SETH predicts nuances of residue disorder from protein embeddings

Dagmar Ilzhoefer[1,*], Michael Heinzinger[1,2,*,#], Burkhard Rost[1,3]

[1] Faculty of Informatics, Bioinformatics & Computational Biology, TUM (Technical University of Munich), Garching/Munich, Germany

[2] Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), TUM Graduate School, Garching, Germany

[3] Institute for Advanced Study (TUM-IAS), TUM (Technical University of Munich), Garching/Munich, Germany; TUM School of Life Sciences Weihenstephan (WZW), TUM (Technical University of Munich), Freising, Germany

[*] These authors have contributed equally to this work.

[#] Correspondence: Michael Heinzinger <mheinzinger@rostlab.org> or <assistant@rostlab.org>

Keywords: protein disorder, residue disorder, IDP, IDR, protein structure prediction, AlphaFold2, protein language model

Abstract

Predictions of millions of protein 3D structures are only a few clicks away since the release of *AlphaFold2* results for entire data sets. However, many proteins have so-called intrinsically disordered regions (IDRs) that do not adopt unique structures in isolation. These IDRs are associated with several diseases, including Alzheimer's Disease. We showed that the absence of reliable *AlphaFold2* predictions correlated only to a limited extent with IDRs. In contrast, many expert methods predict IDRs directly and reliably by combining complex machine learning models with expert-crafted input features and evolutionary information from multiple sequence alignments. Some of these input features are not always available and computationally expensive to generate, limiting their scalability. In this work, we present the novel prediction method SETH that predicts residue disorder from embeddings generated by the protein Language Model ProtT5, which explicitly only uses single sequences as input. Thereby SETH, a relatively shallow convolutional neural network, already outperformed much more complex state-of-the-art solutions while being much faster, allowing to create predictions for the human proteome in fewer than 30 minutes on a machine with one RTX A6000 GPU with 48GB RAM. Trained on a continuous disorder scale, our method captured subtle variations in disorder, thereby providing important information beyond the binary classification of other predictors. The new method is freely publicly available at: https://github.com/DagmarIlz/SETH.

1 Introduction

IDRs crucial for the functioning of organisms. Protein sequence determines protein three-dimensional (3D) structure, which, in turn, determines protein function. While this dogma usually refers to proteins folding into well-defined 3D structures, other proteins do not adopt unique 3D structures in isolation. Instead, these so-called intrinsically disordered proteins (Dunker et al., 2013) with intrinsically disordered regions (IDRs) sample their accessible conformational space, thereby expanding their functional spectrum (Wright and Dyson, 1999; Radivojac et al., 2004; Tompa et al., 2005; Tompa et al., 2006; Tompa et al., 2008; Uversky et al., 2009; Schlessinger et al., 2011), and possibly providing mechanisms to cope with evolutionary challenges (Tantos et al., 2009;Vicedo et al., 2015a;Vicedo et al., 2015b). For long IDRs, the difference between long IDR and long loops (neither helix nor strand) can be reliably predicted from sequences (Schlessinger et al., 2007b). For very short regions, IDRs and loops are technically not distinguishable in a predictive sense. Therefore, IDRs have to be longer than some minimal length Lmin for identification. While the precise value for Lmin remains obscure, Lmin=10 is clearly too short and Lmin=30 is clearly sufficient, as may be many values in between (Schlessinger et al., 2011). Using the more conservative Lmin=30, about 20-50% of all proteins in an organism are predicted to contain IDRs, with higher abundance in eukaryotes, especially in mammals (Romero et al., 1998;Liu et al., 2002;Schlessinger et al., 2011). Every fourth protein has been predicted as completely disordered (Dunker et al., 2008). This ubiquitous nature of disorder highlights its importance for the correct functioning of cells and makes the identification of IDRs crucial for understanding protein function. Alzheimer's disease and Huntington's disease, which are related to malfunctioning of disordered proteins/IDRs upon

mutation, further underline this importance (Dyson and Wright, 2005;Dunker et al., 2008).

Experimental best for IDRs: CheZOD scores. The experimental study of protein disorder remains difficult. X-ray crystallography is challenged by the lack of rigidity and nuclear magnetic resonance (NMR) remains limited to proteins shorter than average (~450 residues) (Howard, 1998;Oldfield et al., 2013;Nwanochie and Uversky, 2019). An additional complication is that upon binding to substrates, IDRs may appear ordered (Nielsen and Mulder, 2019). Arguably, today's best experimental approach toward capturing IDRs are NMR-derived chemical shift Z-scores (CheZOD scores), despite the length-limitation (Nielsen and Mulder, 2019). In contrast to binary measures such as "missing X-Ray coordinates" (Romero et al., 1998), CheZOD scores provide a well-calibrated measure for the nuances of per-residue disorder. CheZOD scores are computed from the difference of chemical shift values obtained in NMR spectroscopy (Howard, 1998) and computed random coil chemical shift values (Nielsen and Mulder, 2020).

Plenty of prediction methods available. The limited scalability of labor-intensive and expensive wet-lab experiments has spawned many computational tools for disorder prediction, including (from old to new): PONDR (Romero et al., 1998;Peng et al., 2005), NORSp (Liu et al., 2002), DISOPRED2 (Ward et al., 2004), IUPred (Dosztanyi et al., 2005), FoldIndex (Prilusky et al., 2005); RONN (Yang et al., 2005), PrDOS (Ishida and Kinoshita, 2007), NORSnet (Schlessinger et al., 2007a), MetaDisorder-MD (Schlessinger et al., 2009), ESpritz (Walsh et al., 2012), MetaDisorder (Kozlowski and Bujnicki, 2012), SPOT-Disorder (Hanson et al., 2016), and ODiNPred (Dass et al., 2020). As for almost every phenotype since the introduction of the combination of machine learning and evolutionary information (EI), derived from multiple sequence alignments (MSAs) (Rost and Sander, 1993), MSA-based predictions were best (Nielsen and Mulder, 2019;Dass et al., 2020). However, using EI/MSAs slows down inference and performs worse for proteins with a limited number of aligned sequences, which complicates the prediction of IDRs, which are inherently difficult to align due to, e.g., reduced sequence conservation in comparison to structured regions (Radivojac et al., 2002;Lange et al., 2016).

The prediction of protein structure from sequence leaped in quality through *AlphaFold2* (Jumper et al., 2021), Nature's method of the year 2021 (Marx, 2022). Although *AlphaFold2* appears to provide accurate predictions for only very few novel "folds", it importantly increases the width of structural coverage (Bordin et al., 2022). One important additional feature of *AlphaFold2*, other than providing the best 3D predictions from sequence/MSAs, is the estimate of performance depending on prediction strength, i.e., measuring prediction reliability as introduced for secondary structure prediction (Rost and Sander, 1993). The resulting score, dubbed pLDDT (predicted local distance difference test), distinguishes formidably between trustworthy and less reliable predictions (Jumper et al., 2021). Low values for pLDDT have been suggested to predict IDRs rather accurately (Wilson et al., 2021;Piovesan et al., 2022) or to predict non-existing proteins (Monzon et al., 2022).

Protein Language Models (pLMs) latest breakthrough. Inspired by recent leaps in Natural Language Processing (NLP), pLMs learn to predict masked tokens given their surrounding sequence of tokens (Asgari and Mofrad, 2015;Alley et al., 2019;Bepler and Berger, 2019;Heinzinger et al., 2019;Bepler and Berger, 2021;Elnaggar et al., 2021;Ofer et al., 2021;Rives et al., 2021;Wu et al., 2021). Toward this end, NLP words/tokens correspond to amino acids, while sentences correspond to full-length proteins in most current pLMs. As no information other than the amino acid sequence is required at any stage (*self-supervised learning*), pLMs efficiently leverage large but unlabeled databases with billions of protein sequences, such as BFD with 2.7b sequences (Steinegger et al., 2019). The information learned by the pLM during so-called (pre-) training can be retrieved and transferred (*transfer learning*) afterwards, by encoding a protein sequence in vector representations (embeddings). In their simplest form, embeddings mirror the last "hidden" states/values of pLMs. In analogy to NLPs implicitly learning grammar, embeddings from pLMs were shown to capture some aspects of the language of life as written in protein sequences (Heinzinger et al., 2019;Ofer et al., 2021), which suffices as exclusive input to many methods predicting aspects of protein structure and function (Asgari and Mofrad, 2015;Alley et al., 2019;Heinzinger et al., 2019;Elnaggar et al., 2021;Heinzinger et al., 2021;Littmann et al.,

2021a;Littmann et al., 2021b;Littmann et al., 2021c;Marquet et al., 2021;Rives et al., 2021).

Here, we bypass the problem of getting EI from MSAs for IDRs, by using embeddings from pre-trained protein language models (pLMs). In particular, we compared embeddings from five pLMs (ESM-1b (Rives et al., 2021), ProtBERT (Elnaggar et al., 2021), SeqVec (Heinzinger et al., 2019), ProtT5 (Elnaggar et al., 2021) and ProSE (Bepler and Berger, 2021)) by training a linear regression on each of the five pLM embeddings individually on predicting the degree of disorder of a residue as defined by the CheZOD scores. The pLMs themselves were not fine-tuned in any way to solving the prediction problem. Only the embeddings from the pLM that performed best when used as input to the linear regression were further used to train and compare four, in part, slightly more complex models, namely a logistic regression (LogReg), another linear regression (LinReg; trained on the full training set, as opposed to the linear regression used to compare the pLMs, only trained on 90% of the training set), a two-layer neural network (ANN), and a two-layer convolutional neural network (CNN; dubbed SETH (**S**elf-supervised **E**mbeddings predic**T** chemical s**H**ift Z-scores)). By training regression and classification models, we also investigated the benefit of training on nuanced CheZOD scores compared to binary disorder classification. The combination of using a rather simplistic model and embeddings from single protein sequences enabled our final method called SETH to predict disorder for the entirety of Swiss-Prot (The UniProt et al., 2021) in approximately 7 hours on a machine with one RTX A6000 GPU with 48GB RAM.

While recent, unpublished work implies that *AlphaFold2*'s pLDDT can be used for binary classification of disorder/order (Wilson et al., 2021;Piovesan et al., 2022), we also analyzed whether or not the *AlphaFold2* pLDDT provides information on nuances of disorder as reflected in the CheZOD scores.

2 Methods

2.1 Data sets

Raw data CheZOD scores. To streamline comparability to existing methods, we used two datasets available from ODiNPred (Dass et al., 2020) for training (file name CheZOD1325 in GitHub; 1325 proteins) and testing (file name CheZOD in GitHub; 117 proteins). Each dataset contains protein sequences and CheZOD scores for each residue. This score measures the degree of disorder of the residue and is calculated from the difference between chemical shift values obtained by NMR spectroscopy (Howard, 1998) and computed random coil chemical shifts (Nielsen and Mulder, 2020). These differences vary considerably between ordered and disordered residues, thereby continuously measuring the nuances of order/disorder for each residue (Nielsen and Mulder, 2020).

Redundancy reduction. To guarantee no relevant redundancy between train and test set, we constructed non-redundant subsets. Firstly, we built profiles (position specific scoring matrices; PSSMs) from multiple sequence alignments (MSAs) obtained through three iterations with *MMSeqs2* (Steinegger and Söding, 2017) (--num-iterations 3; other parameters were left at default values) against proteins in the training set. In the next step, any protein in the training set that shared more than 20% PIDE (percentage pairwise sequence identity) to any test set profile using bi-directional coverage, was removed using high-sensitivity (--s 7.5) search in MMSeqs2 (Steinegger and Söding, 2017). As the training set had been constructed such that any proteins within the set shared a maximum of 50% PIDE (Dass et al., 2020), no redundancy reduction within the training set was performed. Secondly, while embeddings were computed from full-length protein sequences, all residues not having a valid CheZOD score (indicated by CheZOD scores≥900) were excluded during training and evaluation. The resulting training set (dubbed *CheZOD1174*) contained 1174 proteins with a total of 132,545 residues (with an average length of 113 residues these proteins were about 3-4 times shorter than most existing proteins). The resulting dataset for testing (dubbed *CheZOD117*), contained 117 sequences with a total of 13,069 residues (average: 112). Consequently, we did not alter the test set published alongside ODiNPred, which has been used to evaluate 26 disorder prediction methods (Nielsen and Mulder, 2019), enabling a direct comparison of the results. However, as we altered the training data published and used for ODiNPred, to ensure that there was absolutely no overlap between training and test data, some methods from others may have allowed more for homology-based prediction, i.e., for

zooming into the residual levels of redundancy between train and test sets.

Dataset distributions. After preparing the data, the distributions of the CheZOD scores were analyzed for both datasets (Supplementary Figure 1). The CheZOD scores in these sets ranged from -5.6 to 16.2. A previously established threshold of eight differentiates between disorder (CheZOD≤8) and order (CheZOD>8) (Nielsen and Mulder, 2016). In both sets, the CheZOD score distributions were bimodal, but while there was an over-representation of ordered residues in CheZOD1174 (72% ordered), disordered residues were most prevalent in CheZOD117 (31% ordered). As artificial intelligence (AI) always optimizes for similar distributions in train and test, this discrepancy provided an additional safeguard against over-estimation of performance.

2.2 Data representation: embeddings

Five pLMs. Protein sequences from both sets (CheZOD117, CheZOD1174) were encoded as distributed vector representations (embeddings) with the following five pLMs: (1) SeqVec (Heinzinger et al., 2019), based on the NLP algorithm ELMo (Peters et al., 2018), is a stack of bi-directional long short-term memory cells (LSTM (Hochreiter and Schmidhuber, 1997)) trained on a 50% non-redundant version of UniProt (The UniProt et al., 2021) (UniRef50 (Suzek et al., 2015)). (2) ProtBERT (Elnaggar et al., 2021), based on the NLP algorithm BERT (Devlin et al., 2018), trained on the Big Fantastic Database, (Steinegger and Söding, 2018;Steinegger et al., 2019) with over 2.3 billion protein sequences. (3) ESM-1b (Rives et al., 2021), which is conceptually similar to (Prot)BERT (both use a stack of Transformer encoder modules (Vaswani et al., 2017)), but trained on UniRef50. (4) ProtT5-XL-U50 (Elnaggar et al., 2021) (dubbed ProtT5 for simplicity), based on the NLP sequence-to-sequence model T5 (Transformer encoder-decoder architecture) (Raffel et al., 2020), trained on the Big Fantastic Database and fine-tuned on Uniref50. (5) ProSE (Bepler and Berger, 2021), consisting of LSTMs trained on 76M unlabeled protein sequences in UniRef90 and additionally on predicting intra-residue contacts and structural similarity from 28k SCOPe proteins (Fox et al., 2014). While ProtBERT and ESM-1b were trained on reconstructing corrupted tokens/amino acids from non-corrupted (protein) sequence context (i.e., masked language modeling), ProtT5 was trained by teacher forcing, i.e., input and targets were fed to the model, with inputs being corrupted protein sequences and targets being identical to the inputs but shifted to the right (span generation with span size of 1 for ProtT5). In contrast, SeqVec was trained on predicting the next amino acid, given all previous amino acids in the protein sequence (referred to as auto-regressive pre-training). All pLMs, except for ProSE, were optimized only through self-supervised learning, i.e., exclusively using unlabeled sequences for pre-training. In contrast, ProSE was trained on three tasks simultaneously (multi-task learning), i.e., masked language modeling was used to train on 76M unlabeled sequences in UniRef90 and training to predict residue-residue contacts together with structural similarity was performed using 28k labeled protein sequences from SCOPe (Fox et al., 2014).

Embeddings: last hidden layer. Embeddings were extracted from the last hidden layer of the pLMs, with ProtT5 per-residue embeddings being derived from the last attention layer of the model's encoder-side using half-precision. The bio_embeddings package was used to generate the embeddings (Dallago et al., 2021). The resulting output is a single vector for each input residue, yielding an LxN-dimensional matrix (L: protein length, N: embedding dimension; N=1024 for SeqVec/ProtBERT/ProtT5; N=1280 for ESM-1b; N=6165 for ProSE).

Choosing embeddings best suited for IDR prediction. To find the most informative pLM embeddings for the prediction of IDRs/CheZOD score-residue disorder, we randomly chose 90% of the proteins in CheZOD1174 and trained a linear regression model on each of the five pLM embeddings individually to predict continuous CheZOD scores. The linear regressions were implemented with the *LinearRegression* module of scikit-learn (Pedregosa et al., 2011) with all parameters left at default values. Following this, the models were evaluated through two measures of performance, namely the Spearman correlation coefficient (ρ; Eqn. 2) and the AUC (area under the receiver operating characteristic curve; scikit-learn implementation). For this evaluation, the remaining 10% of CheZOD1174 were used (see section 2.5 for details). Only the best-performing pLM was used for further analyses.

Unsupervised embedding analysis. Lastly, the ProtT5 embeddings of CheZOD117 were analyzed in more detail on their disorder signal, by creating a t-distributed stochastic neighbor embedding (t-SNE; (van der Maaten and Hinton, 2008)) using the scikit-learn (Pedregosa et al., 2011) implementation. PCA (principle component analysis (Wold et al., 1987)) initialized the t-SNE to enable higher reliability of the resulting structure (Kobak and Berens, 2019). Furthermore, following a rule of thumb previously established (Kobak and Berens, 2019), the perplexity was chosen at the high value of 130 (1% of the sample size) to emphasize a global data structure (Kobak and Berens, 2019) in order to identify putative clusters of order or disorder. All other parameters were left at default values.

2.3 Disorder predictors and training

We optimized four different models to predict disorder: (1) linear regression (dubbed LinReg), (2) multi-layer artificial neural network (dubbed ANN), (3) two-layer CNN (dubbed SETH) and (4) logistic regression (dubbed LogReg). The models used throughout this work were deliberately kept simple. Firstly, because we tried to get the best possible fast prediction. Secondly, more complex models often do not improve performance significantly despite being computationally more expensive during training and deployment (Elnaggar et al., 2021;Littmann et al., 2021b;Marquet et al., 2021). Three of our models were trained on regression (LinReg, ANN and SETH), mapping from pLM embeddings to continuous CheZOD scores, while LogReg was trained on discriminating disordered from ordered residues (binary classification), mapping from pLM embeddings to binary labels, obtained by applying a previously established threshold of eight (Nielsen and Mulder, 2016) on the CheZOD scores.

LinReg, ANN and LogReg were implemented with scikit-learn (Pedregosa et al., 2011). For LinReg the *LinearRegression* model was used. For ANN, the *MLPRegressor* model was used and lastly, for LogReg, the *LogisticRegression* model was used. For the linear regression, all parameters were left at default levels. For LogReg, the class imbalance between ordered/disordered residues was taken into account by setting "class_weight" to balanced, which causes weighting inversely proportional to class frequency. Furthermore, the maximal number of iterations was chosen to be 400, while all other parameters were left at standard values. For ANN, the tanh was used as the activation function between hidden layers, an identity function was used in the final output layer, the random state was set to 1 to guarantee reproducibility and early stopping was activated. Furthermore, an optimization of the following hyperparameters was performed using scikit-learn's (Pedregosa et al., 2011) GridSearchCV on CheZOD1174: a) number of hidden neurons, b) solver (adam (Kingma and Ba, 2017) versus stochastic gradient descent (SGD) (Rumelhart et al., 1986)), and c) the learning rate. The ρ between the true and predicted CheZOD scores was used as the scoring function in this optimization. The best performing model (in terms of mean ρ), was used for any further analysis, resulting in the stochastic gradient descent, 2 hidden layers with 3 neurons each and a constant learning rate of 0.001 (Supplementary Figure 2). SETH was implemented with PyTorch (Paszke et al., 2019), where the convolutional layers were implemented with "Conv2d". Padding was applied to keep the dimension corresponding to the number of residues constant in the convolutions. MSELoss was used as the loss function and adam was used as the optimizer with a learning rate of 0.001, activating amsgrad (Reddi, 2018). The random seeds used were all set to 42 for reproducibility. For training, 90% of the proteins of CheZOD1174 were randomly chosen. The remaining 10%, the validation data, was used to determine early stopping, after 10 epochs of no improvement. Furthermore, the validation data was used to evaluate the hyperparameter optimization (Supplementary Figure 3). Out of the best performing models in this optimization, the model with the most constraints was chosen as the final model (Supplementary Figure 3, red bar), resulting in a kernel size of (5,1), 28 output channels of the first convolutional layer, the activation function Tanh between the 2 convolutional layers and the weight decay parameter of 0.001 in the optimizer.

2.4 AlphaFold2

AlphaFold2 (Jumper et al., 2021) predicts a reliability for the correctness of each residue prediction, namely, the pLDDT. This score has been correlated to binary descriptions of disorder (Wilson et al., 2021;Piovesan et al., 2022). To analyze *AlphaFold2* predictions against CheZOD scores, we predicted the 3D structures for all proteins in CheZOD117 using *ColabFold* (Mirdita et al., 2022). *ColabFold* speeds up

AlphaFold2 predictions 20-30x by replacing jackhmmer (Johnson et al., 2010) and HHblits (Remmert et al., 2012) in the computationally expensive MSA generation by *MMSeqs2* (Steinegger and Söding, 2017) without losing much in performance. We generated MSAs by searching UniClust30 (Mirdita et al., 2017) and the environment database ColabFoldDB (Mirdita et al., 2022). Neither templates nor Amber force-field relaxation (Hornak et al., 2006) were used, as those have been shown to not improve results significantly, on average (Jumper et al., 2021;Mirdita et al., 2022) while increasing compute time manifold (especially the Amber relaxation).

2.5 Evaluation of disorder predictions

In our performance evaluation, we followed the previous method comparison ((Nielsen and Mulder, 2019); comparing variants of MetaDisorder (Kozlowski and Bujnicki, 2012), SPOT-Disorder (Hanson et al., 2016), AUCPred with and without evolution (Wang et al., 2016), MFDp2 (Mizianty et al., 2013), PrDOS (Ishida and Kinoshita, 2007), RONN (Yang et al., 2005), DISpro (Cheng et al., 2005), DISOPRED2 (Ward et al., 2004), DISOPRED3 (Jones and Cozzetto, 2015), s2D (Sormanni et al., 2015), DynaMine (Cilia et al., 2013), variants of ESpritz (Walsh et al., 2012), DISPROT (VSL2b; (Vucetic et al., 2003)), variants of IUPred (Dosztanyi et al., 2005), variants of DisEMBL (Linding et al., 2003a), GlobPlot (Linding et al., 2003b)) by using the same test set and evaluation metrics, to ease comparability. Therefore, the ρs between the true CheZOD scores and the predicted CheZOD scores (or for LogReg and AlphaFold2, the predicted class probabilities or the pLDDT, respectively) were calculated for all our models.

The ρ and its 95% confidence interval (CI) were estimated over n = 1000 bootstrap sets in all cases (Efron and Tibshirani, 1991). For each bootstrap set, a random sample of the size of the test set (= m) was drawn with replacement from the test set. For each of these sampled sets, the ρ was calculated. If u_i is the rank of the ith value in the true CheZOD scores and v_i the rank of the ith value in the predicted CheZOD scores (or the rank of the respective values for LogReg and AlphaFold2) of the method, the ρ was calculated with Eqn. 2. The final ρ was derived from averaging over those 1000 values and the 95% CI was estimated by computing the standard deviation of the ρs over the sampled sets and multiplying it by 1.96. The standard deviation was calculated with Eqn. 1, where x_i is the ρ of an individual bootstrap set and ⟨x⟩ is the average ρ over all bootstrap sets.

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\langle x\rangle)^2}{n}} \quad \text{(Eqn. 1)}$$

$$\rho = \frac{\sum_{i=1}^{m}\left[\left(u_i-\frac{1}{m}\sum_{j=1}^{m}u_j\right)*\left(v_i-\frac{1}{m}\sum_{j=1}^{m}v_j\right)\right]}{\sqrt{\sum_{i=1}^{m}\left(u_i-\frac{1}{m}\sum_{j=1}^{m}u_j\right)^2*\sum_{i=1}^{m}\left(v_i-\frac{1}{m}\sum_{j=1}^{m}v_j\right)^2}} \quad \text{(Eqn. 2)}$$

Furthermore, the AUC and its 95% CI were estimated for each model evaluated here, again, by applying the same bootstrapping procedure. The AUC for each sampled set was calculated using the scikit-learn implementation. As the AUC requires class labels as its true reference values, continuous CheZOD scores were forced to be binarized using the threshold of eight: disorder CheZOD score≤8 and order CheZOD score>8 (Nielsen and Mulder, 2016).

2.6 Speed analysis

We analyzed the best method introduced here on its runtime for the human proteome and Swiss-Prot (The UniProt et al., 2021). This evaluation was performed on a machine with 2 AMD EPYC™ ROME 7352 CPUs at 2.30GHz each with 24/48 cores, a 256GB RAM (16 x 16GB) DDR4-3200MHz ECC, one RTX A6000 GPU with 48GB RAM, a 278GB SSD scratch disk and a 7.3TB HDD. The runtime reported included all steps required for receiving the predictions: load ProtT5, load the model checkpoint of the best method according to the evaluation, read in the sequences from the FASTA file, create embeddings, create predictions and lastly write the predictions into a file.

3 Results

Success of minimalist: single sequence, simple model. While state-of-the-art (SOTA) methods usually rely on some form of evolutionary information (EI) taken from MSAs as input to predict disorder, the method(s) introduced here use protein Language Models (pLMs) to encode single protein sequences as embeddings that served as the sole input feature for any prediction. To find the most informative pLM concerning disorder prediction, we predicted CheZOD scores through the minimalistic approach of linear regressions on top of embeddings from five pLMs (ProtT5: (Elnaggar et al., 2021), ProSE: (Bepler and Berger, 2021), ESM-1b: (Rives et al., 2021), ProtBERT: (Elnaggar et al., 2021), SeqVec: (Heinzinger et al., 2019)). Through the simplicity of a linear regression, we could focus on the

most informative signal for disorder in the embeddings. We established a random baseline through inputting random embeddings with 1024 dimensions sampled from a standard normal distribution. Following the recent assessment of 26 methods (Nielsen and Mulder, 2019), we calculated the ρ between true and predicted CheZOD scores and the AUC for 10% of CheZOD1174, not used in training, to evaluate the models. We calculated 95% CIs through bootstrapping (Efron and Tibshirani, 1991). The ρ describes the agreement between predicted and observed CheZOD scores, while the AUC provides some indication for the success of the binary classification into order and disorder.

Embeddings from all pLMs outperformed the random baseline, both for the correlation (Fig. 1A; ρ) and the binary projection of CheZOD scores (Fig. 1B; AUC). The simplest pLM-type included here, namely SeqVec, performed consistently and statistically significantly worse than all other pLMs (Fig. 1). The other four embeddings (ProSE, ESM-1b, ProtT5, ProtBERT) did not differ to an extent that was statistically significant, given the small data set. However, since the linear regression trained on ProtT5 reached the numerical top both in ρ and AUC, we used only embeddings from ProtT5 for further analyses.

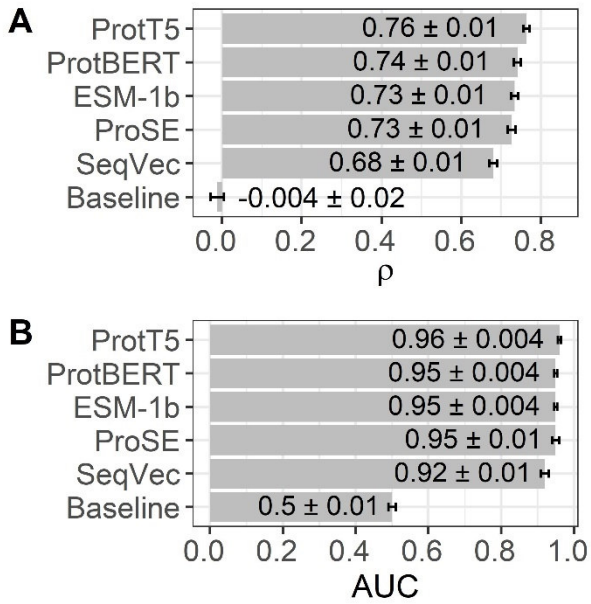

Fig. 1: Linear regression on five pLMs successful. Performance for training on 90% of set *CheZOD1174* (Dass et al., 2020) and testing on the remaining 10% simply using linear regressions fed by raw embeddings (without further optimization) from five protein language models (pLMs), namely: ProtT5 (Elnaggar et al., 2021), ProtBERT (Elnaggar et al., 2021), ESM-1b (Rives et al., 2021), ProSE (Bepler and Berger, 2021), SeqVec (Heinzinger et al., 2019). The sixth rows display the baseline/random predictions computed on 1024 dimensional embeddings sampled randomly from a standard normal distribution. Panel (A) depicts the Spearman correlation coefficient (ρ; Eqn. 2), calculated using the observed and predicted CheZOD scores, while Panel (B) required to first project predictions onto a binary state of disorder (CheZOD score≤8) /order (CheZOD score>8) and measures the area under the receiver operating characteristic curve (AUC). The errors mark the 95% confidence intervals approximated by multiplying 1.96 with the bootstrap standard deviation (Methods section 2.5).

ProtT5 captures disorder without any optimization. Next, we analyzed which information about disorder ProtT5 had already learned during self-supervised pre-training, i.e., before seeing any disorder-related labels. Towards this end, 1024-dimensional embeddings for all residues in all CheZOD117 sequences were projected to two dimensions using t-SNE (Fig. 2).

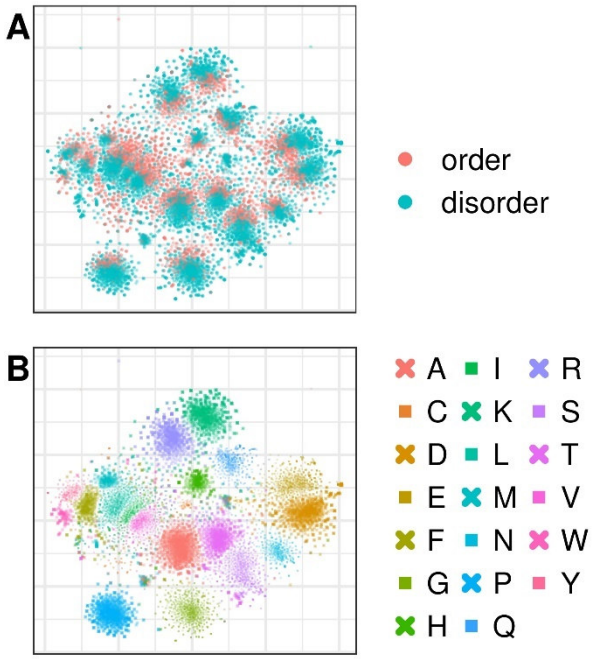

Fig. 2: Information of embeddings revealed by t-SNE. The t-SNE dimensionality reduction (van der Maaten and Hinton, 2008) was performed on the 1024-dimensional ProtT5 (Elnaggar et al., 2021) residue-level embeddings for all sequences in test set *CheZOD117* (13,069 residues; (Dass et al., 2020)), extracted from the last attention layer of ProtT5. Panel (A) shows the embeddings colored by order (CheZOD score>8; red) and disorder (CheZOD score≤8, blue) (Nielsen and Mulder, 2016). Panel (B) shows the same t-SNE projection but with coloring by the 20 standard

amino acid types (here shown in one-letter code; A=Alanine, C=Cysteine, D=Aspartic acid, E=Glutamic acid, F=Phenylalanine, G=Glycine, H=Histidine, I=Isoleucine, K=Lysine, L=Leucine, M=Methionine, N=Asparagine, P=Proline, Q=Glutamine, R=Arginine, S=Serine, T=Threonine, V=Valine, W=Tryptophan, Y=Tyrosine).

The resulting image suggests some level of separation between ordered (red) and disordered (blue) residues (Fig. 2A: red colors oriented toward the Fig. center in each cluster), indicating that even raw ProtT5 embeddings already captured some aspects of disorder without having used any such annotations (ProtT5 only learned to predict masked amino acid tokens (Elnaggar et al., 2021)). However, the major signal seemingly did not cluster the disorder/order phenotype. Coloring residues according to amino acid type showed that this "phenotype" provided the predominant signal for generating the clusters (Fig. 2B).

SETH (CNN) outperformed other supervised models. Next, we compared different AI-models inputting ProtT5 embeddings to predict continuous CheZOD scores (1-3) or binary disorder (4): (1) LinReg: linear regression trained on the full training set instead of the 90% subset used before, (2) ANN: 2-layer neural network with about three times more free parameters than LinReg, (3) SETH: 2-layer CNN, (4) LogReg: logistic regression analyzing the effect of transferring the regression problem to a classification problem by predicting the mapping of the continuous CheZOD scores to either disorder (CheZOD scores≤8) or order (CheZOD scores>8).

As we evaluated all four models on the same test set (CheZOD117) and used the same metrics (ρ and AUC) as a recent method comparison (Nielsen and Mulder, 2019), we could add the performances of our methods to this comparison (Fig. 3). We also evaluated the ODiNPred web application (Dass et al., 2020) on CheZOD117 and added this to the comparison. When considering the mean ρ (Fig. 3A), our methods SETH, LinReg and ANN outperformed all others not using EI (below dashed line in Fig. 3), as well as the methods relying on EI (above dashed line in Fig. 3). When requiring a statistically significant difference at the 95% CI (±1.96 standard errors) for the ρ, our methods (SETH, LinReg, ANN and LogReg) significantly outperformed all others, except for ODiNPred. When evaluating the performance based on the mean AUC, all our methods, even the simplistic linear regression, outperformed all other evaluated methods. Due to the already high AUC levels, the absolute improvement of our models in terms of AUC was numerically smaller compared to the improvement in the ρ and, consequently, often not statistically significant.

Comparing the models introduced here based on the AUC or the ρ revealed no significant difference between LogReg, LinReg, ANN and SETH. However, SETH had the highest mean ρ and, together with LinReg, the highest mean AUC.

For a more detailed analysis of the correlation between ground truth (CheZOD scores from NMR and computed estimates (Nielsen and Mulder, 2020)) and predictions, we plotted the true and predicted CheZOD scores (or for LogReg the true CheZOD scores and the predicted probability for the class "order") for CheZOD117 against each other in a 2D histogram for all four models (Fig. 4).

SETH (Fig. 4A), ANN (Fig. 4B) and LinReg (Fig. 4C) agreed well with the ground truth (matching the high ρ-values). However, SETH's and ANN's points clustered more closely around the diagonal than LinReg's. Additionally, the plots revealed that SETH, LinReg and ANN tended to overestimate residue order, as indicated by the higher prediction density above the diagonal. In contrast to our other models, most of the pairs of LogReg's predicted order probability vs. true CheZOD scores fell into two flat clusters at 0 and 1, confirming that LogReg tended to predict extreme values optimal for classification. Thereby, it lost the subtle nuances of a residue's "(dis-) orderdness". This was also reflected in the lower mean ρ of LogReg, compared to SETH, LinReg or ANN.

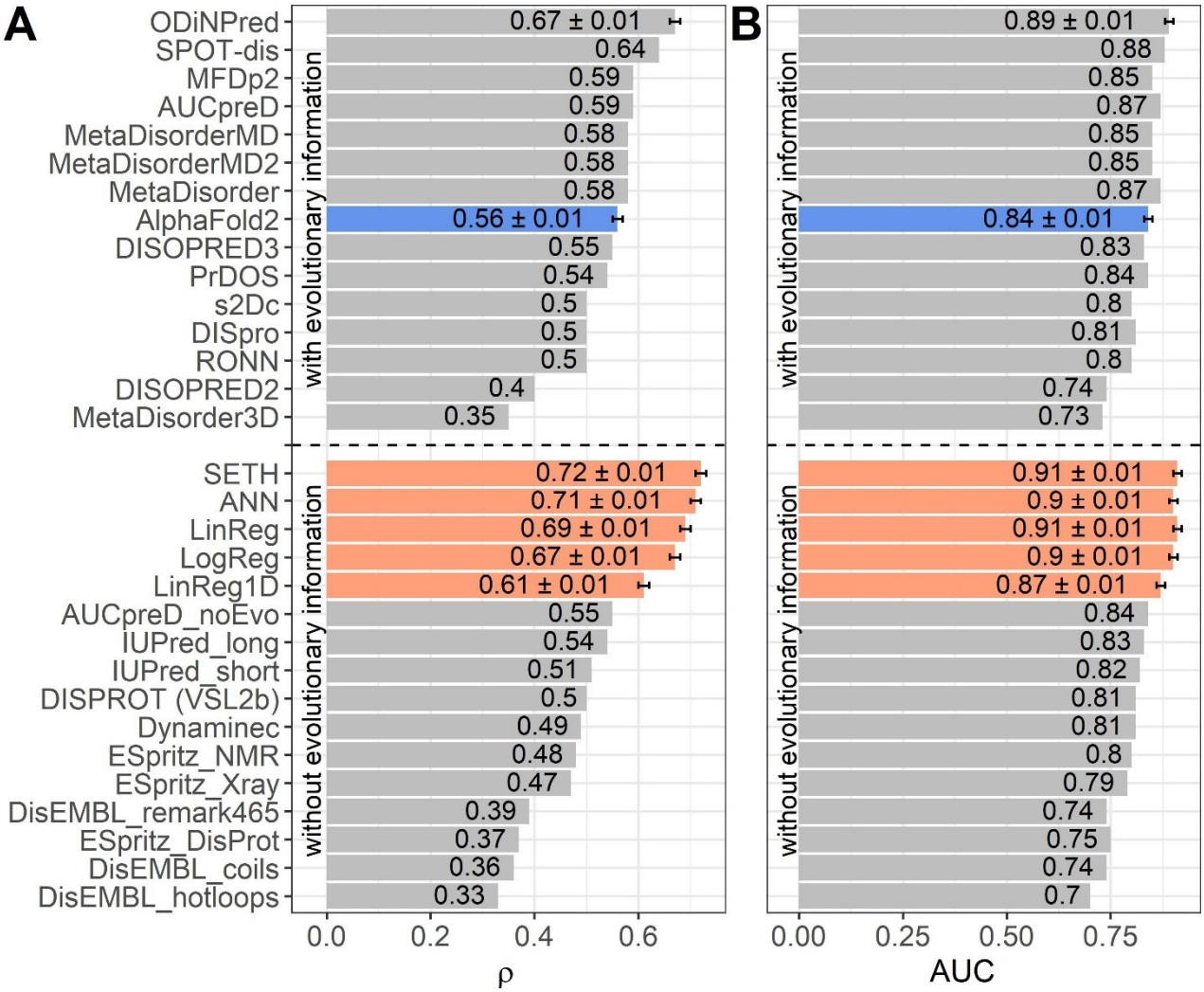

Fig. 3: pLM-based methods performed outstandingly. Data: set *CheZOD117* (Dass et al., 2020). All methods introduced here (SETH, ANN, LinReg, LogReg, LinReg1D) in red, the ODiNPred web application in grey (Dass et al., 2020), for *AlphaFold2* ((Jumper et al., 2021); in blue), we used its pLDDT score as proxy for order, all other performances were taken from the previous comparison (Nielsen and Mulder, 2019) using the same test set. While three of our models (SETH, ANN, LinReg/LinReg1D) and ODiNPred were trained on continuous chemical shift Z-scores (CheZOD scores), LogReg is a logistic regression trained on binary classification of order/disorder (ODiNPred used more proteins for training than our pLM-based models). The dotted line separates models using evolutionary information from multiple sequence alignments (above line) from single sequence-based methods (below line). Error bars mark the 95% confidence interval, approximated by bootstrapping for our methods, *AlphaFold2* and the ODiNPred web application. Panel (A): Spearman correlation coefficient (ρ; Eqn. 2), Panel (B): Area under the receiver operating characteristic curve (AUC) after binary projection of CheZOD scores (order: CheZOD score>8, disorder: CheZOD score≤8; (Nielsen and Mulder, 2016)).

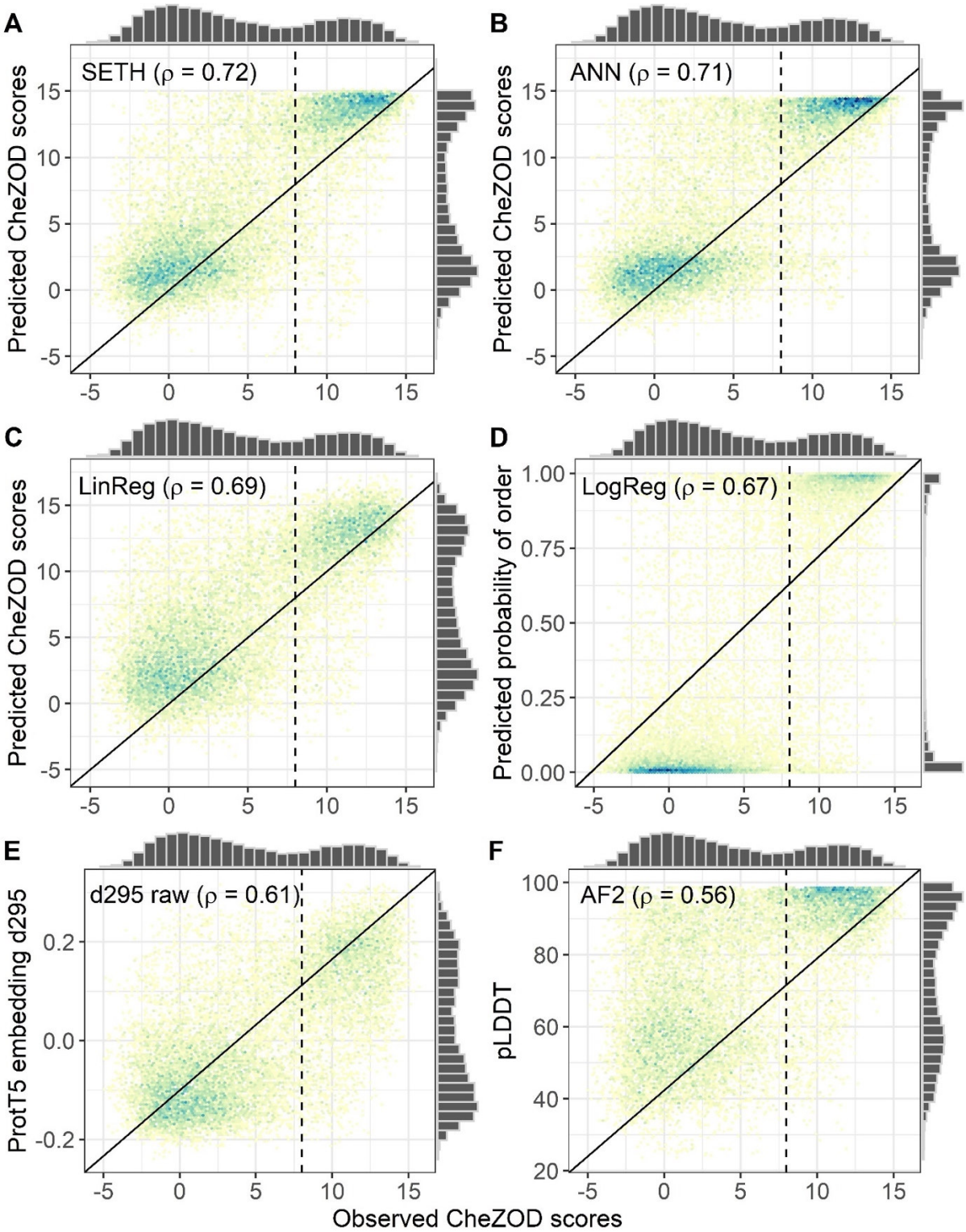

Fig. 4: Nuances about per-residue CheZOD scores accurately predicted. Data: set CheZOD117 (13,069 residues; (Dass et al., 2020)). Histogram of the observed CheZOD score (x-axis) against predictions for six methods (y-axis). Additionally, a marginal histogram is visible for each axis. Methods introduced here: (A) SETH, (B) ANN, (C) LinReg, (D) LogReg, (E) raw embedding dimension 295 (d295 raw) of ProtT5, which is the dimension with the highest regression coefficient in LinReg. (F) AlphaFold2 (AF2; (Jumper et al., 2021)). The black diagonal in each plot marks the optimal regression fit. Vertical dotted black lines separate ordered (CheZOD score>8) from disordered residues (CheZOD score≤8; (Nielsen and Mulder, 2016)). The Spearman correlation coefficient (ρ) was estimated from bootstrapping in all panels. For all panels except panel (D), the same colors correspond to the same number of points in an area.

SETH very fast. Although three of our models outperformed all other methods compared, we would always pick the best of the three, namely SETH, when analyzing proteins and proteomes. Thus, we next analyzed the speed of SETH. On a machine with one RTX A6000 GPU with 48GB RAM, predicting the nuances of disorder for each residue of the entire human proteome (20,352 proteins) from the individual protein sequences took approximately 23 minutes. For Swiss-Prot (566,969 proteins; (The UniProt et al., 2021)), it took approximately seven hours. Consequently, SETH can predict disorder for approximately 15-20 proteins in 1 second.

The most informative embedding dimension / LinReg1D. After training, we also analyzed the regression coefficients of LinReg to better understand how ProtT5 embeddings affected the prediction. For the dimension with the highest regression coefficient (dimension 295 of 1024; Supplementary Figure 4), we subsequently plotted the raw embedding values against the true CheZOD scores (Fig. 4E) to visualize the information on order/disorder in the embeddings without supervised training. The Spearman correlation for this single dimension (ρ = 0.61) was almost the same as that for all 1024-d used by LinReg (ρ = 0.69), showing that the pLM already learned aspects of disorder during self-supervised pre-training, i.e., without ever seeing such labels. However, in contrast to LinReg, the single dimension without supervised training avoided to overestimate residue order (no accumulation of high density above the diagonal; Fig. 4).

In order to explicitly quantify the influence of this single most informative dimension, we additionally trained and evaluated a linear regression inputting only this 295th embedding dimension (dubbed LinReg1D). LinReg1D reached a ρ of 0.61 (LinReg ρ=0.69) and an AUC of 0.87 (LinReg=0.91, Fig. 3). Therefore, this single dimension accounted for 89% or 96% of the performance of LinReg, when considering the ρ or the AUC respectively. As only a linear transformation was performed from the raw values to LinReg1D, both showed the same ρ when correlated with the true CheZOD scores (Fig. 4, Supplementary Figure 5).

When comparing LinReg1D to the predictors evaluated in the large-scale comparison of disorder predictors (Nielsen and Mulder, 2019) and ODiNPred, even this extremely reduced model outperformed all other methods not using EI and only fell short compared to the two best-performing methods using EI (SPOT-Disorder and ODiNPred), when looking at both the AUC and the ρ (Fig. 3). However, compared to our other methods (SETH, LinReg, ANN, LogReg), LinReg1D performed significantly worse.

AlphaFold2 captured CheZOD scores less well than older and newer top methods. AlphaFold2's predicted reliability pLDDT was recently reported to capture some aspects of binary disorder (Wilson et al., 2021;Piovesan et al., 2022). However, we discovered that the ρ between AlphaFold2's pLDDT and CheZOD scores clearly neither reached the levels of the top expert solutions (SETH, LinReg, ANN, LogReg, LinReg1D or ODiNPred; Fig. 3A) trained on CheZOD scores, nor that of many other methods using EI (Nielsen and Mulder, 2019). However, given that AlphaFold2's pLDDT was optimized to estimate the reliability of the 3D prediction, the agreement between the pLDDT and CheZOD scores was still remarkable (ρ=0.56). For instance, AlphaFold2 exceeded all disorder predictors not using EI evaluated in the large-scale comparison of disorder prediction methods (Nielsen and Mulder, 2019). Looking at the correlation between pLDDT scores and CheZOD scores in more detail (Fig. 4F) revealed that disordered residues (CheZOD≤8) were occasionally predicted with high confidence (pLDDT>80), which explained the rather low ρ. The performance of AlphaFold2's pLDDT for the binary projection of CheZOD scores measured with the AUC was largely similar, i.e., AlphaFold2 performed similar to the best existing methods not using EI but fell short of expert solutions relying on EI and all models introduced here (Fig. 3B). The distance between AlphaFold2 and the best model was, however, not as large as for the ρ.

4 Discussion

We introduced SETH, a CNN for predicting the continuum of residue disorder defined by CheZOD scores (Nielsen and Mulder, 2020), exclusively using embeddings from the pLM ProtT5 (Elnaggar et al., 2021) as input. We evaluated SETH, along with three even simpler pLM-based models, using identical (or slightly more conservative: Methods section 2.1 redundancy reduction) conditions as used recently to evaluate 26 disorder prediction methods (Nielsen and Mulder, 2019). We also added AlphaFold2's predicted reliability pLDDT to this comparison motivated by

recent findings (Wilson et al., 2021;Piovesan et al., 2022).

Supervised models picked up class imbalance. The datasets used here for training (*CheZOD1174*) and testing (*CheZOD117*) were chosen to ease comparability to existing methods (Nielsen and Mulder, 2019). However, we removed 151 sequences from ODiNPred's (Dass et al., 2020) training set that shared more than 20% sequence similarity to the test set *CheZOD117* to avoid information leakage. The resulting training and test sets differed substantially in their distributions of CheZOD scores (Supplementary Figure 1), with the training set having more ordered residues than the test set (train: 72% ordered vs. test: 31%). Since our regression models did not use any notion of classes, class imbalance was not corrected during training. This might explain why our supervised regression models trained on this imbalanced data (SETH, LinReg and ANN) mildly over-predicted the degree of residue order compared to the raw embedding values of dimension 295 (Fig. 4).

Similar performance for five pLMs. We trained simple linear regression models to predict CheZOD scores using embeddings from five pLMs, namely ProtT5 (Elnaggar et al., 2021), ProSE (Bepler and Berger, 2021), ESM-1b (Rives et al., 2021), ProtBERT (Elnaggar et al., 2021) and SeqVec (Heinzinger et al., 2019). Only SeqVec performed significantly worse than the others when evaluated on 10% of the training set (Fig. 1). The other four performed within the 95% CI of each other for this small data set. Since ProtT5 consistently reached the highest mean ρ and mean AUC and since it also outperformed other pLMs in previous studies (Heinzinger et al., 2021;Marquet et al., 2021;Stärk et al., 2021), we decided to focus on ProtT5 in our subsequent work.

Simple classification model LogReg struggled where SETH excelled. We carefully tested the effect of increasing the model complexity when inputting only embeddings. The simplistic linear regression (LinReg), along with the more complex ANN and SETH (CNN), established that the treatment of disorder as a regression problem improved over the supervised training on binary assignments (disorder/order; LogReg; Fig. 3). This was interesting because except for ODiNPred (Dass et al., 2020), most SOTA prediction methods for disorder realize a binary classification.

An ideal prediction method clusters pairs observed/predicted around the diagonal (Fig. 4). Qualitatively, SETH and ANN came closest to this, while LinReg had more spread-out clusters (Fig. 4A-C). In contrast to an ideal prediction, LogReg generated two clusters, one around probability 0 (disorder) and the other around 1 (order; Fig. 4D). Although such a bifurcation is expected for a logistic regression trained to classify, the off-diagonal shift of the data showed that LogReg struggled to capture subtle degrees of disorder/order, showing the limitations of training binary classification models on this problem. This qualitative analysis was supported by the ρ (Fig. 3A: SETH highest, LogReg lowest). However, the ρ was relatively similar between all four models, including LogReg. Nonetheless, SETH was consistently superior by all criteria (Fig. 3, 4).

Significant improvement over SOTA with simpler models. When comparing the performance of more complex solutions to the four models introduced here (SETH, LinReg, ANN, LogReg), we showed that fast and simple methods trained on top of ProtT5 embeddings numerically outperformed all existing approaches in terms of mean AUC (Fig. 3B), irrespective of whether they use evolutionary information (EI) from MSAs or not. In terms of mean ρ (Fig. 3A), three of our methods (SETH, LinReg and ANN) still outperformed all other models, while LogReg performed on par with ODiNPred. In terms of statistical significance for the ρ at the CI=95% level, all our models and ODiNPred significantly outperformed all other models evaluated.

For our best model SETH, the numerical improvement over ODiNPred was larger for the ρ than for the AUC. However, considering relative values, the improvement for the ρ was 15% of the maximally possible improvement, while that for AUC was 18%. Consequently, the seemingly small absolute improvement in AUC originated mostly from an already high performance which rendered any further improvement increasingly difficult.

Disorder successfully predicted by single embedding dimension. ProtT5 embeddings have 1024 dimensions for each residue. One of those 1024 (dimension 295) carried 86%-96% of the signal of the entire vector. Investigating further why this embedding dimension contains specific information

about one aspect of proteins, namely disorder, and if something similar can be observed for other dimensions and other tasks, could give valuable insights into pLMs in general.

AlphaFold2 pLDDT not competitive as proxy for CheZOD score disorder. It has recently been suggested that AlphaFold2's pLDDT score is a reasonable indicator for disorder (Wilson et al., 2021;Piovesan et al., 2022). By plotting pLDDT against CheZOD scores (Fig. 4), we confirmed this basic correlation. In generalizing from the previous analyses using binary definitions of disorder/order, to the continuous scale of CheZOD scores, we additionally revealed that AlphaFold2 was often certain about a predicted structure (high pLDDT) even for highly disordered residues. This might indicate AlphaFold2 predictions to be overly optimistic in some of these cases, as disordered residues by definition do not adapt a rigid structure, making highly reliable predictions in this region unlikely. One possible explanation for this might be that while AlphaFold2 was only trained on single protein domains, some of these proteins were measured as homo- or heteromers. Consequently, the AlphaFold2 predictions might be biased in regions that are disordered in isolation but become rigid upon interaction. Furthermore, the mean pLDDT is trivially higher for shorter than for longer proteins (Monzon et al., 2022). Since the mean protein length for the test set *CheZOD117* was only 112, this could also explain some outliers. More detailed analyses of AlphaFold2 predictions were beyond the scope of this work. Lastly, in agreement with recent results (Wilson et al., 2021;Piovesan et al., 2022), we found AlphaFold2 to be clearly outperformed by disorder specialists (Fig. 3, 4). Clearly, given the many times higher runtime (we ran AlphaFold2 (without the MSA generation step and using early stopping when one of five models reached a pLDDT≥85) and SETH on the machine with one RTX A6000 GPU with 48GB RAM and AlphaFold2 took approximately 170 times as long as SETH), SETH appeared by far a better method for predicting disorder as defined by CheZOD scores than AlphaFold2. Even when AlphaFold2 predictions will be available for all proteins, the degree to which SETH outperformed pLDDT scores and the speed of the method may suggest to always use SETH instead of AlphaFold2 to predict CheZOD-like disorder.

5 Conclusions

We introduced several novel methods exclusively using embeddings from the protein Language Model ProtT5 (Elnaggar et al., 2021) to predict per-residue protein disorder/order as proxied by CheZOD scores (based on NMR measurements (Nielsen and Mulder, 2020)). The best approach, dubbed SETH, captured fine-grained nuances of disorder on a continuous scale and outperformed the state-of-the-art methods (Nielsen and Mulder, 2019) using evolutionary information derived from multiple sequence alignments (Fig. 3). The solution was so successful because the unoptimized embeddings carried important information about disorder (Fig. 2), to the extent that mostly one of the 1024 dimensions mattered (Fig. 4E). Since SETH exclusively uses embeddings of single protein sequences, it easily scales to the analysis of entire proteomes, e.g., (dis-) order of all human proteins can be predicted in fewer than 30 minutes. Although the break-through AlphaFold2 (Jumper et al., 2021) 3D predictions may soon become available for most proteins, and although the per-residue reliability score (pLDDT) of AlphaFold2 predictions somehow correlates with disorder, the correlation between low pLDDT and CheZOD scores (Fig. 4) was so much inferior to the predictions of SETH to suggest the investment of fewer than 2 minutes per 1,000 proteins.

6 Software and prediction availability

SETH is available to download at https://github.com/DagmarIlz/SETH and available for online execution (no setup on your machine required) at https://colab.research.google.com/drive/1vDWh5YI_BPxQg0ku6CxKtSXEJ25u2wSq?usp=sharing. The predictions of SETH for Swiss-Prot (The UniProt et al., 2021) and the human proteome are available at https://doi.org/10.5281/zenodo.6673817.

7 Funding

This work was supported by the Bavarian Ministry of Education through funding to the TUM and by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium für Bildung und Forschung), by two grants from BMBF (031L0168 and program "Software Campus 2.0 (TUM) 2.0"

01IS17049) as well as by a grant from Deutsche Forschungsgemeinschaft (DFG-GZ: RO1320/4-1).

8 Acknowledgments

Particular thanks to Keith Dunker (Indiana U) for his help in realizing the importance of the topic of IDRs and IDPs. Further thanks primarily to Tim Karl (TUM) for invaluable help with hardware and software and to Inga Weise (TUM) for crucial support with many other aspects of this work. Thanks to Maria Littmann and Michael Bernhofer (both TUM) for listening, helping and contributing their insights in discussions. Last, but not least, thanks to all those who contribute toward maintaining public databases, and to all experimentalists who enabled this analysis by making their data publicly available.

9 Abbreviations

3D: three-dimensional, i.e., co-ordinates of all atoms/residues in a protein; **AI:** artificial intelligence; ***AlphaFold2***: AI-based method reliably predicting protein 3D structure from EI/MSAs (Jumper et al., 2021); **ANN:** artificial feed-forward neural network; **AUC**: area under the receiver operating characteristic curve; **CheZOD scores**: chemical shift Z-scores (Nielsen and Mulder, 2019); **CI:** confidence interval, here typically used as the 95% CI implying an interval between ±1.96*StandardError; **CNN:** convolutional neural network; **EI:** evolutionary information; **IDR**: intrinsically disordered region; **LogReg:** logistic regression; **LSTM:** long short-term memory cell; **MSA**: multiple sequence alignment; **NLP:** Natural Language Processing; **NMR:** nuclear magnetic resonance; **PIDE**, percentage pairwise sequence identity; **pLDDT:** predicted local distance difference test from *AlphaFold2* (Jumper et al., 2021); **pLM:** protein language model; **SOTA**: state-of-the-art; **t-SNE**: t-distributed stochastic neighbor embedding; **ρ**: Spearman correlation coefficient.

10 References

Alley, E.C., Khimulya, G., Biswas, S., Alquraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* 16, 1315-1322.

Asgari, E., and Mofrad, M.R.K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE* 10, e0141287.

Bepler, T., and Berger, B. (2019). Learning protein sequence embeddings using information from structure. *arXiv:1902.08661 [cs, q-bio, stat]*.

Bepler, T., and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. *Cell Systems* 12, 654-669.e653.

Bordin, N., Sillitoe, I., Nallapareddy, V., Rauer, C., Lam, S.D., Waman, V.P., Sen, N., Heinzinger, M., Littmann, M., Kim, S., Velankar, S., Steinegger, M., Rost, B., and Orengo, C. (2022). AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *bioRxiv*, 2022.2006.2002.494367.

Cheng, J., Sweredoski, M.J., and Baldi, P. (2005). Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and Knowledge Discovery* 11, 213-222.

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W.F. (2013). From protein sequence to dynamics and disorder with DynaMine. *Nature Communications* 4, 2741.

Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A.X., Yang, K.K., Min, S., Yoon, S., Morton, J.T., and Rost, B. (2021). Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. *Current Protocols* 1.

Dass, R., Mulder, F.a.A., and Nielsen, J.T. (2020). ODiNPred: comprehensive prediction of protein order and disorder. *Scientific Reports* 10, 14780.

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

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433-3434.

Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, Z., Dyson, H.J., Forman-Kay, J., Fuxreiter, M., Gsponer, J., Han, K.-H., Jones, D.T., Longhi, S., Metallo, S.J., Nishikawa, K., Nussinov, R., Obradovic, Z., Pappu, R.V., Rost, B., Selenko, P., Subramaniam, V., Sussman, J.L., Tompa, P., and Uversky, V.N. (2013). What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. *Intrinsically Disordered Proteins* 1, e24157.

Dunker, A.K., Silman, I., Uversky, V.N., and Sussman, J.L. (2008). Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology* 18, 756-764.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology* 6, 197-208.

Efron, B., and Tibshirani, R. (1991). Statistical Data Analysis in the Computer Age. *Science* 353, 390-395.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2021). ProtTrans: Towards

Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *arXiv:2007.06225 [cs, stat]*.

Fox, N.K., Brenner, S.E., and Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* 42**,** D304-309.

Hanson, J., Yang, Y., Paliwal, K., and Zhou, Y. (2016). Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics***,** btw678.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* 20**,** 723.

Heinzinger, M., Littmann, M., Sillitoe, I., Bordin, N., Orengo, C., and Rost, B. (2021). "Contrastive learning on protein embeddings enlightens midnight zone". Bioinformatics).

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation* 9**,** 1735-1780.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65**,** 712-725.

Howard, M.J. (1998). Protein NMR spectroscopy. *Current Biology* 8**,** R331-R333.

Ishida, T., and Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research* 35**,** W460-W464.

Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics* 11**,** 431-431.

Jones, D.T., and Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics (Oxford, England)* 31**,** 857-863.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.a.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596**,** 583-589.

Kingma, D.P., and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications* 10**,** 5416.

Kozlowski, L.P., and Bujnicki, J.M. (2012). MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics* 13**,** 111.

Lange, J., Wyrwicz, L.S., and Vriend, G. (2016). KMAD: knowledge-based multiple sequence alignment for intrinsically disordered proteins. *Bioinformatics* 32**,** 932-936.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. (2003a). Protein Disorder Prediction. *Structure* 11**,** 1453-1459.

Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. (2003b). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic acids research* 31**,** 3701-3708.

Littmann, M., Bordin, N., Heinzinger, M., Schütze, K., Dallago, C., Orengo, C., and Rost, B. (2021a). Clustering FunFams using sequence embeddings improves EC purity *Bioinformatics* 37**,** 3449–3455.

Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T., and Rost, B. (2021b). Embeddings from deep learning transfer GO annotations beyond homology. *Sci Rep* 11**,** 1160.

Littmann, M., Heinzinger, M., Dallago, C., Weissenow, K., and Rost, B. (2021c). Protein embeddings and deep learning predict binding residues for various ligand types. *Scientific Reports* 11**,** 23916.

Liu, J., Tan, H., and Rost, B. (2002). Loopy proteins appear conserved in evolution. *Journal of Molecular Biology* 322**,** 53-64.

Marquet, C., Heinzinger, M., Olenyi, T., Dallago, C., Erckert, K., Bernhofer, M., Nechaev, D., and Rost, B. (2021). Embeddings from protein language models predict conservation and variant effects. *Human Genetics*.

Marx, V. (2022). Method of the Year: protein structure prediction. *Nat Methods* 19**,** 5-10.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold - Making protein folding accessible to all. *bioRxiv***,** 2021.2008.2015.456425.

Mirdita, M., Von den driesch, L., Galiez, C., Martin, M.J., Söding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* 45**,** D170-D176.

Mizianty, M.J., Peng, Z., and Kurgan, L. (2013). MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. *Intrinsically disordered proteins* 1**,** e24428-e24428.

Monzon, V., Haft, D.H., and Bateman, A. (2022). Folding the unfoldable: using AlphaFold to explore spurious proteins. *Bioinformatics Advances* 2**,** vbab043.

Nielsen, J.T., and Mulder, F.a.A. (2016). There is Diversity in Disorder—"In all Chaos there is a Cosmos, in all Disorder a Secret Order". *Frontiers in Molecular Biosciences* 3.

Nielsen, J.T., and Mulder, F.a.A. (2019). Quality and bias of protein disorder predictors. *Scientific reports* 9**,** 5137.

Nielsen, J.T., and Mulder, F.a.A. (2020). "Quantitative Protein Disorder Assessment Using NMR Chemical Shifts," in *Intrinsically Disordered Proteins,* eds. B.B. Kragelund & K. Skriver. (New York, NY: Springer US), 303-317.

Nwanochie, E., and Uversky, V.N. (2019). Structure determination by single-particle cryo-electron microscopy: Only the sky (and intrinsic disorder) is the limit. *International journal of molecular sciences* 20**,** 4186.

Ofer, D., Brandes, N., and Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal* 19, 1750-1758.

Oldfield, C.J., Xue, B., Van, Y.-Y., Ulrich, E.L., Markley, J.L., Dunker, A.K., and Uversky, V.N. (2013). Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochimica Et Biophysica Acta* 1834, 487-498.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., Devito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (Year). "PyTorch: An Imperative Style, High-Performance Deep Learning Library", eds. H. Wallach, H. Larochelle, A. Beygelzimer, F.D. Alché-Buc, E. Fox & R. Garnett: Curran Associates, Inc.).

Pedregosa, F., Varoquaux , G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., Dunker, A.K., and Obradovic, Z. (2005). Optimizing long intrinsic disorder predictors with protein evolutionary information. *J Bioinform Comput Biol* 3, 35-60.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations.

Piovesan, D., Monzon, A.M., and Tosatto, S.C.E. (2022). Intrinsic Protein Disorder, Conditional Folding and AlphaFold2. *bioRxiv*.

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., and Sussman, J.L. (2005). FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 21, 3435-3438.

Radivojac, P., Obradovic, Z., Brown, C.J., and Dunker, A.K. (2002). Improving sequence alignments for intrinsically disordered proteins. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 589-600.

Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D., and Dunker, A.K. (2004). Protein flexibility and intrinsic disorder. *Protein Science* 13, 71-80.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P.J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv*.

Reddi, S.J., Kale, S., & Kumar, S. (2018). On the Convergence of Adam and Beyond. *International Conference on Learning Representations.*

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* 9, 173-175.

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. *Proceedings of the National Academy of Sciences* 118, e2016239118.

Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E., Guilliot, S., and Dunker, A.K. (1998). Thousands of proteins likely to have long disordered regions. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 437-448.

Rost, B., and Sander, C. (1993). Prediction of Protein Secondary Structure at Better than 70% Accuracy. *Journal of Molecular Biology* 232, 584-599.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. *nature* 323, 533-536.

Schlessinger, A., Liu, J., and Rost, B. (2007a). Natively unstructured loops differ from other loops. *PLoS Computational Biology* 3, e140.

Schlessinger, A., Punta, M., and Rost, B. (2007b). Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics* 23, 2376-2384.

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., and Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE* 4, e4433.

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., and Rost, B. (2011). Protein disorder—a breakthrough invention of evolution? *Current Opinion in Structural Biology* 21, 412-418.

Sormanni, P., Camilloni, C., Fariselli, P., and Vendruscolo, M. (2015). The s2D Method: Simultaneous Sequence-Based Prediction of the Statistical Populations of Ordered and Disordered Regions in Proteins. *Journal of Molecular Biology* 427, 982-996.

Stärk, H., Dallago, C., Heinzinger, M., and Rost, B. (2021). Light attention predicts protein location from the language of life. *Bioinformatics Advances* 1.

Steinegger, M., Mirdita, M., and Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods* 16, 603-606.

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 35, 1026-1028.

Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications* 9, 2542.

Suzek, B.E., Wang, Y., Huang, H., Mcgarvey, P.B., Wu, C.H., and Consortium, U. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31, 926-932.

Tantos, A., Friedrich, P., and Tompa, P. (2009). Cold stability of intrinsically disordered proteins. *FEBS Lett* 583, 465-469.

The Uniprot, C., Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., Bye-a-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L.G., Garmiri, P., Georghiou, G.,

Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Jokinen, P., Joshi, V., Jyothi, D., Lock, A., Lopez, R., Luciani, A., Luo, J., Lussi, Y., Macdougall, A., Madeira, F., Mahmoudy, M., Menchi, M., Mishra, A., Moulang, K., Nightingale, A., Oliveira, C.S., Pundir, S., Qi, G., Raj, S., Rice, D., Lopez, M.R., Saidi, R., Sampson, J., Sawford, T., Speretta, E., Turner, E., Tyagi, N., Vasudev, P., Volynkin, V., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.-C., Bolleman, J., Boutet, E., Breuza, L., Casals-Casas, C., De Castro, E., Echioukh, K.C., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Keller, G., Kerhornou, A., Lara, V., Le Mercier, P., Lieberherr, D., Lombardot, T., Martin, X., et al. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* 49, D480-D489.

Tompa, P., Dosztanyi, Z., and Simon, I. (2006). Prevalent structural disorder in E. coli and S. cerevisiae proteomes. *J Proteome Res* 5, 1996-2000.

Tompa, P., Prilusky, J., Silman, I., and Sussman, J.L. (2008). Structural disorder serves as a weak signal for intracellular protein degradation. *Proteins* 71, 903-909.

Tompa, P., Szasz, C., and Buday, L. (2005). Structural disorder throws new light on moonlighting. *Trends Biochem Sci* 30, 484-489.

Uversky, V.N., Oldfield, C.J., Midic, U., Xie, H., Xue, B., Vucetic, S., Iakoucheva, L.M., Obradovic, Z., and Dunker, A.K. (2009). Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* 10 Suppl 1, S7.

Van Der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 2579-2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is all you need", in: *Proceedings of the 31st International Conference on Neural Information Processing Systems.* (Long Beach, California, USA: Curran Associates Inc.).

Vicedo, E., Gasik, Z., Dong, Y.-A., Goldberg, T., and Rost, B. (2015a). Protein disorder reduced in Saccharomyces cerevisiae to survive heat shock. *F1000Res* 4, 1222.

Vicedo, E., Schlessinger, A., and Rost, B. (2015b). Environmental Pressure May Change the Composition Protein Disorder in Prokaryotes. *PLoS One* 10, e0133990.

Vucetic, S., Brown, C.J., Dunker, A.K., and Obradovic, Z. (2003). Flavors of protein disorder. *Proteins: Structure, Function, and Bioinformatics* 52, 573-584.

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.

Wang, S., Ma, J., and Xu, J. (2016). AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics (Oxford, England)* 32, i672-i679.

Ward, J.J., Sodhi, J.S., Mcguffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology* 337, 635-645.

Wilson, C.J., Choy, W.-Y., and Karttunen, M. (2021). "AlphaFold2: A role for disordered protein prediction?". Bioinformatics).

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 37-52.

Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology* 293, 321-331.

Wu, Z., Johnston, K.E., Arnold, F.H., and Yang, K.K. (2021). Protein sequence design with deep generative models. *Current Opinion in Chemical Biology* 65, 18-27.

Yang, Z.R., Thomson, R., Mcneil, P., and Esnouf, R.M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21, 3369-3376.