutionary information. A notable pattern here is that most of these methods use pre-computed features. First, the performance of the model is reliant on its upstream dependencies. For example, if a model uses MSAs as input, one would expect its performance to

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010

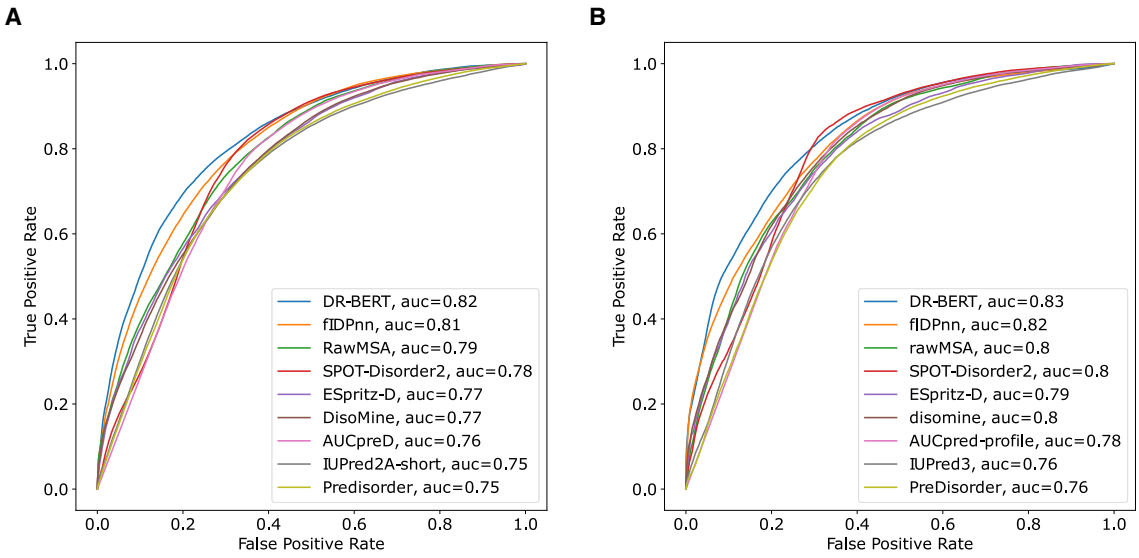


Figure 2. Receiver operating characteristic (ROC) curves on the CAID datasets

The ROC curves of DR-BERT and other models on test sets from (A) CAID 1 and (B) CAID 2. The legends display the area under the curve (AUC) for each model. The models are ordered based on the AUC in CAID 1.

deteriorate for proteins that do not have many known homologs. In addition, the presence of multiple third-party techniques in a prediction pipeline makes it more difficult to optimize computational efficiency. In contrast, DR-BERT is a fully self-contained model that does not rely on any additional information besides the amino acid sequence of a protein.

Despite not requiring any additional information, the receiver operating characteristic (ROC) curves (Figure 2) on both the CAID 1 and CAID 2 test sets demonstrate that DR-BERT outperforms all of the other methods in predicting disordered regions. For CAID 1, DR-BERT ranks first in terms of area under the ROC curves (AU-ROC) with a value of 0.82. The scores then incrementally decrease with flDPnn, RawMSA, and SPOT-Disorder2. For CAID 2, DR-BERT is again the highest ranking method followed by flDPnn and a three-way tie between rawMSA, SPOT-Disorder2, and DisoMine. The ROC curves also show that DR-BERT offers particularly evident improvements in the lower range of false positive rates. However, as the disordered region dataset is imbalanced with more ordered residues than disordered ones, the ROC curves may show an overly optimistic view of the classifiers.[26] Therefore, we also calculate F1 and Matthews correlation coefficients (MCCs) for each model. These scores, along with the AU-ROC scores are shown in Figure 3A for CAID 1 and Figure 3B for CAID 2. Again, DR-BERT scores the highest on both metrics with an F1 of 0.55 and MCC of 0.43 for CAID 1 and an F1 of 0.56 and MCC of 0.43 for CAID 2. The precision-recall plots in Figure S1 also show that DR-BERT performs better than the other methods in balancing the trade-off between precision and recall.

To determine the statistical significance of DR-BERT's improvement over the existing methods, we performed a resampling analysis similar to that of Hu et al.[19] and Necci et al.[7] Specifically, we resampled 25% of the test set 20 times. For each resample, we calculated the AU-ROC, F1, and MCC scores for DR-BERT and the other methods. Finally, we performed Wilcoxon tests comparing the scores obtained by DR-BERT to those of each of the other methods, with the alternative hypothesis being that DR-BERT's score is greater than the other method. The p values from these hypothesis tests, shown in Figure S2, confirm that DR-BERT performs significantly better than existing methods across the different scoring metrics for both CAID 1 and CAID 2.

This supremacy of DR-BERT over methods that use evolutionary and structural features suggests that these features can be successfully learned by the model either during pretraining or fine-tuning. In fact, it has been previously shown that pretrained protein language models are able to extract structural information from amino acid sequences.[27,28] However, these results alone do not elucidate the contribution of pretraining to the success of DR-BERT.

Pretraining and fine-tuning

To better understand the role that pretraining plays in extracting the information relevant to disordered region prediction, we interrogated DR-BERT models at two stages: (a) after only pretraining and (b) after pretraining and fine-tuning. At both of these stages, we extracted the embeddings from the encoder block for each residue in the test set. Using t-SNE, we projected these embeddings down to two dimensions.[29] Then, we calculated kernel density estimates (KDEs) separately for ordered and disordered residues. These KDEs are shown in Figure 4A for the pretrained model and in Figure 4B for the model that was pretrained and fine-tuned. The plot for the pretrained model shows about 20 different clusters of ordered residues and 15 distinct clusters of disordered residues. Upon further investigation, we see that each cluster corresponds to an individual amino acid. There are a few exceptions to this for disordered residues. For instance, there is no clear disordered cluster for the amino acid tryptophan (W). This is because tryptophan is one of the most

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010

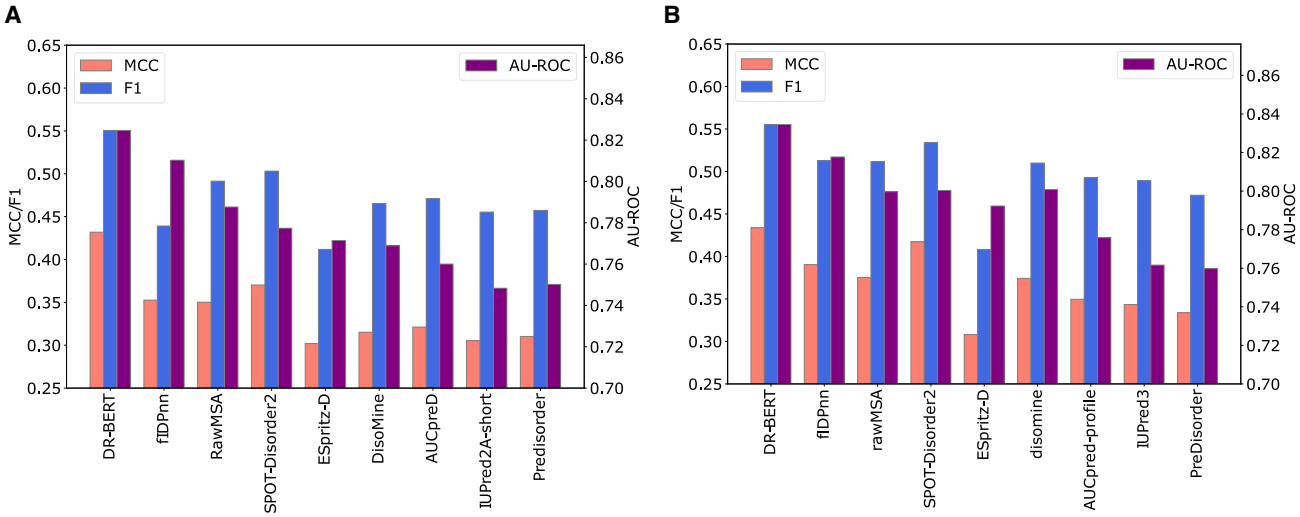


Figure 3. Comparing the results of DR-BERT with other top-performing methods on the CAID datasets

(A) The MCC, F1, and AU-ROC scores of DR-BERT and the top-performing methods from CAID 1, evaluated on the test split.

(B) The MCC, F1, and AU-ROC scores for corresponding methods evaluated on the CAID 2 test data.

order-promoting amino acids and is rarely encountered inside IDRs.[30] However, the clear overall pattern in Figure 4A is that most ordered clusters are accompanied by an adjacent disordered cluster for the same amino acid. This is in contrast to the null model where the disorder/order residue labels are shuffled, shown in Figure S4. On the other hand, the plot for the fine-tuned embeddings depicts a different story. While the embeddings are not clustered by amino acid, the disordered residues are all clustered together and are well separated from the ordered residues. The difference between the pretrained and fine-tuned embeddings highlights that pretraining a protein language model is sufficient to extract some information regarding disordered regions in proteins. Fine-tuning the model allows it to then home in on the differences between disordered and ordered residues to more efficiently separate them. This result gives credence to an observation we made in the study by Nambiar et al.[12] where we noted that pretraining a protein language model allows it to learn general but biologically relevant information from amino acid sequences, whereas fine-tuning gives the model more information about one characteristic but at the expense of generality.

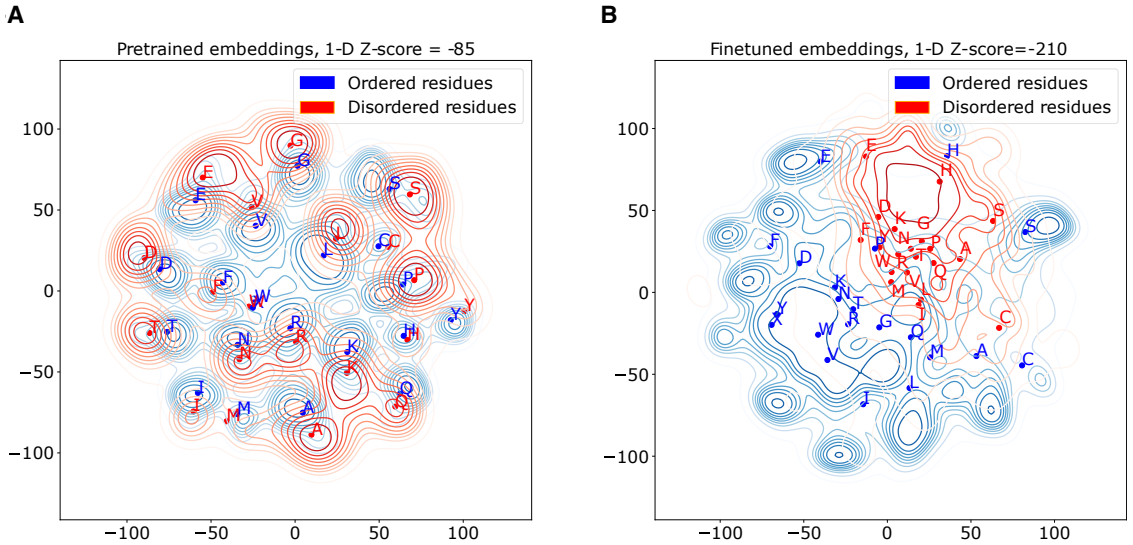


Figure 4. Plotting the embeddings of ordered and disordered residues

(A) A t-SNE projection of the pretrained embeddings of residues in the CAID 1 test set. The plot shows the kernel density estimates of ordered residues in blue and disordered residues in red. The labeled points indicate the mean position of each amino acid. This plot should be compared to the null model shown in Figure S4.

(B) A similar plot but with embeddings from a model fine-tuned to predict disordered regions. For both plots, the two-sample Z-test is performed after reducing the dimensionality of the embedding to 1D.

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010

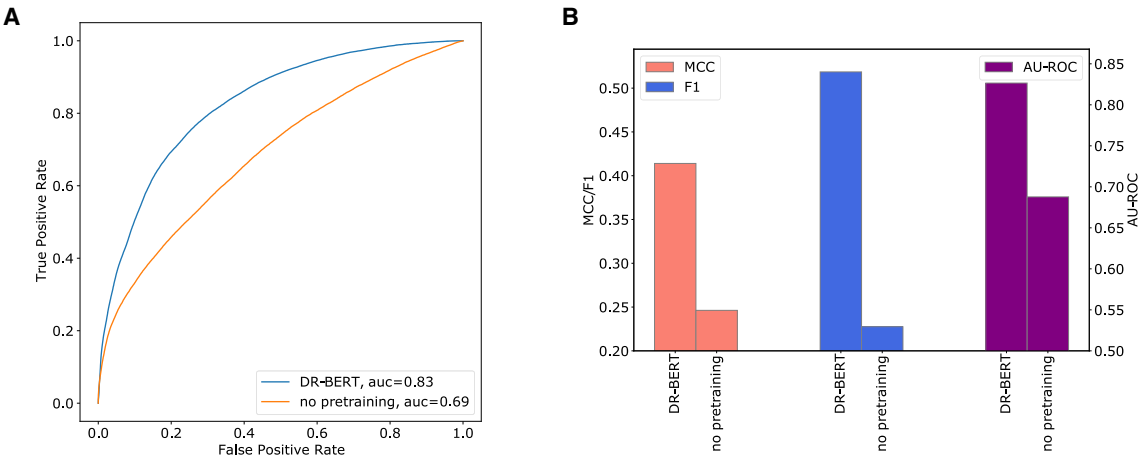


Figure 5. Comparison of the results of DR-BERT with its version without pretraining

(A) The ROC plots of DR-BERT and the non-pretrained model, evaluated on the CAID 1 test set. The AUC is presented in the legend.

(B) The MCC, F1, and AU-ROC scores of DR-BERT and the version without pretraining.

In addition, we wanted to quantify the advantage of pretraining for predicting disordered regions. To do this, we trained a model with an identical architecture to DR-BERT to predict disordered regions without any pretraining. The results of this non-pretrained model evaluated on CAID 1, shown in Figure 5, show that pretraining DR-BERT gives it a considerable advantage. In fact, in the absence of pretraining, our model lags behind the models from the CAID competition, shown in Figure 3 and S3. This showcases the advantage of pretraining for transformer neural networks, especially for low data regimes.

A case study: The disordered region in RPB6 protein

Evaluating DR-BERT on a large annotated dataset gave us confidence in DR-BERT's ability to make accurate predictions regarding disordered regions. However, it is also useful to illustrate a potential use-case by focusing on the predictions of the model for an IDR within a single protein. Doing so gives us the opportunity to gain insight into how the context of a particular sequence is used by the attention heads of the model (see STAR Methods) to make predictions for different residues in the same protein. We decided to illustrate this using RPB6 protein as an example. RPB6 is a subunit of an RNA polymerase in fission yeast. It is known to bind to the general transcription factor, TFIIS.[31] This example allows us to test our disordered region prediction for a protein that is known to perform an important function. Figure 6B shows that DR-BERT predicts with high confidence that the N-terminal tail of RPB6 is in fact disordered. Indeed, NMR spectroscopy shows that not only does RPB6 have a flexible N-terminal tail, but this tail is also used to bind to the p62 subunit of the TFIIH transcription factor.[32] Figure 6A shows DR-BERT's predictions overlaid on the NMR-determined structure of the complex between RPB6 and the TFIIH p62 PH domain (PDB: 7DTI). To analyze how DR-BERT uses sequence context to make its predictions, we extracted the self-attention heads for each of the six layers in DR-BERT's encoder as it processed the 130 amino acid long RPB6 sequence. Each attention map is represented as a 130× 130 matrix M where M_{ij} gives a numerical score as to how much that particular attention head focuses on amino acid j when determining the context relevant for amino acid i. A sample of these attention maps for each layer in DR-BERT is displayed in Figure 6C (a complete table is shown on Figure S5). We observed that the features learned by the attention maps of the initial layer do not have any clear high-level patterns. However, attention maps from layers two to four display some distinct patterns. For example, in layer 4, the attention map reveals that the relevant context for each residue includes a large window of surrounding residues in addition to several smaller windows at intervals on either side of the residue in question. By layer 5, at least one attention map divides the residues into two distinct groups: one group consists of residues 1 to 50, and the other group comprises residues 50 to 130. Each group only considers residues within its respective group as relevant context, while ignoring residues in the other group. Comparing this attention map to the disorder scores by sequence position, we can see that the division between the two groups occurs at the transition between the ordered and disordered regions of the protein. This observation is similar to demonstrations in computer vision, where deep neural networks learn features hierarchically, with the initial layers detecting simpler, disjoint features, and layers toward the end of the neural network detecting high-level features directly related to the model's training task.

DISCUSSION

In this study, we introduce DR-BERT, a protein language model for predicting disordered regions in proteins. DR-BERT is first pretrained on the masked language modeling task before it is fine-tuned to predict disordered regions.

This fine-tuned model is benchmarked using the CAID evaluation data and significantly surpasses the other models. This improvement over models that use biophysical and structural information supports the hypothesis that pretraining protein language models enables them to learn biologically relevant information in a self-supervised manner without any provided

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010

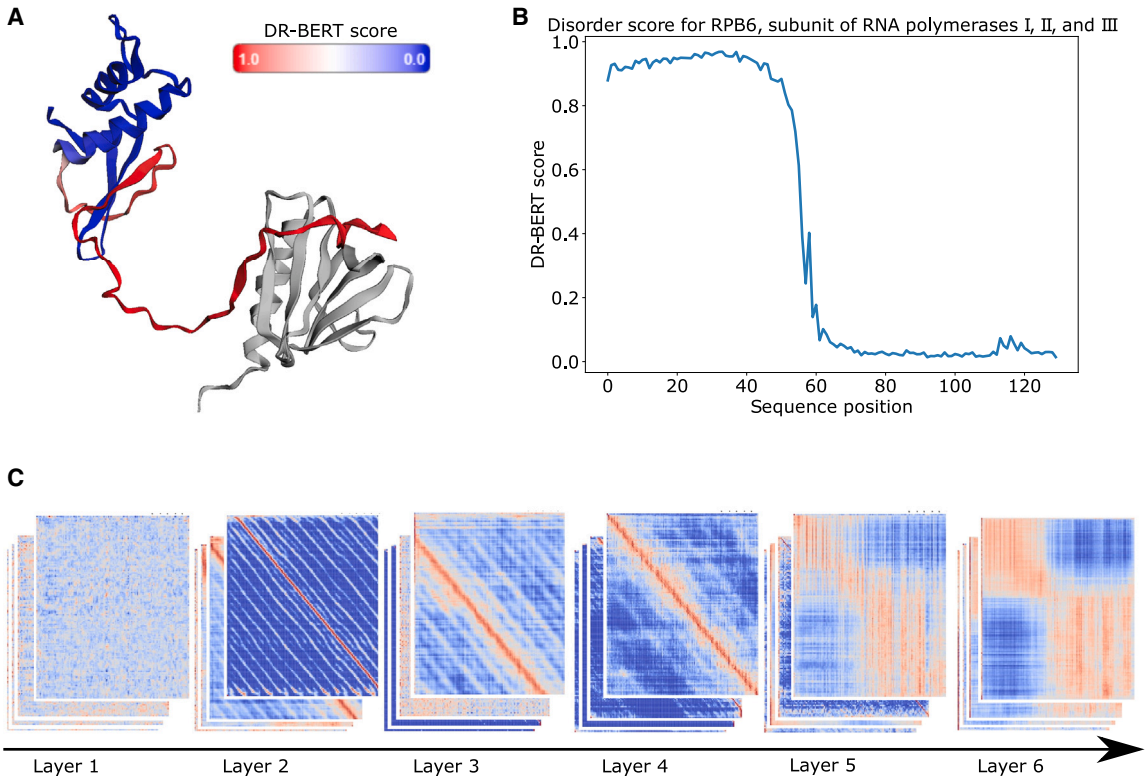


Figure 6. Application of DR-BERT to RPB6, a subunit of RNA polymerase

(A) The three-dimensional structure of RPB6 as it binds to the TFIIH p62 PH domain (PDB:7DTI). The protein is colored by the DR-BERT score, which represents the probability that a given residue is disordered.

(B) A plot of the DR-BERT scores for RPB6 shown for each position along the amino acid sequence.

(C) A sample of DR-BERT's self-attention maps for each of the 6 layers in the model as it processes the RPB6 sequence.

annotations. This learned information should include both evolutionary and biophysical information.

This hypothesis is further validated as we show that the embeddings of the pretrained model are able to differentiate between disordered and ordered residues without access to any annotations during training. Furthermore, we showed that a model with an identical architecture as DR-BERT suffers a large loss in performance when the pretraining step is skipped.

Finally, we took a closer look at how DR-BERT makes predictions for RPB6. Through this exercise, we saw that DR-BERT extracts patterns hierarchically, with higher-level features extracted by attention heads in deeper layers of the neural network. This is similar to the behavior that had been observed in computer vision.

To verify that DR-BERT was not overfitting on the training data, we excluded from the training set proteins that were clustered with proteins in the test set with 25% sequence similarity. The clustering in this process was performed using CD-HIT.[33]

Given the high performance of DR-BERT on the disordered region prediction task, we also investigated its ability to perform related tasks from CAID 2. This included evaluating DR-BERT on a disordered region dataset where X-ray annotations were removed (disorder-noX) and a dataset where PDB residues were incorporated (disorder-PDB). The results shown in Figure 7 show that DR-BERT performs well on the disorder-noX test set, placing first on the AU-ROC and area under precision-recall plot metrics and second to SPOT-Disorder2 on F1 score and MCC. However, DR-BERT shows a weaker performance on the disorder-PDB set, being outperformed by tying for third place with RawMSA on AU-ROC and area under precision recall plot and placing fourth on F1 score and MCC. Given that DR-BERT was trained on the vanilla disordered region dataset from DisProt, it is not surprising that DR-BERT's performance dropped on some of these variants. In addition to variants of disordered region annotations, we also evaluated DR-BERT on predicting protein-binding regions. Protein-binding disordered regions are regions in disordered proteins that bind to structured partners and potentially allow the disordered protein to bind to multiple partners.[34] DR-BERT achieved an AU-ROC of 0.75, F1 score of 0.45, and MCC of 0.32 beating other protein-binding predictors from CAID 2.

To further validate our results, we also ran an additional evaluation where we modified the CAID 1 and CAID 2 disorder sets so that no protein should have more than 25% identity to any other protein (even within the set). These results, shown in Figure S7, show no significant changes to performance, confirming the reliability of our results.

The success of DR-BERT, in addition to the insight into how DR-BERT makes predictions, leads us to believe that protein language models could play an important role in the next

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010

CellPress

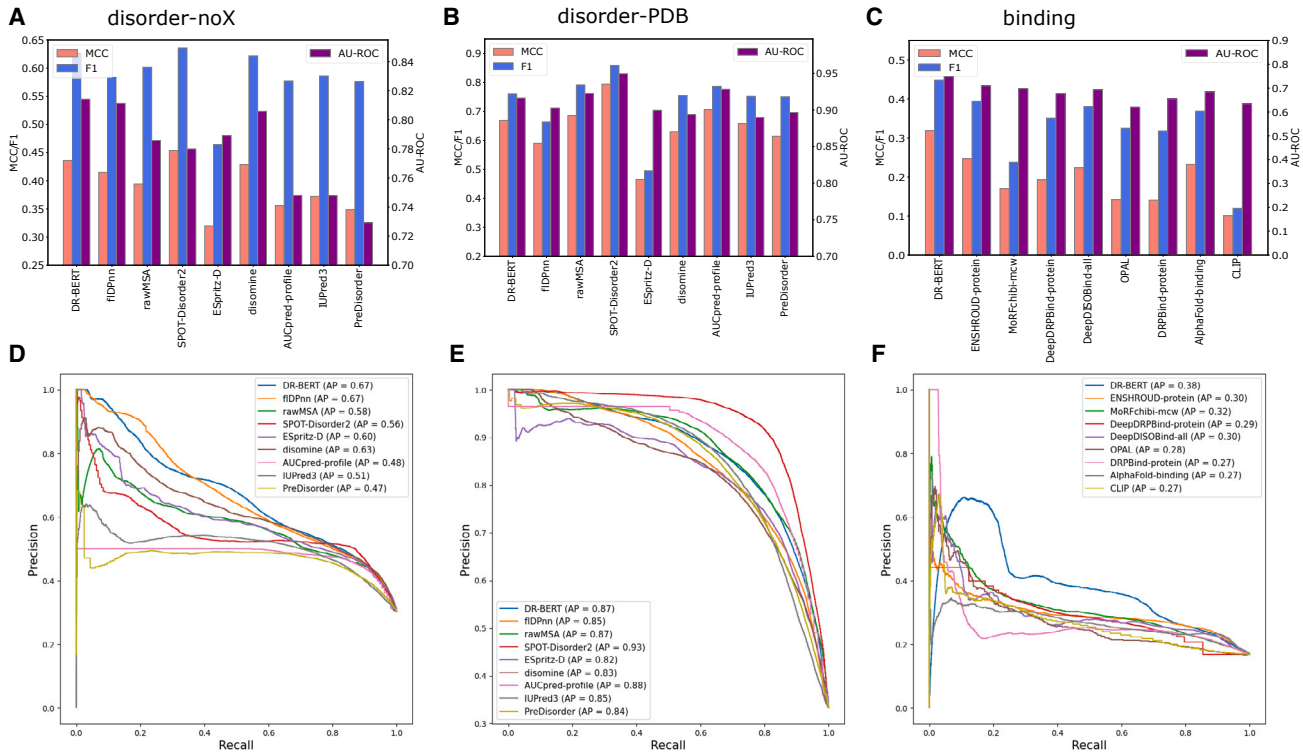


Figure 7. The results on the additional test sets from CAID 2

(A–C) Show the AU-ROC, F1 score, and MCC on the disorder-noX, disorder-PDB, and protein binding test sets.

(D–F) Show precision-recall plots that correspond to the same tasks.

generation of neural networks for predicting disordered regions. In fact, after completing our study, we found that a similar model to DR-BERT was presented in a recent preprint by Redl et al.[35] However, there are significant differences in our studies, including our investigation of the effect of pretraining on the success of the protein language model and the insight into the features extracted by the attentional layers. In addition, DR-BERT is significantly smaller than the model proposed by Redl et al.[35] (with 15x fewer parameters), which may make DR-BERT more accessible to users without access to high-performance GPUs. An alternative approach to the one shown in our study would be to extract embeddings from a pretrained model and pass them to a downstream classifier without fine-tuning the embeddings. This approach, which is used by the SETH model, makes it more efficient to train models on downstream tasks using embeddings from a large pretrained language model.[36] However, as shown in Figure S6, DR-BERT is able to surpass the performance of a larger language model where the embeddings are not fine-tuned. Moreover, each prediction would still involve a forward pass on the large model, which would be slower than using a small protein language model like DR-BERT.

In order to maximize the accessibility of our model, we have made a web-app (accessible at https://huggingface.co/spaces/nambiar4/DR-BERT) where anyone can use DR-BERT to make disordered region predictions.[37] Due to the small size of our model, we are currently able to run our server using only 2 CPU cores and 16 GB of RAM. In addition to the web-app, users who want more control over their predictions can run the pretraining, fine-tuning, and prediction scripts we have made available at https://github.com/maslov-group/DR-BERT.

While we have shown that a protein language model with no additional information is sufficient to make accurate predictions of disordered regions in proteins, a direction worth exploring in the future is whether combining the information learned by protein language models with biophysical properties and the outputs of other models might further improve performance. One additional input that could be particularly interesting is the per-residue confidence score provided by AlphaFold when making sequence-to-structure predictions since it has been shown that disordered regions are often assigned a low confidence score by AlphaFold.[38]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data processing
  - Model architecture
  - Pretraining
  - Disordered region prediction

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.str.2024.04.010.

ACKNOWLEDGMENTS

This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign. This work was partially supported by NSF grant #2107344. Part of this work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357. J.M.F. and S.L. have been supported by the James Scholar Honors Program and the Illinois Scholars Undergraduate Research Program. We thank Mark Hopkins, Anna Ritz, Ashley Blystone, and Desiree Odgers for insightful discussions.

AUTHOR CONTRIBUTIONS

All authors designed the study. S.M. supervised the study and A.N., J.M.F., and S.L. performed simulations and calculations. All authors discussed and wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 7, 2023
Revised: June 16, 2023
Accepted: April 8, 2024
Published: May 2, 2024

REFERENCES

1. Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. Ber. Dtsch. Chem. Ges. *27*, 2985–2993.
2. Uversky, V.N. (2002). Natively unfolded proteins: A point where biology waits for physics. Protein Sci. *11*, 739–756.
3. Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. J. Mol. Biol. *293*, 321–331.
4. Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., et al. (2014). Classification of intrinsically disordered regions and proteins. Chem. Rev. *114*, 6589–6631.
5. Wright, P.E., and Dyson, H.J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. Nat. Rev. Mol. Cell Biol. *16*, 18–29.
6. Zhao, B., and Kurgan, L. (2022). Deep learning in prediction of intrinsic disorder in proteins. Comput. Struct. Biotechnol. J. *20*, 1286–1294.
7. Necci, M., Piovesan, D.; CAID Predictors; DisProt Curators, and Tosatto, S.C.E. (2021). Critical assessment of protein intrinsic disorder prediction. Nat. Methods *18*, 472–481.
8. Hanson, J., Paliwal, K.K., Litfin, T., and Zhou, Y. (2019). SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. Dev. Reprod. Biol. *17*, 645–656.
9. Tang, Y.-J., Pang, Y.-H., and Liu, B. (2021). IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. Bioinformatics *36*, 5177–5186.
10. Tang, Y.-J., Pang, Y.-H., and Liu, B. (2022). DeepIDP-2L: protein intrinsically disordered region prediction by combining convolutional attention network and hierarchical attention network. Bioinformatics *38*, 1252–1260.
11. Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C.J., Aspromonte, M.C., Davey, N.E., Davidović, R., Dosztányi, Z., et al. (2017). DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res. *45*, D219–D227.
12. Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., and Ritz, A. (2020). Transforming the language of life: transformer neural networks for protein prediction tasks. In Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics (ACM Digital Library), pp. 1–8.
13. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. USA *118*, e2016239118.
14. Asgari, E., and Mofrad, M.R. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One *10*, e0141287.
15. Stärk, H., Dallago, C., Heinzinger, M., and Rost, B. (2021). Light attention predicts protein location from the language of life. Bioinform. Adv. *1*, vbab035.
16. Hie, B.L., Yang, K.K., and Kim, P.S. (2022). Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. Cell Syst. *13*, 274–285.e6.
17. Lupo, U., Sgarbossa, D., and Bitbol, A.-F. (2022). Protein language models trained on multiple sequence alignments learn phylogenetic relationships. Nat. Commun. *13*, 6298.
18. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H.; UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932.
19. Hu, G., Katuwawala, A., Wang, K., Wu, Z., Ghadermarzi, S., Gao, J., and Kurgan, L. (2021). flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat. Commun. *12*, 4438.
20. Mirabello, C., and Wallner, B. (2019). rawMSA: End-to-end deep learning using raw multiple sequence alignments. PLoS One *14*, e0220182.
21. Orlando, G., Raimondi, D., Codicè, F., Tabaro, F., and Vranken, W. (2022). Prediction of disordered regions in proteins with recurrent neural networks and protein dynamics. J. Mol. Biol. *434*, 167579.
22. Walsh, I., Martin, A.J.M., Di Domenico, T., and Tosatto, S.C.E. (2012). ESpritz: accurate and fast prediction of protein disorder. Bioinformatics *28*, 503–509.
23. Wang, S., Ma, J., and Xu, J. (2016). AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. Bioinformatics *32*, i672–i679.
24. 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 Res. *46*, W329–W337.
25. Deng, X., Eickholt, J., and Cheng, J. (2009). PreDisorder: ab initio sequence-based prediction of protein disordered regions. BMC Bioinf. *10*, 436.
26. Davis, J., and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (ACM Digital Library), pp. 233–240.
27. Bhattacharya, N., Thomas, N., Rao, R., Dauparas, J., Koo, P.K., Baker, D., Song, Y.S., and Ovchinnikov, S. (2020). Single layers of attention suffice to predict protein contacts. Preprint at bioRxiv. https://doi.org/10.1101/2020.12.21.423882.
28. Singh, J., Paliwal, K., Litfin, T., Singh, J., and Zhou, Y. (2022). Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. Sci. Rep. *12*, 7607.
29. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research *9*.
30. Campen, A., Williams, R.M., Brown, C.J., Meng, J., Uversky, V.N., and Dunker, A.K. (2008). TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept. Lett. *15*, 956–963.

31. Ishiguro, A., Nogi, Y., Hisatake, K., Muramatsu, M., and Ishihama, A. (2000). The Rpb6 subunit of fission yeast RNA polymerase II is a contact target of the transcription elongation factor TFIIS. Mol. Cell Biol. *20*, 1263–1270.

32. Okuda, M., Suwa, T., Suzuki, H., Yamaguchi, Y., and Nishimura, Y. (2021). Three human RNA polymerases interact with TFIIH via a common RPB6 subunit. Nucleic Acids Res. *50*, 1–16.

33. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

34. Mészáros, B., Simon, I., and Dosztányi, Z. (2009). Prediction of Protein Binding Regions in Disordered Proteins. PLoS Comput. Biol. *5*, e1000376.

35. Redl, I., Fisicaro, C., Dutton, O., Hoffmann, F., Henderson, L., Owens, B.M., Heberling, M., Paci, E., and Tamiola, K. (2023). ADOPT: intrinsic protein disorder prediction through deep bidirectional transformers. NAR Genom Bioinform *5*, lqad041.

36. Ilzhöfer, D., Heinzinger, M., and Rost, B. (2022). SETH predicts nuances of residue disorder from protein embeddings. Front. Bioinform. *2*, 1019597.

37. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. arxiv. https://doi.org/10.48550/arXiv.1910.03771.

38. Alderson, T.R., Pritišanac, I., Moses, A.M., and Forman-Kay, J.D. (2022). Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. Proc Natl Acad Sci USA *120*, e2304302120.

39. Quaglia, F., Mészáros, B., Salladini, E., Hatos, A., Pancsa, R., Chemes, L.B., Pajkos, M., Lazar, T., Peña-Díaz, S., Santos, J., et al. (2022). DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. Nucleic Acids Res. *50*, D480–D487.

40. Van Rossum, G., and Drake, F.L., Jr. (1995). Python reference manual (Centrum voor Wiskunde en Informatica Amsterdam).

41. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. Adv. Neural Inf. Process. Syst. *32*, 8024–8035. Curran Associates, Inc.

42. McKinney, W., et al. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference, Austin, TX, *445*, pp. 51–56.

43. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. In Advances in Neural Information Processing Systems, *34*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J.W. Vaughan, eds. *(Curran Associates, Inc), pp. 29287–29303.*

44. Del Conte, A., Bouhraoua, A., Mehdiabadi, M., Clementel, D., Monzon, A.M., CAID predictors, Tosatto, S.C.E., and Piovesan, D. (2023). CAID prediction portal: a comprehensive service for predicting intrinsic disorder and binding regions in proteins. Nucleic Acids Res. *51*, W62–W69.

45. Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR.

46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.U., and Polosukhin, I. (2017). Attention is All you Need. In Advances in Neural Information Processing Systems, *30*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. *(Curran Associates, Inc).*

47. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

48. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR. abs/1907.11692.

49. Kindratenko, V., Mu, D., Zhan, Y., Maloney, J., Hashemi, S.H., Rabe, B., Xu, K., Campbell, R., Peng, J., and Gropp, W. (2020). HAL: Computer System for Scalable Deep Learning (Association for Computing Machinery), pp. 41–48.

50. Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. *21*, 6–13.

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
UniRef90 Protein Dataset	UniProt Knowledgebase Suzek et al.[18]	https://www.uniprot.org/help/uniref
DisProt Disordered Region Dataset	DisProt Quaglia et al.[39]	https://disprot.org/download
CAID 1 and CAID 2 Test Set	Critical Assessment of protein Intrinsic Disorder prediction Experiment Necci et al.[7]	https://caid.idpcentral.org/challenge
Software and algorithms		
Python 3.9	Python Software Foundation Van Rossum and Drake[40]	https://www.python.org
Pytorch library	PyTorch Foundation Paszke et al.[41]	https://pytorch.org/
Transformers library	Hugging Face, Inc Wolf et al.[37]	https://huggingface.co/docs/transformers/index
Pandas library	NumFOCUS McKinney et al.[42]	https://pandas.pydata.org/
DR-BERT	This paper	https://github.com/maslov-group/DR-BERT

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Sergei Maslov (maslov@illinois.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All original code has been deposited at https://github.com/maslov-group/DR-BERT and is publicly available as of the date of publication.
- All data are generated from the dataset provided in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data processing

From UniRef90, we sampled 6,564,742 proteins at random for the training dataset and 250,000 proteins for the validation dataset. Per the construction of UniRef data, the validation set is designed to contain no examples with above 90% sequence similarity to any of the examples in the training set.[18] We used UniRef90 because we hope that down the road, we can use DR-BERT to predict the effects of mutations on disordered regions. It has been observed in [43] that the redundancy provided by UniRef90 allows for better variant effect prediction. Furthermore, as we show on Figure S8, a model pretrained on UniRef50 achieves the same performance as the UniRef90 model used in Figure 3. So, it does not appear that the higher redundancy in UniRef90 impedes the model.

For the finetuning task, 2,419 sequences were taken from DisProt Version 9.2, June 2022, with replicate proteins removed to create the 2,386 example dataset, and their disordered regions were recorded into boolean arrays as the ground truth labels.[39] To construct the train/validation/test splits for CAID 1, the Disprot and CAID 1 datasets were combined and clustered for 25% similarity using the CD-HIT algorithm. Then, only clusters without any CAID 1 proteins were used for training and validation. This resulted in 1,569 proteins in the train set and 156 sequences in the validation set. The 652 proteins from the CAID 1 competition dataset were used as the

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010

test set. To construct the train/validation/test splits for the CAID 2 experiment, we similarly combined the CAID 2, CAID 1 and Disprot datasets, and clustered for 25% similarity. Then, only clusters without any CAID 2 proteins were used for training and validation. The sizes for the train, validation, and test sets were 2013, 216, and 348 respectively.[44] The disorderd region data from CAID 2 comes from two types of experimential experimental techniques: circular dichroism spectroscopy and x-ray. In addition to the standard disorder dataset combining data from both techniques, we also tested on a test set where X-ray annotations were removed (disorder-noX). Furthermore, we also used a test set where in addition to using both techniques to annotate disordered regions, only regions where a structure was observed in the Protein Data Bank were labelled as ordered (disorder-PDB). Finally, we also included a test set for protein binding disordered regions.[44]

Model architecture

DR-BERT uses the Bidirectional Encoder Representations from Transformers (BERT) coupled with token classification heads trained on disordered region labels[45] and was trained using PyTorch and HuggingFace.[37,40–42] Based on the Robustly Optimized BERT pretraining Approach (RoBERTa), the model consists of an embedding layer connected with a Encoder Block with 6 encoder layers. The embedding layer consists of two main component layers: a word embedding layer and a positional embedding layer. The word embedding layer takes the tokenized sequence of amino acids and maps each token to a 768 dimensional vector. In contrast, the positional embedding layer captures the spatial information of the tokens to preserve the notion of context within the sequence.[46] After a dropout layer is applied to decrease the potential for overfitting,[47] the embedding, consisting of a 768 dimensional vector for each amino acid token, is used by the Transformer encoder layers. The RoBERTa transformer layer consists of a self-attention layer and a feed-forward network layer. The self-attention mechanism described in [46], captures the relationship between different tokens in a sequence. Each attention layer consists of 12 heads, which can each capture different contextual information in parallel. The final output from the encoder layers is 1026 vectors, each of length 768, where the first corresponds to a standard summary [CLS] token and the last corresponds to a [SEP] separator token. Many of the hyperparameters used in this paper, including the hidden size of 768 and 12 attention heads, are based on our previous work in ref.[12]

Pretraining

Pretraining of DR-BERT used masked language modeling (MLM): in each example, the model is tasked with identifying some hidden tokens. Following RoBERTa, the masks are set independently during epochs, and 15% of tokens are replaced with a [MASK] token for each example, with cross-entropy loss being applied for every batch of proteins.[48] Pretraining lasted for approximately 11 epochs, allowing the model to see 70 million examples. The batch size was set to 10 examples per device, and the model was trained on 2 NVIDIA V100s.[49]

Disordered region prediction

To finetune DR-BERT, we applied a token classification training method. A classification layer is trained and applied to each positional embedding output Then, a softmax function is applied to transform the embedding into probability space, taking the rounded result as the predicted label. Then, cross-entropy loss is applied between the predicted labels and the ground truths. The classification training lasted 10 epochs, with the best-performing checkpoint on the validation dataset chosen as the final model. The learning rate was empirically chosen to be $2e^{-6}$ (against $2e^{-5}$ and $2e^{-7}$) using the cosine scheduler with hard restarts, as opposed to a linear scheduler. To compare against similar models, DR-BERT was tested on the CAID dataset, which we ensured to be disjoint from both the training and evaluation datasets. We also tested on the CAID 2 dataset, which was again ensured to be disjoint from the training and evaluation datasets.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluation metrics

The primary evaluation metrics used for DR-BERT were Area Under the Receiver Operating Characteristic Curve (AU-ROC), F1 score and the Matthews Correlation Coefficient (MCC). The receiver operating curve is given by observing the change in the true positive to false positive ratio as the probability decision threshold is varied. Therefore, ROC-AUC for a perfect classifier would be 1.0 and a random classifier would have an area of 0.5. F1 scores are computed as a flattened vector of all predicted disorder binary labels against their ground truth, and is given by

$$F1 := \frac{2 * Precision * Recall}{Precision+Recall}, Recall := \frac{TP}{TP+FN}, Precision := \frac{TP}{TP+FP}$$

The MCC score offers a metric that is stable in imbalanced datasets.[50] Because of the MCC formula's false-positive symmetry, the MCC metric is invariant on which class is considered to be negative or positive. As the DisProt dataset has approximately 3 times as many ordered labels as disordered, the MCC metric is an appropriate metric to characterize the model's performance. MCC is defined as:

$$MCC := \frac{TP * TN - FP * FN}{\sqrt{(TP+FP) * (TP+FN) * (TN+FN) * (TN+FN)}}$$

Please cite this article in press as: Nambiar et al., DR-BERT: A protein language model to annotate disordered regions, Structure (2024), https://doi.org/10.1016/j.str.2024.04.010

For a fair comparison between methods, when these evaluation metrics were run, only test sequences that successfully ran on all methods were used. In addition, we attempted to emulate the evaluation strategy of the CAID competitions. In particular, when reporting F1 and MCC, we use the binary labels reported by CAID whenever available for a method since CAID identifies the threshold that maximizes F1 score for a particular method.[7] In the case of methods where a binary label was not provided by CAID and for DR-BERT, we identify the threshhold that maximizes F1 score ourselves. The threshhold for protein binding is calculated independently from the threshold for disordered region prediction (including disorder, disorder-PDB and disorder-noX). However, the same variant of DR-BERT was used for both disordered region and protein binding prediction.