frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 10 October 2022
DOI 10.3389/fbinf.2022.1019597

OPEN ACCESS

EDITED BY
Rachel Kolodny,
University of Haifa, Israel

REVIEWED BY
Castrense Savojardo,
University of Bologna, Italy
Jianzhao Gao,
Nankai University, China

*CORRESPONDENCE
Michael Heinzinger,
mheinzinger@rostlab.org

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Protein Bioinformatics,
a section of the journal
Frontiers in Bioinformatics

RECEIVED 15 August 2022
ACCEPTED 20 September 2022
PUBLISHED 10 October 2022

CITATION
Ilzhöfer D, Heinzinger M and Rost B (2022), SETH predicts nuances of residue disorder from protein embeddings.
Front. Bioinform. 2:1019597.
doi: 10.3389/fbinf.2022.1019597

SETH predicts nuances of residue disorder from protein embeddings

Dagmar Ilzhöfer[1†], Michael Heinzinger[1,2*†] and Burkhard Rost[1,3,4]

[1]Faculty of Informatics, TUM (Technical University of Munich), 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, Germany, [4]TUM School of Life Sciences Weihenstephan (WZW), TUM (Technical University of Munich), Freising, Germany

Predictions for millions of protein three-dimensional structures are only a few clicks away since the release of *AlphaFold2* results for UniProt. 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 three recent disorder measures of *AlphaFold2* predictions (pLDDT, "experimentally resolved" prediction and "relative solvent accessibility") correlated to some extent with IDRs. However, expert methods predict IDRs more reliably by combining complex machine learning models with expert-crafted input features and evolutionary information from multiple sequence alignments (MSAs). MSAs are not always available, especially for IDRs, and are computationally expensive to generate, limiting the scalability of the associated tools. Here, we present the novel method SETH that predicts residue disorder from embeddings generated by the protein Language Model ProtT5, which explicitly only uses single sequences as input. Thereby, our method, relying on a relatively shallow convolutional neural network, outperformed much more complex solutions while being much faster, allowing to create predictions for the human proteome in about 1 hour on a consumer-grade PC with one NVIDIA GeForce RTX 3060. Trained on a continuous disorder scale (CheZOD scores), our method captured subtle variations in disorder, thereby providing important information beyond

Abbreviations: 3D, three-dimensional, i.e., coordinates of all atoms/residues in a protein; AI, artificial intelligence; *AlphaFold2*: AI-based method reliably predicting protein 3D structure from MSAs (Jumper et al., 2021); ANN, an artificial feed-forward neural network; AUC, area under the receiver operating characteristic curve; CheZOD scores, chemical shift Z-scores from NMR (Nielsen and Mulder, 2019); CI, confidence interval, here typically used as the 95% CI implying an interval between ±1.96*Standard Error; CNN, convolutional neural network; ColabFold, protocol for fast execution of *AlphaFold2* (Mirdita et al., 2022); GPU, graphical processing unit; IDP, intrinsically disordered proteins (Dunker et al., 2013); IDR, intrinsically disordered regions (Dunker et al., 2013); LinReg, a linear regression model; LogReg, a logistic regression model; MSA, multiple sequence alignment; NMR, nuclear magnetic resonance; PCA, principle component analysis; PIDE, percentage pairwise sequence identity; pLDDT, predicted local distance difference test from *AlphaFold2* (Jumper et al., 2021); pLM, protein Language Model; ProtT5, particular pLM (Elnaggar et al., 2021); RSA, relative solvent accessible surface area of a residue; SETH, a CNN for continuous disorder prediction (our best model); SOTA, state-of-the-art; t-SNE, t-distributed stochastic neighbor embedding; ρ, Spearman correlation coefficient.

the binary classification of most methods. High performance paired with speed revealed that SETH's nuanced disorder predictions for entire proteomes capture aspects of the evolution of organisms. Additionally, SETH could also be used to filter out regions or proteins with probable low-quality *AlphaFold2* 3D structures to prioritize running the compute-intensive predictions for large data sets. SETH is freely publicly available at: https://github.com/Rostlab/SETH.

KEYWORDS

protein disorder, residue disorder, IDP, IDR, protein structure prediction, *AlphaFold2*, protein language model

Introduction

IDRs crucial for life

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 [IDPs (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). The difference between long IDRs 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). Additionally, 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).

CheZOD scores best characterize IDRs experimentally

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).

Many prediction methods available

The limited scalability of labor-intensive and expensive wet-lab experiments has spawned many computational tools predicting IDRs, 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), PreDisorder (Deng et al., 2009), MetaDisorder-MD (Schlessinger et al., 2009), ESpritz (Walsh et al., 2012), MetaDisorder (Kozlowski and Bujnicki, 2012), AUCpreD (Wang et al., 2016), SPOT-Disorder (Hanson et al., 2016), SPOT-Disorder-Single (Hanson et al., 2018), SPOT-Disorder2 (Hanson et al., 2019), rawMSA (Mirabello and Wallner, 2019), ODiNPred (Dass et al., 2020) and flDPnn (Hu et al., 2021). 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 out-performed methods not using MSAs (Nielsen and Mulder, 2019; Dass et al., 2020). However, using MSAs slows down inference and performs worse for proteins in small families. This 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).

Besides these methods directly predicting disorder, *AlphaFold2* (Jumper et al., 2021), Nature's method of the year 2021 (Marx, 2022), which provided a leap in the quality of protein structure predictions from MSAs and increases the width of structural coverage (Bordin et al., 2022), also provides measures indicative of IDRs. One of these, the pLDDT (predicted local distance difference test), estimates the performance of *AlphaFold2* depending on prediction strength, i.e., it measures prediction reliability as introduced for secondary structure prediction (Rost and Sander, 1993). However, instead of measuring it from the class output, *AlphaFold2* uses different objective functions and predicts its own reliability. The pLDDT distinguishes formidably well between trustworthy and less reliable predictions (Jumper et al., 2021). Additionally, low values for pLDDT have been suggested to predict IDRs rather accurately (Akdel et al., 2021; Wilson et al., 2021; Piovesan et al., 2022) or to predict non-existing proteins (Monzon et al., 2022). Furthermore, the "experimentally resolved" prediction of *AlphaFold2* should also contain information on disorder, since missing coordinates in experimentally recorded structures were an established definition of disorder (Dunker et al., 1998; Monastyrskyy et al., 2014). Lastly, the relative solvent accessible surface area of a residue [RSA (Connolly, 1983; Rost and Sander, 1994)] and its window average, calculated for *AlphaFold2* structure predictions, were also reported to be disorder predictors (Akdel et al., 2021; Piovesan et al., 2022; Redl et al., 2022).

Here, we bypassed the problem of generating MSAs for IDRs, by using embeddings from pre-trained protein language models (pLMs). Inspired by recent leaps in Natural Language Processing (NLP), pLMs learn to predict masked amino acids (tokens) given their surrounding protein sequence (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, amino acids correspond to words/tokens in NLP, 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 more than two billion sequences (Steinegger et al., 2019). The information learned by the pLM during so-called (pre-) training can be retrieved and transferred afterwards (transfer learning), 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 capture some aspects of the language of life as written in protein sequences (Alley et al., 2019; Heinzinger et al., 2019; Ofer et al., 2021; Rives 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; Littmann et al., 2021a; Littmann et al., 2021b; Littmann et al., 2021c; Elnaggar et al., 2021; Heinzinger et al., 2021; Marquet et al., 2021; Rives et al., 2021).

First, we compared to which extent 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)] could predict the degree of disorder of a residue as defined by CheZOD scores. Toward that end, we fit a minimal machine learning model (linear regression) on each of the five pLM embeddings. No pLM was fine-tuned in any way. The best performing embeddings served as input to partly a little 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 pLMs, which was 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 the final method SETH to predict disorder for the entire Swiss-Prot with over 566,000 proteins (The UniProt Consortium et al., 2021) in approximately 7 h on a machine with one RTX A6000 GPU with 48 GB vRAM.

Since recent work showed that *AlphaFold2*'s (smoothed) pLDDT and (smoothed) RSA can be used to predict disorder (Akdel et al., 2021; Wilson et al., 2021; Piovesan et al., 2022; Redl et al., 2022), we tested *AlphaFold2* on CheZOD scores (following the advice of John Jumper, we also analyzed "experimentally resolved" predictions). Furthermore, we investigated the agreement between the disorder predictions of our best method and the pLDDT for 17 organisms, to establish SETH as a speed-up pre-filter for *AlphaFold2*. Lastly, we visually analyzed whether the predicted disorder spectrum carried any information about the evolution of 37 organisms.

Methods

Data sets

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; 1,325 proteins) and testing (file name CheZOD in GitHub; 117 proteins). Each dataset contains protein sequences and CheZOD scores for each residue. The CheZOD 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 (CheZOD1174 and CheZOD117)

To avoid overestimating performance through pairs of proteins with too similar sequences between training and testing sets, we constructed non-redundant subsets. Firstly, we built profiles (position specific scoring matrices; PSSMs) from multiple sequence alignments (MSAs) for proteins in the test set, obtained through three iterations with *MMSeqs2* [(Steinegger and Söding, 2017); using default parameters, except for "--num-iterations 3", an established number of iterations, also applied in ColabFold (Mirdita et al., 2022) and enabling sensitive but still fast sequence searches (Steinegger and Söding, 2017)] against proteins in the training set. Next, any protein in the training set with >20% PIDE (percentage pairwise sequence identity) to any test set profile using bi-directional coverage [with default coverage threshold of 80%, focusing on joining proteins with similar domain composition (Hauser et al., 2016)] was removed using *MMSeqs2* high-sensitivity (--s 7.5) search. The value PIDE<20% was, for simplicity, concluded from an earlier analysis of the reach of homology-based inference for the structural similarity of protein pairs (Rost, 1999). The training set had been constructed such that all protein pairs had <50% PIDE (Dass et al., 2020), and we did not reduce the redundancy within the training set any further. Secondly, we removed all residues without valid CheZOD scores [indicated by CheZOD scores≥900; for all models apart from SETH, they were removed after embedding generation, while for SETH (CNN) they were removed before, to enable undisturbed passing of information from neighboring residues]. The resulting training set (dubbed *CheZOD1174*) contained 1,174 proteins with a total of 132,545 residues (at 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 length 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, we altered the training data published and used for ODiNPred, to reduce the overlap between training and testing.

Dataset distributions

After preparing the data, we analyzed the distributions of the CheZOD scores for both *CheZOD117* and *CheZOD1174* (Supplementary Figure S1). The CheZOD scores in these sets ranged from -5.6 to 16.2. Nielsen and Mulder had previously established a threshold of eight to differentiate between disorder (CheZOD score≤8) and order (CheZOD score>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 the training set *CheZOD1174* (72% ordered), disordered residues were most prevalent in the test set *CheZOD117* (31% ordered). As artificial intelligence (AI) always optimizes for similar distributions in train and test, the train-test set discrepancy provided an additional safeguard against over-estimating performance.

Input embeddings

Five pLMs

Protein sequences from both sets (*CheZOD117*, *CheZOD1174*) were encoded as distributed vector representations (embeddings) using 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 Consortium 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 BFD, the Big Fantastic Database (Steinegger and Söding, 2018; Steinegger et al., 2019), with over 2.1 billion protein sequences. 3) ESM-1b (Rives et al., 2021), 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 BFD 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 one 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 = 1,024 for SeqVec/ProtBERT/ProtT5; N = 1,280 for ESM-1b; N = 6,165 for ProSE).

Choosing embeddings best suited for IDR prediction

To find the most informative pLM embeddings for predicting 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 to predict continuous CheZOD scores. To simplify the comparison and "triangulation" of our results, we also compared these five embedding-based models to inputting the standard one-hot encodings (i.e., 20 instead of 1,024/1280/6,165 input units per residue). One-hot encodings represent each residue/sequence position by a 20-dimensional vector, for the 20 standard amino acids essentially contained in all proteins. Each position in the vector corresponds to one amino acid, i.e., the elements of the vector are binary: one for the position in the vector corresponding to the encoded amino acid, zero otherwise. The special case "X" (unknown amino acid) was encoded in a 20-dimensional vector containing only 0s. The linear regressions were implemented with the *LinearRegression* module of scikit-learn (Pedregosa et al., 2011) with all parameters left at default values. We evaluated the models on the remaining 10% of *CheZOD1174* using the Spearman correlation coefficient (ρ; Eq. 2) and the AUC (area under the receiver operating characteristic curve; Eq. 3; see Methods *Evaluation*).

Unsupervised embedding analysis

Lastly, we analyzed the ProtT5 embeddings of *CheZOD117* in more detail by creating a t-distributed stochastic neighbor embedding [t-SNE (van der Maaten and Hinton, 2008)] using scikit-learn (Pedregosa et al., 2011). 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 the global data structure (Kobak and Berens, 2019) in order to identify putative clusters of order or disorder (defaults for all other parameters).

New disorder prediction methods

We optimized four 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 to gain speed and avoid over-fitting. Three of our models were trained on regression (LinReg, ANN and SETH), while LogReg was trained on discriminating disordered from ordered residues (binary classification: disorder: CheZOD score≤8, order: CheZOD score>8 (Nielsen and Mulder, 2016)).

SETH was implemented in PyTorch (Paszke et al., 2019) using *Conv2d* for the convolutional layers, MSELoss as loss function and Adam as optimizer (learning rate of 0.001), activating amsgrad (Reddi et al., 2018). Additionally, we padded to receive one output per residue and set all random seeds to 42 for reproducibility. Lastly, we randomly split *CheZOD1174* into training (90% of proteins: optimize weights) and validation (10%: for early-stopping after 10 epochs without improvement, hyper-parameter optimization: of the best performing models, we chose that with the most constraints (Supplementary Figure S3, red bar), resulting in a kernel size of (5,1), 28 output channels of the first convolutional layer, the activation function Tanh between the two convolutional layers and the weight decay parameter of 0.001 in the optimizer). Details for LinReg, ANN and LogReg are in Supplementary Material S1.1 and Supplementary Figure S2.

AlphaFold2

AlphaFold2 (Jumper et al., 2021) predicts a reliability for each residue prediction, namely, the pLDDT. This score and its running average over a window of consecutive residues have been claimed to predict disorder (Akdel et al., 2021; Wilson et al., 2021; Piovesan et al., 2022; Redl et al., 2022). Another objective function predicted by AlphaFold2, namely, the "experimentally resolved" prediction (Jumper et al., 2021) is likely also informative, as missing coordinates in experimental structures have been used to define disorder (Dunker et al., 1998; Monastyrskyy et al., 2014). To analyze these *AlphaFold2* predictions against CheZOD scores, we applied *ColabFold* (Mirdita et al., 2022) to predict 3D structures for all proteins in *CheZOD117*. *ColabFold* speeds up *AlphaFold2* predictions 40-60x mostly 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). We used neither templates nor Amber force-field relaxation (Hornak et al., 2006), as those do not significantly improve performance (Jumper et al., 2021; Mirdita et al., 2022) although increasing runtime manifold (especially the Amber relaxation). As *ColabFold* currently does not support outputting the "experimentally resolved" head, we added this feature by averaging over the sigmoid of the raw "experimentally resolved" logits output of *AlphaFold2* for each atom in a residue. After having generated the pLDDT values and the "experimentally resolved" predictions, we additionally calculated the smoothed pLDDT for each residue, using a sliding window of 21 consecutive residues following previous findings (Akdel et al., 2021). While sliding the window over the sequence, the center residue of the window was always assigned the mean of all values within the window (instead of padding, windows closer than 10 residues to the N- and C-terminus were shrunk asymmetrically, e.g., for the N-terminal position i (i = 1,. . .,10, starting at i = 1 at the N-terminus): averaging over (i-1) positions left of i).

It has also been reported that the window-averaged RSA calculated from AlphaFold2's 3D predictions correlates with IDRs (Akdel et al., 2021; Piovesan et al., 2022; Redl et al., 2022). Consequently, we also added this measure to our evaluation. With the pipeline provided alongside one of the analyses reporting the RSA as a disorder predictor (Piovesan et al., 2022), we calculated the RSA for the residues of the *CheZOD117* set, leaving all parameters at default. Then we smoothed the RSA by averaging over 25 consecutive residues as suggested elsewhere (Piovesan et al., 2022). While the use of the "experimentally resolved" predictions is new here (thanks to John Jumper for the recommendation), all other ways of processing *AlphaFold2* predictions to predict disorder were taken from other work.

Evaluation

We followed a previous analysis in evaluating our performance (same evaluation measures and test set) for easy comparability (Nielsen and Mulder, 2019). This allowed a direct comparison to (alphabetical list): AUCPred with and without evolution (Wang et al., 2016), DisEMBL (Linding et al., 2003), DISOPRED2 (Ward et al., 2004), DISOPRED3 (Jones and Cozzetto, 2015), DISpro (Cheng et al., 2005), DynaMine (Cilia et al., 2014), DISPROT/VSL2b (Vucetic et al., 2005), ESpritz (Walsh et al., 2012), GlobPlot (Linding et al., 2003), IUPred (Dosztanyi et al., 2005), MetaDisorder (Kozlowski and Bujnicki, 2012), MFDp2 (Mizianty et al., 2013), PrDOS (Ishida and Kinoshita, 2007), RONN (Yang et al., 2005), s2D (Sormanni et al., 2015), SPOT-Disorder (Hanson et al., 2016). We added results for flDPnn (Hu et al., 2021) and ODiNPred (Dass et al., 2020) using the publicly available web-servers. SPOT-Disorder2 (Hanson et al., 2019) predictions were custom-generated by the program's developers for all but one protein in test set CheZOD117 (10010: failed run).

We estimated the Spearman correlation, ρ, and its 95% confidence interval (CI) over n = 1,000 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 *i*th value in the ground truth CheZOD scores and v_i the rank of the *i*th value in the predicted CheZOD scores (or the rank of the respective predictive values for LogReg and *AlphaFold2*) of the method, the ρ was calculated with Eq. (1). The final ρ was derived from averaging over those 1,000 values and the 95% CI was estimated by computing the standard deviation of the ρ over the sampled sets and multiplying it by 1.96. The standard deviation was calculated with Eq. 2, where x_i is the ρ of an individual bootstrap set and ⟨x⟩ is the average ρ over all bootstrap sets.

$$\rho\ (Spearman\, correlation) = \frac{\sum_{i=1}^{m}\left[\left(u_i - \frac{1}{m}\sum_{j=1}^{m}u_j\right)^{*}\left(v_i - \frac{1}{m}\sum_{j=1}^{m}v_j\right)\right]}{\sqrt{\sum_{i=1}^{m}\left(u_i - \frac{1}{m}\sum_{j=1}^{m}u_j\right)^{2} {}^{*}\sum_{i=1}^{m}\left(v_i - \frac{1}{m}\sum_{j=1}^{m}v_j\right)^{2}}} \tag{1}$$

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \langle x\rangle\right)^{2}}{n}} \tag{2}$$

Furthermore, the AUC and its 95% CI were estimated for each model evaluated here, again, by applying the same bootstrapping procedure. As the AUC requires binarized ground truth class labels, continuous CheZOD scores were binarized using the threshold of eight (disorder CheZOD score≤8 and order CheZOD score>8 (Nielsen and Mulder, 2016)) for the calculation of the AUC (Eq. (3); scikit-learn implementation). In Eq. 3, I[.] is the indicator function, $m^{+/-}$ are the number of ordered/disordered samples in the test set (classifying the samples according to the ground truth class label) and $y_i^{+/-}$ is the *i*th predicted value in the ordered/disordered samples.

$$\text{AUC} = \frac{\sum_{j=1}^{m^-}\sum_{i=1}^{m^+}\left(I\left[y_i^{+} > y_j^{-}\right]\right)}{m^{+*}m^{-}} \tag{3}$$

Lastly, we plotted the receiver operating characteristic curve for our models (SETH, LinReg/LinReg1D, ANN and LogReg), as well as for *AlphaFold2*'s pLDDT (Supplementary Figure S5).

Additional tests

Runtime

We analyzed the runtime for the best method introduced here (SETH), by clocking the predictions for the human proteome (20,352 proteins) and the Swiss-Prot database [566,969 proteins (The UniProt Consortium et al., 2021)]. This evaluation was performed on a machine with two AMD EPYC™ ROME 7352 CPUs at 2.30 GHz each with 24/48 cores, a 256 GB RAM (16 × 16 GB) DDR4-3200 MHz ECC, one RTX A6000 GPU with 48GB RAM, a 278 GB SSD scratch disk and a 7.3 TB HDD. However, the final model constituting SETH can also easily be deployed on any machine holding a GPU with ≥8 GB RAM at some cost in speed, allowing to run SETH, e.g., in *Google Colab*. To reflect this, we also benchmarked the speed for running the entire human proteome on a smaller GPU (single NVIDIA GeForce RTX 3060 with 12 GB vRAM). Lastly, we benchmarked the speed on our test set *CheZOD117* on an AMD Ryzen 5 5500U CPU, to reflect that SETH can even efficiently be run without a GPU for small sets. All values for runtime included all steps required: 1) load ProtT5, 2) load SETH model checkpoint, 3) read sequences from FASTA files, 4) create embeddings, 5) create predictions and 6) write results into a file.

Comparison: CheZOD score predictions and pLDDT in 17 organisms

From the AlphaFold database with 3D predictions (Jumper et al., 2021), we downloaded the available files ending in "F1-model-v2. pdb" for all proteins listed in UniProt (The UniProt Consortium et al., 2021) for 19 organisms (Supplementary Table S2). A few files (0.3% of proteins) appeared with seemingly corrupted format (no separator between some values) and were removed. For all others, we extracted the pLDDT values.

For three organisms (*Leishmania infantum*, *Schistosoma mansoni* and *Plasmodium falciparum*) we predicted disorder with SETH using the sequences provided in UniProt (The UniProt Consortium et al., 2021); for the remaining 16 organisms (Supplementary Table S2) we used the sequences present in Swiss-Prot, due to already having generated this data (see Runtime). Due to GPU resources, we did not predict disorder for proteins with >9,000 residues (0% of *Leishmania infantum* + *Schistosoma mansoni*, 0.7% - 40 proteins - of *Plasmodium falciparum*, 0.004% - 25 proteins - of Swiss-Prot). None of the proteins for the CheZOD sets (CheZOD117, CheZOD1174) were that long (for obvious reasons related to the length-limitation of NMR).

To compare disorder predictions and pLDDT, only the subset of the data where both *AlphaFold2* and disorder predictions were available were used. The resulting set contained 17 of the above downloaded 19 organisms (Supplementary Table S2; two organisms: no overlap in the predictions available for *AlphaFold2* and SETH) with 105,881 proteins containing a total of 47M residues. We referred to this data set as the *17-ORGANISM-set*.

Spectrum of predicted CheZOD score distributions for entire organisms

The spectra of predicted subcellular location reveal aspects pertaining to the evolution of species (Marot-Lassauzaie et al., 2021). Consequently, we tried the same concept on predicted CheZOD scores for Swiss-Prot. For technical reasons (GPU memory), we excluded proteins longer than 9,000 residues from our analysis. In the entire Swiss-Prot, 0.004% of the proteins reached this length and were excluded. For the other 99.996%, we first converted all predicted CheZOD score distributions (consisting of all disorder predictions of all residues within one organism) of all Swiss-Prot organisms into vectors by counting CheZOD scores in eight bins (-15, -11.125, -7.25, -3.375, 0.5, 4.375, 8.25, 12.125, 16). After normalization (dividing raw counts by all residues in the organism), we PCA-plotted 37 organisms of Swiss-Prot with at least 1,500 proteins [(Wold et al., 1987); to keep clarity in the plot, some organisms with at least 1,500 proteins were neglected), using the standard implementation of R (prcomp (R Core Team, 2021)].

Results

Success of minimalist: Single sequence, simple model

While state-of-the-art (SOTA) methods usually rely on MSA input to predict IDRs, the methods introduced here use pLMs to encode single protein sequences as embeddings that served as the sole input feature for any prediction. To find the most informative pLM for IDRs, 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). Following the recent assessment of 26 methods (Nielsen and Mulder, 2019), we calculated the Spearman correlation coefficient ρ between true and predicted CheZOD scores and the AUC for 10% of *CheZOD1174*, not used in training, to evaluate the models.

Embeddings from all pLMs outperformed the random baseline and the one-hot encodings, both for the correlation (Figure 1B; ρ) and the binary projection of CheZOD scores (Figure 1A; AUC). The simplest pLM-type included here, namely *SeqVec*, performed consistently and statistically significantly worse than all other pLMs (Figure 1). The other four embeddings (ProSE, ESM-1b, ProtT5, ProtBERT) did not differ to a statistically significant extent, 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.

ProtT5 captured 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, t-SNE projected the 1024-dimensional embeddings onto two dimensions (Figure 2). This suggested some level of separation between ordered (red) and disordered (blue) residues (Figure 2A: red colors oriented toward the center in each cluster), indicating that even raw ProtT5 embeddings

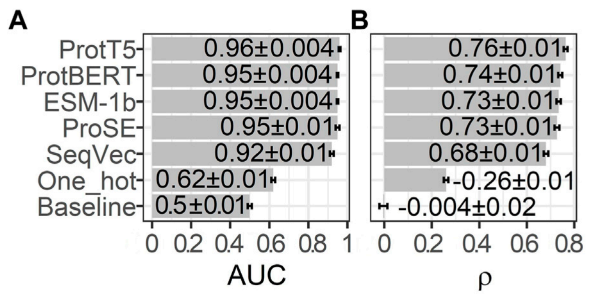

FIGURE 1
Performance estimates for training on 90% of *CheZOD1174* (Dass et al., 2020) and testing on the remaining 10% using linear regressions fed by 20-dimensional one-hot encodings or raw embeddings (without further optimization) from five protein language models (pLMs): ProtT5 (Elnaggar et al., 2021), ProtBERT (Elnaggar et al., 2021), ESM-1b (Rives et al., 2021), ProSE (Bepler and Berger, 2021), SeqVec (Heinzinger et al., 2019). The seventh row displays the performance of the baseline/random model computed on 1024-dimensional embeddings sampled randomly from a standard normal distribution. **(A)** 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; Eq. 3), while **(B)** depicts the Spearman correlation coefficient (ρ; Eq. 2), calculated using the observed and predicted CheZOD scores. The errors mark the 95% confidence intervals approximated by multiplying 1.96 with the bootstrap standard deviation (Methods).

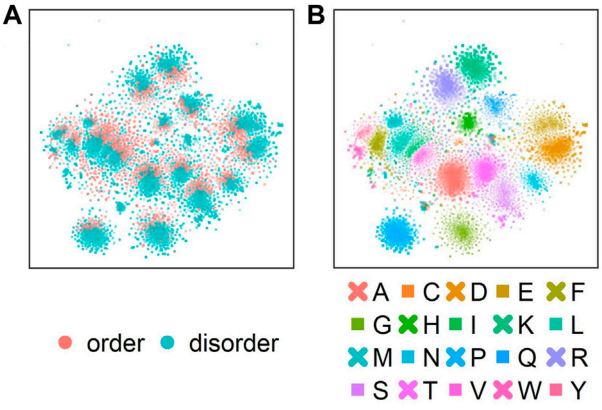

FIGURE 2
t-SNE dimensionality reduction (van der Maaten and Hinton, 2008) performed on the 1024-dimensional ProtT5 (Elnaggar et al., 2021) residue-level embeddings extracted from the last attention layer of ProtT5 for all sequences in test set *CheZOD117* [13,069 residues; (Dass et al., 2020)]. **(A)** shows the embeddings colored by order (CheZOD score>8; red) and disorder [CheZOD score≤8, blue; (Nielsen and Mulder, 2016)]. **(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).

already captured some aspects of disorder without seeing 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. Instead, the primary 20 clusters corresponded to the 20 amino acids (Figure 2B).

SETH (CNN) outperformed other supervised models

Next, we trained four AI models, inputting ProtT5 embeddings: three predicted continuous CheZOD scores (LinReg, ANN, SETH), one predicted binary disorder (LogReg). We could add the performance of our methods to a recent method comparison (Nielsen and Mulder, 2019) since we used the same performance metrics and test set (*CheZOD117;* Figure 3). We also added the ODiNPred web application (Dass et al., 2020), the flDPnn webserver (Hu et al., 2021) and the performance of the new method ADOPT ESM-1b (Redl et al., 2022), which also uses pLM embeddings. Additionally, the program's developers ran SPOT-Disorder2 (Hanson et al., 2019) for us, which, however, failed to run for one test set protein. The performance on the remaining 116 proteins was: ρ = 0.63 ± 0.01 and AUC = 0.88 ± 0.01. When considering the mean ρ (Figure 3A), our methods SETH and ANN numerically outperformed all others, both those not using MSAs (below dashed line in Figure 3), and those using MSAs (above dashed line in Figure 3). When requiring a statistically significant difference at the 95% CI (±1.96 standard errors) for the ρ, our methods (SETH, ANN, LinReg and LogReg) significantly outperformed all others, except for ODiNPred and ADOPT ESM-1b. When evaluating the performance based on the mean AUC, SETH and the simplistic LinReg outperformed all other evaluated methods. Due to the already high AUC levels of many methods, the absolute improvement of our models (SETH, ANN, LinReg and LogReg) to SOTA methods in terms of AUC was often not statistically significant.

The differences between the models introduced here (LogReg, LinReg, ANN and SETH) were not statistically significant (neither for AUC nor for ρ). However, SETH had the highest mean ρ and, together with LinReg, the highest mean AUC. For a more detailed analysis, 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 (Supplementary Figure S6). SETH, ANN and LinReg agreed well with the ground truth. However, 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. observed CheZOD

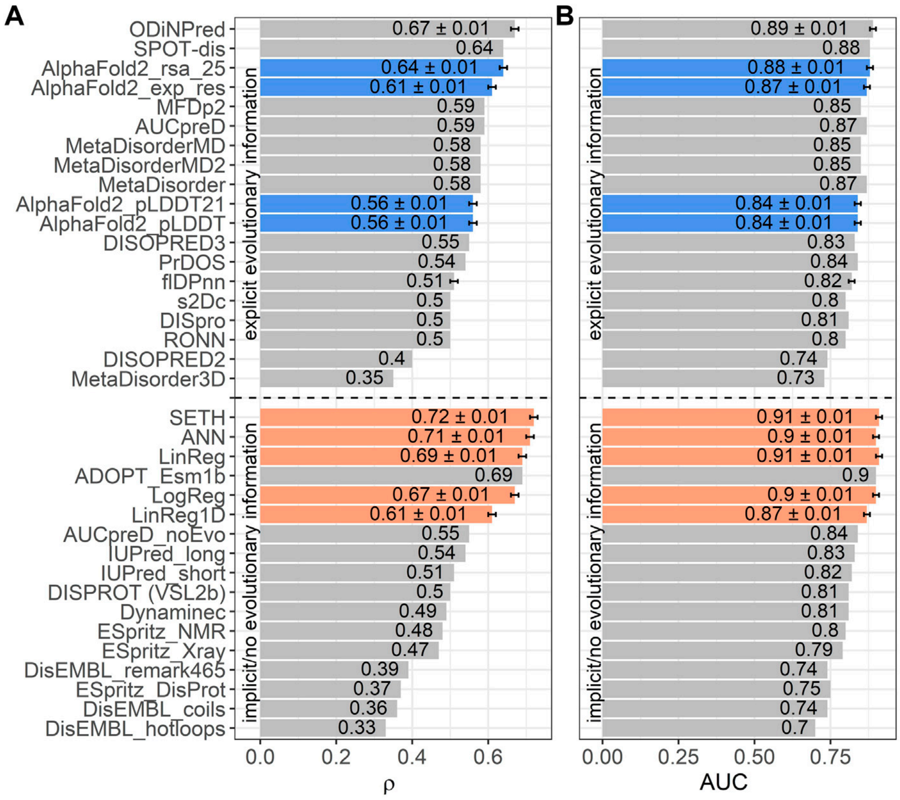

FIGURE 3
Data: set *CheZOD117* (Dass et al., 2020). Performances of all methods introduced here (SETH, ANN, LinReg, LogReg, LinReg1D) in orange, the ODiNPred web application in grey (Dass et al., 2020), ADOPT ESM-1b in grey (Redl et al., 2022), the fIDPnn server in grey (Hu et al., 2021) and four disorder measures derived from *AlphaFold2* (Jumper et al., 2021) in blue [these are: AlphaFold2_*pLDDT*, AlphaFold2_pLDDT21: smoothed over 21-consecutive residues (Akdel et al., 2021), AlphaFold2_exp_res: experimentally resolved prediction, AlphaFold2_rsa_25: running average over relative solvent accessibility averaged over 25 consecutive residues (Piovesan et al., 2022)]. All other performances were taken from the previous comparison (Nielsen and Mulder, 2019) using the same test set (see Methods *Evaluation*). While three of our models (SETH, ANN, LinReg/LinReg1D), ADOPT ESM-1b and ODiNPred were trained on continuous chemical shift Z-scores (CheZOD scores), the logistic regression, LogReg, was trained on a binary classification of order/disorder (CheZOD score>8/≤8). ODiNPred and ADOPT ESM-1b used more proteins for training than our models. The horizontal dotted line separates models using MSAs (above line) from single sequence-based methods (below line). Error bars mark the 95% confidence interval, approximated by bootstrapping for our methods, *AlphaFold2*, the ODiNPred web application and the fIDPnn server (Methods). Panel (A): Performance measured with the spearman correlation coefficient (ρ; Eq. 2) between the ground truth and the prediction. Panel (B): Performance measured with the area under the receiver operating characteristic curve (AUC; Eq. 3) after the binary projection of the ground truth CheZOD scores [order: CheZOD score>8, disorder: CheZOD score≤8; (Nielsen and Mulder, 2016)].

scores fell into two flat clusters at 0 and 1, confirming that LogReg tended to predict extreme values optimal for classification. The removal of short disordered residues (i.e. less than 30 consecutive residues with observed CheZOD scores≤8) did not change the Spearman correlation significantly (Supplementary Figure S7).

Shortcomings of SETH

For SETH, our best model (outperforming all others in ρ and AUC; Figure 3), we added another analysis classifying each residue in *CheZOD117* into one of three classes according to the observed CheZOD scores: ordered (CheZOD score>8), long disorder (residues in a disorder (CheZOD score≤8) stretch with≥30 residues) and short disorder (disordered stretches with<30 residues). Firstly, SETH clearly missed short disorder (Figure 4: predicted values for this class were approximately uniformly distributed in (0,15), with a ρ of only 0.41 ± 0.04). Secondly, SETH overestimated order (Figure 4 and also Supplementary Figure S6), as there was a shift of the distributions of ordered and long disordered residues to the right from the observed to the predicted scores. Thirdly, SETH predicted several residues as ordered, for which the ground truth CheZOD scores suggested long consecutive regions of disorder (Figure 4B). For a subset of proteins, for which at least one-third of all residues were in long IDRs but

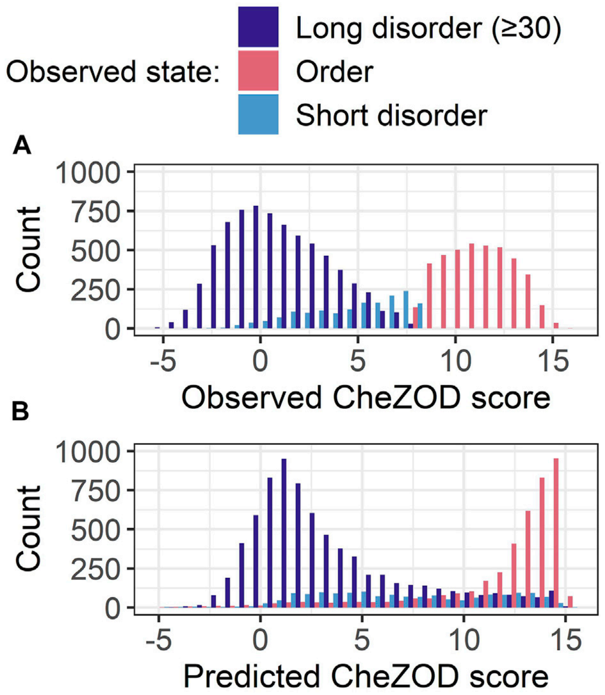

FIGURE 4
Data: set *CheZOD117* (Dass et al., 2020). All residues in *CheZOD117* were classified into one of three classes: ordered (chemical shift Z-score (CheZOD score) > 8), short disorder and long disorder (disordered residues (CheZOD score≤8) in a disordered region≥30), using the ground truth labels. **(A)** Distribution of the observed CheZOD scores (i.e., the ground truth labels) in the three classes. **(B)** Distribution of SETH's predicted CheZOD scores in the three classes. SETH is a CNN trained on continuous chemical shift Z-scores and outperformed all state-of-the-art methods evaluated, as well as all our other methods.

SETH predicted order, *AlphaFold2*'s pLDDT largely supported our predictions of order (Supplementary Figure S8). For two of these ten proteins, we found DisProt annotations (Quaglia et al., 2022), showing disorder to order transition regions (i.e., regions that can change from disorder to order, e.g., upon binding) overlapping with the regions of wrongly predicted order (Supplementary Figure S8). Lastly, SETH's predicted CheZOD scores<0 indicated long IDRs (only this class has high counts below 0, Figure 4). This suggested zero as a second more conservative threshold for classifying disorder, to filter out short linker regions falsely labeled as disorder.

SETH blazingly fast

Using SETH for analyzing proteins and proteomes requires top performance (Figure 3) and speed. 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 min. For Swiss-Prot [566,969 proteins (The UniProt Consortium et al., 2021)], it took approximately 7 hours. As a rule of thumb, SETH could predict disorder for approximately 10–20 proteins in 1 s, depending on the protein length. Even on smaller GPUs such as a single NVIDIA GeForce RTX 3060 with 12 GB vRAM, computing predictions for the human proteome still took only an hour. Lastly, even on an AMD Ryzen 5 5500U CPU, performing predictions for our test set *CheZOD117* (average protein length 112) only took 12 min, showing that for small sets a GPU is not even necessary.

One of 1,024 embedding dimension outperformed most methods (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 1,024; Supplementary Figure S4), we subsequently plotted the raw embedding values against the true CheZOD scores (Supplementary Figure S6E) 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 LinReg (ρ = 0.69; LinReg used all 1,024 dimensions in training), 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 overestimating residue order (no accumulation of high density above the diagonal; Supplementary Figure S6).

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 AUC = 0.91, Figure 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.

When comparing LinReg1D to the other methods evaluated in the large-scale comparison of disorder predictors (Nielsen and Mulder, 2019), ODiNPred and ADOPT ESM-1b, even this extremely reduced model outperformed all other methods not using MSAs apart from ADOPT ESM-1b and only fell short compared to the two best-performing methods using MSAs (SPOT-Disorder (Hanson et al., 2016) and ODiNPred), when looking at both the AUC and the ρ (Figure 3). However, compared to our other methods (SETH, LinReg, ANN, LogReg) LinReg1D performed significantly worse.

AlphaFold2 correlated less with CheZOD scores than top methods

AlphaFold2's (smoothed) predicted reliability pLDDT and its (smoothed) predicted RSA have recently been reported to capture some aspects of IDRs (Akdel et al., 2021; Wilson et al., 2021; Piovesan et al., 2022; Redl et al., 2022). However, the ρ between *AlphaFold2*'s (smoothed) pLDDT and CheZOD scores clearly neither reached the levels of the top expert solutions (SETH, LinReg, ANN, LogReg, LinReg1D, ODiNPred or ADOPT ESM-1b; Figure 3A) trained on CheZOD scores, nor that of many other methods using MSAs (Nielsen and Mulder, 2019). Looking at the correlation between pLDDT scores and CheZOD scores in more detail (Supplementary Figure S6G) revealed that disordered residues (CheZOD score≤8) were occasionally predicted with high confidence (pLDDT>80) explaining the rather low ρ. *AlphaFold2's* "experimentally resolved" prediction (AlphaFold2_exp_res, Figure 4) correlated better with CheZOD scores, reaching the top 10 methods. Even better was the smoothed RSA value (ρ = 0.64; AlphaFold2_rsa_25, Figure 4), although still falling behind the top expert solutions (SETH, LinReg, ANN, LogReg, ODiNPred or ADOPT ESM-1b).

SETH disorder predictions correlated with *AlphaFold2* pLDDT

We analyzed the fitness of SETH as a fast pre-filter to distinguish between proteins/regions with low and high mean pLDDTs of *AlphaFold2* (Figure 5). For proteins from 17 model organisms, SETH's predictions correlated well with the *AlphaFold2* pLDDT (ρ = 0.67; Figure 5A, per-organism details: Supplementary Figure S10). This trend remained after binarizing disorder using a CheZOD threshold of 8 (Figure 5B). If the goal were to predict the classification of all proteins into those with mean pLDDT≥70 (*wanted*) and pLDDT<70 (*unwanted*), depending on the threshold in the mean predicted CheZOD score (number on the curve in Figure 5C), this will result in different pairs of wanted proteins incorrectly missed (*y*-axis, Figure 5C) given the proteins correctly ignored (*x*-axis, Figure 5C). For instance, at a threshold of eight in the mean predicted CheZOD scores, a quarter of all proteins could be avoided at an error rate of only 5% (proteins missed with pLDDT≥70). The accuracy at this threshold was 0.86. This might be relevant to prioritize/filter data in large-scale *AlphaFold2* predictions.

More importantly, the comparison of the *AlphaFold2* pLDDT and SETH's predictions could also be used to find out more about the causes of lacking reliable *AlphaFold2* predictions. For instance, a lack of reliable *AlphaFold2* predictions was often due to disorder in proteins since low pLDDT values were mostly present for disordered residues (Figure 5B). However, providing Figure 5B at the organism level (Supplementary Figure S11) revealed that for some organisms, especially those with rather low mean pLDDT values (Supplementary Figure S9), SETH predicted many residues as ordered for which *AlphaFold2's* pLDDT was low. There were even cases, where nearly the entire protein was predicted to be ordered, but *AlphaFold2* could not predict any reliable 3D structure (Supplementary Figure S12).

Evolutionary information captured in CheZOD score distributions

Encouraged by the finding that the spectrum of predicted subcellular locations (in 10 classes) captures aspects of evolution (Marot-Lassauzaie et al., 2021), here, we converted the CheZOD score predictions for an entire organism into a single 8-dimensional vector containing the binned normalized counts of predicted CheZOD scores. A simple PCA (Wold et al., 1987) on the resulting vectors for 37 organisms revealed a clear connection from the micro-molecular level of per-residue predictions of CheZOD-disorder to the macro-molecular level of species evolution (Figure 6). Firstly, eukaryotes and prokaryotes (Bacteria + Archaea) were clearly separated. Secondly, even within these major groups, there appeared some relevant separation into phyla for the bacteria and into kingdoms for the eukaryotes. However, based on these limited samples, it also seemed like some groups could not be separated completely according to their disorder spectra, e.g., the fungi and the metazoa.

Discussion

We introduced SETH, a shallow CNN, for predicting the continuum of residue disorder defined by CheZOD scores (i.e., the difference between observed chemical shifts from NMR and computed random coil chemical shifts (Nielsen and Mulder, 2020)). SETH's exclusive input are embeddings from the pLM ProtT5 (Elnaggar et al., 2021). Using performance measures and data sets proposed in a recent analysis (Nielsen and Mulder, 2019), SETH outperformed three even simpler (fewer parameters) models introduced here, along with 26 other disorder prediction methods. Predictions of *AlphaFold2* have recently been shown to capture IDRs (Akdel et al., 2021; Wilson et al., 2021; Piovesan et al., 2022; Redl et al., 2022). However, we found the correlation between *AlphaFold2* predictions and CheZOD scores to be much lower than for SETH.

Redundancy-reduction affects performance estimates, not performance

We chose our datasets (training *CheZOD1174*, and testing *CheZOD117*) and performance measures (Eqn. (2) and (3) following a recent analysis (Nielsen and Mulder, 2019).

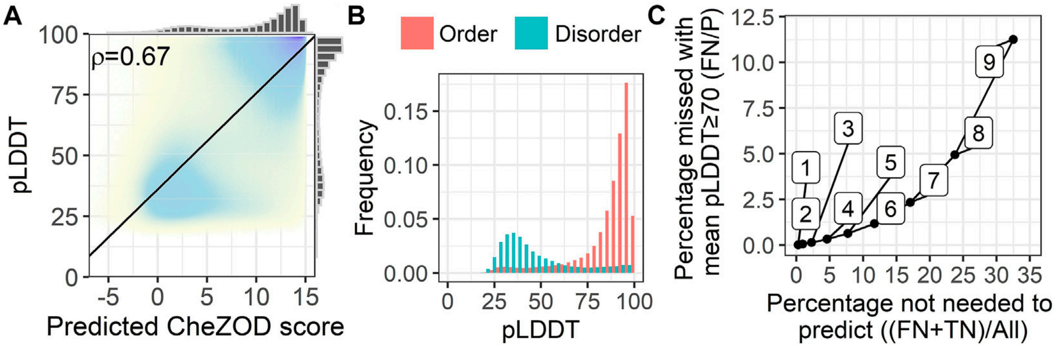

FIGURE 5

SETH's predictions correlated with AlphaFold2's pLDDT. Data: 17-ORGANISM-set (47,400,204 residues from 105,881 proteins in 17 organisms). **(A)** 2-dimensional histogram of *AlphaFold2* pLDDT (Jumper et al., 2021) against SETH disorder predictions (black line: optimal regression fit, marginal histograms on each axis; number: overall Spearman correlation coefficient ρ, Eq. 2). **(B)** Histograms of the pLDDT of *AlphaFold2*, for the classes order (predicted CheZOD>8) and disorder (predicted CheZOD≤8). **(C)** Cost versus gain analysis using SETH as a pre-filter for *AlphaFold2*. *Y*-axis—Cost: The percentage of proteins with a mean predicted CheZOD score below a certain threshold (thresholds marked as numbers on the curve), but mean pLDDT≥70 (FN) out of all proteins with mean pLDDT≥70 (P). This gives the percentage of proteins with a pLDDT≥70 missed using the SETH CheZOD score prediction as a pre-filter. *X*-axis—Gain: The percentage of proteins with mean CheZOD score < threshold (FN + TN) out of all proteins (All). This is the percentage of proteins in the entire dataset for which *AlphaFold2* will not have to be run at all, or defines a list of priority: first run *AlphaFold2* on the proteins with lower SETH disorder. For instance, with threshold 8, a quarter of all *AlphaFold2* predictions can be avoided at an error rate of only 5%.

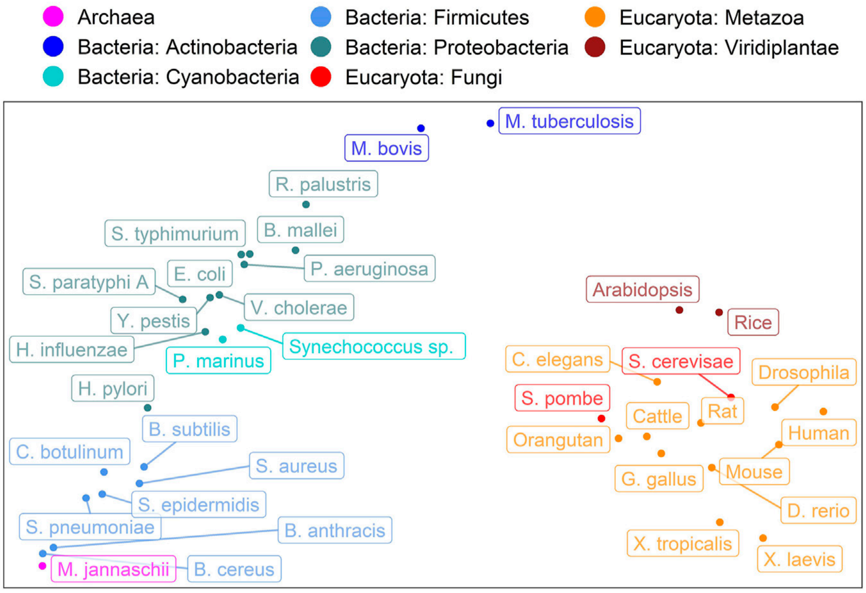

FIGURE 6

Evolution captured by spectrum of predicted CheZOD scores. Data: SETH predictions for proteins from 37 organisms taken from Swiss-Prot (The UniProt Consortium et al., 2021). The plot shows the PCA (Wold et al., 1987) of the binned spectrum of predicted CheZOD scores (8 bins, meaning 8-dimensional vectors; each vector describes one organism). The colors indicate the super-kingdoms (Eukaryota: reds, Bacteria: blues, Archaea: violet) as well as the Phyla for the Bacteria and the Kingdoms for the Eukaryotes, as given in UniProt (The UniProt Consortium et al., 2021). The organism names were shortened (for complete organism names see Supplementary Table S4).

However, since adequate redundancy-reduction is *sine qua non* to correctly estimate performance, we additionally removed 151 sequences from ODiNPred's (Dass et al., 2020) training set which had, based on alignments with 80% coverage, over 20% pairwise sequence identity (PIDE) with proteins in the test set (see Methods 2.1). Unfortunately, the threshold of sequence identity T (here 20% PIDE) crucially depends on the phenotype (here disorder). In the lack of sufficiently large data sets to establish T for disorder (and many other phenotypes, including protein-protein interactions (Park and Marcotte, 2012; Hamp and Rost, 2015)), developers should be as conservative as possible. However, there is a trade-off: chose T too low, lose proteins for training/testing, chose T too high, risk substantially over-estimating performance. With our threshold, we try to balance both. However, we only removed proteins which were aligned with 80% coverage, meaning there might still be some information leakage on a smaller level (3 test proteins with PIDE>20% to training proteins at a coverage of 10%; Supplementary Table S5). However, this leakage should be negligible, since none of the aligned proteins lie above the HSSP-curve (Rost, 1999). Even if there would still be some minor leakage of information, this might only balance out the over-estimates of performance of other methods, since over-estimating performance has become many times more common with the rise of AI with immense numbers of free parameters (often 10-times more parameters than samples), which can often easily zoom into residual sequence similarity between train and test set. Also considering that we used a quite conservative T, other methods tested on the same test set might more likely overestimate their performance. We cannot answer whether this over-estimate of any method is statistically insignificant or significant. That depends on many aspects of the method.

Supervised models picked up class imbalance

The training and test sets resulting from redundancy reduction differed substantially in their distributions of CheZOD scores (Supplementary Figure S1; note the test set *CheZOD117* had not been changed, only the training set). In a binary projection, the fraction of ordered residues was 72% for the training and 31% for the testing set. Our regression models did not use any notion of classes. Thus, we could not correct for class imbalance. This might explain why our supervised regression models trained on this imbalanced data (SETH, LinReg, LinReg1D and ANN) mildly over-predicted the degree of residue order compared to the raw embedding values of dimension 295 (Supplementary Figure S6).

Simple classification model LogReg struggled where SETH excelled

We tested the effect of increasing the model complexity when inputting only embeddings. For an ideal prediction method observed and predicted CheZOD scores would perfectly correlate, i.e., in a scatter plot with observed on the *x*-axis and predicted on the *y*-axis, perfect methods would cluster all points around the diagonal (Supplementary Figure S6). Qualitatively, our two most complex methods SETH and ANN came closest to this, followed by the simpler model LinReg, with more spread-out clusters (Supplementary Figure S6A–C). In contrast to an ideal prediction, the simplest model LogReg generated two clusters, one around probability 0 (disorder) and the other around 1 (order; Supplementary Figure S6D). 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. This qualitative analysis was supported by the ρ (Figure 3A: SETH highest, LogReg lowest). Therefore, we established that the treatment of disorder as a regression problem (SETH, ANN, LinReg) improved over the supervised training on binary assignments (disorder/order; LogReg; Figure 3). This was interesting because except for ODiNPred (Dass et al., 2020) and ADOPT (Redl et al., 2022), most SOTA disorder prediction methods realize a binary classification. However, the ρ was still similar between all our four models, including LogReg. Likewise, the performance on binarized CheZOD scores (order: CheZOD score>8, disorder: CheZOD score≤8), measured with the AUC did also not vary significantly. Nonetheless, SETH was consistently superior by all criteria (Figure 3).

Simpler, better, faster

The simplicity of a machine learning model can be proxied by the number of free parameters. Our top performing models SETH, ANN, LinReg and LogReg did not reach anywhere near the simplicity of earlier IDR prediction methods such as NORS (Liu et al., 2002) or IUPred (Dosztanyi et al., 2005) or recent adaptations of *AlphaFold2* predictions (Akdel et al., 2021; Wilson et al., 2021; Piovesan et al., 2022; Redl et al., 2022) when neglecting *AlphaFold2's* training and only considering the disorder prediction from *AlphaFold2's* output. Then, *AlphaFold2* binary disorder prediction would only need three parameters: choice of feature (e.g., RSA vs. pLDDT), averaging window (e.g., 25 for RSA) and a threshold (RSA < T). However, we still constrained the size of our models (Supplementary Table S3). The comparison to one-hot encodings clearly demonstrated the benefit of increasing model complexity by inputting high dimensional pLM embeddings (Figure 1). Lastly, our simplification of LinReg (LinReg1D) based on one of the 1,024 dimensions of ProtT5 (Elnaggar et al., 2021), namely dimension 295 that carried 86%–96% of the signal of the entire 1024-dimensional vector (Figure 3), reached the simplicity of very basic predictors. Still, it outperformed most complex methods.

Two of our models numerically reached higher AUC values than all other methods compared (SETH and LinReg, Figure 3B), irrespective of whether they use MSAs or not. When considering the ρ (Figure 3A), again two of our methods (SETH and ANN) outperformed all others. In terms of statistical significance for the ρ at the CI = 95% level, all our models along with ODiNPred (Dass et al., 2020) and ADOPT ESM-1b (Redl et al., 2022) significantly outperformed all others. Of these top performers, only ODiNPred relies on MSAs, i.e., this is the only top performer for which we first need to create informative MSAs before we can analyze the disorder content of a newly sequenced proteome. Even using tools such as the blazingly fast *MMseqs2* (Steinegger and Söding, 2017), this will still slow down the analysis. In contrast, ADOPT ESM-1b also only requires pLM embeddings as input. Given the larger model used by ADOPT ESM-1b and the larger size of ESM-1b (Rives et al., 2021) compared to ProtT5 (Elnaggar et al., 2021) used by our tools, we expect the difference in speed to favor SETH more than that in performance.

AlphaFold2 not competitive to pLM-based methods as proxy for CheZOD disorder

AlphaFold2's pLDDT correlates with binary descriptions of IDRs (Akdel et al., 2021; Wilson et al., 2021; Piovesan et al., 2022). In principle, we confirmed this for CheZOD scores reflecting non-binary disorder (Supplementary Figure S6G). However, we also found *AlphaFold2* to often be certain about a predicted structure (high pLDDT) even for regions where CheZOD scores suggest long IDRs (≥30 residues; Supplementary Figure S7). 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 well-structured upon interaction. This hypothesis was supported by a very limited analysis comparing the pLDDT to DisProt annotations [(Quaglia et al., 2022); Supplementary Figure S8]. Furthermore, the mean pLDDT is trivially higher for shorter than for longer proteins (Monzon et al., 2022). As proteins in the test set were shorter than average (mean sequence length in *CheZOD117*: 112), this trivial length-dependence might also explain some outliers.

Comparing several ways to utilize *AlphaFold2* predictions as a direct means to predict CheZOD scores revealed the window-averaged of the RSA to correlate even better with CheZOD scores than the prediction of "experimentally resolved" and the (smoothed) pLDDT (Figure 3). It outperformed all but two (ODiNPred (Dass et al., 2020), SPOT-dis (Hanson et al., 2016)) of the methods not based on pLMs. However, all four methods introduced here (SETH, ANN, LinReg, LogReg) and ADOPT ESM-1b (Redl et al., 2022) topped this.

Concluding, 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 48 GB 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 for the many proteins where *AlphaFold2* predictions are already available, the degree to which SETH outperformed disorder measures derived from *AlphaFold2* and the speed of SETH suggest to always use SETH instead of *AlphaFold2* to predict CheZOD-like disorder.

Agreement between SETH's disorder predictions and *AlphaFold2's* pLDDT

AlphaFold2's recent release of structure predictions (28 July 2022), expanding the *AlphaFold2* database to over 200 million predictions, has considerably expanded the structural coverage in the protein Universe. However, each day new proteins and proteomes are discovered and will require *AlphaFold2* 3D predictions. Could SETH help to prioritize how to run *AlphaFold2*, e.g., choosing the proteins most likely to have high pLDDTs (i.e., ordered proteins) first and leaving the rest for later, or completely neglecting the rest (i.e., disordered proteins)? Toward this end, we analyzed a large set of residues from 17 organisms and found the correlation between SETH's predictions and *AlphaFold2*'s pLDDT (Figure 5A) to be much higher than the correlation between the pLDDT and the ground truth CheZOD scores (ρ(AlphaFold2_pLDDT, ground truth) = 0.56 vs. ρ(AlphaFold2_pLDDT, SETH) = 0.67). This confirmed the agreement in over-prediction of order for SETH and *AlphaFold2* (Supplementary Figure S8) because if SETH and *AlphaFold2* make the same mistakes, a higher correlation is expected. These findings are at the base of using SETH to pre-filter or prioritize *AlphaFold2* predictions, e.g., using SETH protein mean CheZOD scores<8 to deprioritize or exclude some proteins will reduce costs for *AlphaFold2* by one-quarter at an error rate of only 5%.

The comparisons between SETH and *AlphaFold2* also might help to rationalize some predictions, e.g., for organisms with low mean pLDDT values, SETH often predicted order where *AlphaFold2* could not predict reliable 3D structures (Supplementary Figures S9, 11). Such cases might suggest that there are some "principles of protein structure formation" not yet captured by the outstanding *AlphaFold2*. More detailed studies will have to address this speculation.

CheZOD score disorder not equal to binary disorder

Most methods developed in the field of disorder predictions are trained on binary data: *disordered* (IDR: intrinsically disordered regions/IDP: intrinsically disordered proteins) as opposed to *well-structured/ordered*. Although this is standard procedure for machine learning, the situation for disorder is slightly different. There, we assume the set of all experimentally known 3D structures as deposited in the PDB (Burley et al., 2019) to be more representative of all well-ordered proteins than DisProt (Vucetic et al., 2005; Quaglia et al., 2022) of all disordered proteins, as the diversity of disorder is much more difficult to capture experimentally. Thus, for disorder we have many reasons to doubt that today's experimental data are representative. This creates a "Gordian knot": how do we train on unknown data? In previous work, we tried cutting through this knot by training on data differing from DisProt data (long loops, low contact density), but testing on DisProt (Liu et al., 2002; Schlessinger et al., 2007a; Schlessinger et al., 2007b), as, for instance, the successful method IUPred did for contacts (Dosztanyi et al., 2005). Instead, here we used CheZOD scores (Nielsen and Mulder, 2016, 2019; Dass et al., 2020) introduced by Nielsen and Mulder as the "secret order in disorder". The CheZOD perspective appealed to us because of three reasons. Firstly, it provides details or nuances for each residue. Secondly, it partially eradicates the need for a minimal threshold of continuous regions: most loops (non-regular secondary structure) of, e.g., 5–15 residues are absolutely unrelated to what we consider disorder, while loops with over 30 consecutive residues clearly fall into two distinct classes of long-loops in regular structures and disordered regions (Schlessinger et al., 2007a). Thirdly, the non-binary classification allowed to describe an entire organism by an 8-dimensional vector that captured evolution (Figure 6).

A recent large-scale evaluation of disorder prediction methods (Nielsen and Mulder, 2019; Dass et al., 2020) and one of CAID's [Critical Assessment of Protein Intrinsic Disorder Prediction, (Necci et al., 2021)] top methods SPOT-Disorder2 show that methods for binary disorder prediction capture information about CheZOD scores. Inversely, SETH, trained on CheZOD scores, appears to capture aspects of binary disorder (as suggested by some preliminary results from the second round of CAID). On the other hand, another one of the CAID-top methods for predicting binary disorder flDPnn (Hu et al., 2021), did not reach top rank for CheZOD scores (Figure 3). Consequently, CheZOD scores might be the "secret order in disorder", but they probably capture aspects somehow orthogonal to binary disorder.

Spectra of predicted CheZOD-disorder capture rudimentary aspects of evolution

Spectra of predicted protein location capture aspects of the evolution of eukaryotes (Marot-Lassauzaie et al., 2021). Additionally, the fraction of intrinsically disordered proteins in a proteome has been revealed as a marker for important aspects in the evolution of species (Dunker et al., 1998; Liu et al., 2002; Fuxreiter et al., 2008; Pentony and Jones, 2009; Uversky et al., 2009; Brown et al., 2011; Schlessinger et al., 2011; Vicedo et al., 2015a; Vicedo et al., 2015b). However, the single number (fraction of IDP in proteome) was too simplistic for analyses as applied to the location spectrum based on 10-dimensional vectors representing ten different subcellular compartments. The crucial step was the prediction of non-binary CheZOD scores and the idea to bin those into a spectrum with eight bins leading to 8-dimensional vectors subjected to straightforward PCA (Wold et al., 1987). Surprisingly, this already revealed a connection between the micro molecular level of per-residue CheZOD score predictions and the macro level of the evolution of species (Figure 6). Minimally, this finding suggests that adjusting—increasing or reducing - the composition of disordered residues in proteins is a tracer of or proxy for evolutionary events. Possibly, these changes might play a role in speciation. However, at this point, the latter remains speculation. Clearly, the analysis revealed another interesting simple feature relating the micro and macro level, i.e., connecting the machinery of the proteins that shape life to the carriers of these molecular machines, namely the organisms.

Conclusion

We introduced four relatively simple 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 NMR derived chemical shift Z-scores (CheZOD scores (Nielsen and Mulder, 2020)). The best approach, dubbed SETH, captured fine-grained nuances of disorder on a continuous scale and, in our hands, appeared to outperform all compared state-of-the-art methods [(Nielsen and Mulder, 2019; Dass et al., 2020; Redl et al., 2022); Figure 3]. Our solutions were so successful because the unoptimized embeddings carried important information about disorder (Figure 2), to the extent that mostly one of the 1,024 dimensions mattered (Supplementary Figure S6E). 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 about 1 hour on a consumer-grade PC with one NVIDIA GeForce RTX 3060. Therefore, it enables large-scale analyses of disorder, which allowed us to show that CheZOD score distributions capture evolutionary information (Figure 6). Although the break-through *AlphaFold2* (Jumper et al., 2021) 3D predictions are now available for most proteins, and although we could show that disorder measures of *AlphaFold2* predictions correlate with CheZOD scores, the correlation was significantly inferior to the predictions of SETH, suggesting the investment of fewer than 3 min per 1,000 proteins.

Data availability statement

SETH is available to download at https://github.com/Rostlab/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 Consortium et al., 2021) and the human proteome are available at https://doi.org/10.5281/zenodo.6673817. The datasets presented in this study (training set: CheZOD1174 and test set: CheZOD117) can be found in online repositories: https://github.com/Rostlab/SETH.

Author contributions

MH provided the AlphaFold2 structures (including pLDDT and implementation to derive the "Experimentally resolved" prediction), relative solvent accessibility values calculated from AlphaFold2 predictions, Embeddings, SETH's predictions, and performed the MMSeqs2 clustering. MH also provided a draft of the code provided at https://github.com/Rostlab/SETH and at https://colab.research.google.com/drive/1vDWh5YI_BPxQg0ku6CxKtSXEJ25u2wSq?usp=sharing, which was refined by DI. DI trained and tested all models and analyzed the data with help of MH. DI wrote the first manuscript draft and generated the figures; iteratively, MH, DI, and BR refined the manuscript. BR helped in the study planning, design and setting the focus. All authors read and approved the final manuscript.

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).

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 and thank you to Konstantin Schütze (TUM) for his input in the review process. Also, thanks to the constructive criticism from the reviewers and to Dr. Yaoqi Zhou for contributing SPOT-Disorder2 results. 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. The manuscript has been published as a preprint (Ilzhoefer et al., 2022).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2022.1019597/full#supplementary-material

References

Akdel, M., Pires, D. E. V., Porta Pardo, E., Jänes, J., Zalevsky, A. O., Mészáros, B., et al. (2021). A structural biology community assessment of AlphaFold 2 applications. bioRxiv, 2021.2009.2026.461876. doi:10.1101/2021.09.26.461876

Alley, E. C., Khimulya, G., Biswas, S., Alquraishi, M., and Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* 16, 1315–1322. doi:10.1038/s41592-019-0598-1

Asgari, E., and Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE* 10, e0141287. doi:10.1371/journal.pone.0141287

Bepler, T., and Berger, B. (2019). "Learning protein sequence embeddings using information from structure," in The International Conference on Learning Representations, New Orleans, United States. *arXiv:1902.08661 [cs, q-bio, stat]*.

Bepler, T., and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. *Cell. Syst.* 12, 654–669.e3. e653. doi:10.1016/j.cels.2021.05.017

Bordin, N., Sillitoe, I., Nallapareddy, V., Rauer, C., Lam, S. D., Waman, V. P., et al. (2022). AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. bioRxiv, 2022.2006.2002.494367. doi:10.1101/2022.06.02.494367

Brown, C. J., Johnson, A. K., Dunker, A. K., and Daughdrill, G. W. (2011). Evolution and disorder. *Curr. Opin. Struct. Biol.* 21, 441–446. doi:10.1016/j.sbi.2011.02.005

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., et al. (2019). RCSB protein data bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* 47, D464–D474. doi:10.1093/nar/gky1004

Cheng, J., Sweredoski, M. J., and Baldi, P. (2005). Accurate prediction of protein disordered regions by mining protein structure data. *Data Min. Knowl. Discov.* 11, 213–222. doi:10.1007/s10618-005-0001-y

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2014). The DynaMine webserver: Predicting protein dynamics from sequence. *Nucleic Acids Res.* 42, W264–W270. doi:10.1093/nar/gku270

Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221, 709–713. doi:10.1126/science.6879170

Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A. X., et al. (2021). Learned embeddings from deep learning to visualize and predict protein sets. *Curr. Protoc.* 1, e113. doi:10.1002/cpz1.113

Dass, R., Mulder, F. a. A., and Nielsen, J. T. (2020). ODiNPred: Comprehensive prediction of protein order and disorder. *Sci. Rep.* 10, 14780. doi:10.1038/s41598-020-71716-1

Deng, X., Eickholt, J., and Cheng, J. (2009). PreDisorder: Ab initio sequence-based prediction of protein disordered regions. *BMC Bioinforma.* 10, 436. doi:10.1186/1471-2105-10-436

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *Bert: Pre-Training of deep bidirectional transformers for language understanding*. Minneapolis, Minnesota: Association for Computational Linguistics.

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. doi:10.1093/bioinformatics/bti541

Dunker, A. K., Babu, M. M., Barbar, E., Blackledge, M., Bondos, S. E., Dosztányi, Z., et al. (2013). What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. *Intrinsically Disord. Proteins* 1, e24157. doi:10.4161/idp.24157

Dunker, A. K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., et al. (1998). Protein disorder and the evolution of molecular recognition: Theory, predictions and observations. *Pac. Symp. Biocomput.* 3, 473–484.

Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008). Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* 18, 756–764. doi:10.1016/j.sbi.2008.10.002

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.* 6, 197–208. doi:10.1038/nrm1589

Efron, B., and Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science* 353, 390–395. doi:10.1126/science.253.5018.390

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., et al. (2021). ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (10), 7112–7127. doi:10.1109/TPAMI.2021.3095381

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 Res.* 42, D304–D309. doi:10.1093/nar/gkt1240

Fuxreiter, M., Tompa, P., Simon, I., Uversky, V. N., Hansen, J. C., and Asturias, F. J. (2008). Malleable machines take shape in eukaryotic transcriptional regulation. *Nat. Chem. Biol.* 4, 728–737. doi:10.1038/nchembio.127

Hamp, T., and Rost, B. (2015). More challenges for machine-learning protein interactions. *Bioinformatics* 31, 1521–1525. doi:10.1093/bioinformatics/btu857

Hanson, J., Paliwal, K. K., Litfin, T., and Zhou, Y. (2019). SPOT-Disorder2: Improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics, Proteomics Bioinforma.* 17, 645–656. doi:10.1016/j.gpb.2019.01.004

Hanson, J., Paliwal, K., and Zhou, Y. (2018). Accurate single-sequence prediction of protein intrinsic disorder by an ensemble of deep recurrent and convolutional architectures. *J. Chem. Inf. Model.* 58, 2369–2376. doi:10.1021/acs.jcim.8b00636

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* 33, 685–692. doi:10.1093/bioinformatics/btw678

Hauser, M., Steinegger, M., and Söding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* 32, 1323–1330. doi:10.1093/bioinformatics/btw006

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., et al. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinforma.* 20, 723. doi:10.1186/s12859-019-3220-8

Heinzinger, M., Littmann, M., Sillitoe, I., Bordin, N., Orengo, C., and Rost, B. (2021). Contrastive learning on protein embeddings enlightens midnight zone. *Bioinformatics* 4 (2), lqac043. doi:10.1093/nargab/lqac043

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

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. doi:10.1002/prot.21123

Howard, M. J. (1998). Protein NMR spectroscopy. *Curr. Biol.* 8, R331–R333. doi:10.1016/s0960-9822(98)70214-3

Hu, G., Katuwawala, A., Wang, K., Wu, Z., Ghadermarzi, S., Gao, J., et al. (2021). flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.* 12, 4438. doi:10.1038/s41467-021-24773-7

Ilzhoefer, D., Heinzinger, M., and Rost, B. (2022). SETH predicts nuances of residue disorder from protein embeddings. bioRxiv, 2022.2006.2023.497276.

Ishida, T., and Kinoshita, K. (2007). PrDOS: Prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* 35, W460–W464. doi:10.1093/nar/gkm363

Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinforma.* 11, 431. doi:10.1186/1471-2105-11-431

Jones, D. T., and Cozzetto, D. (2015). DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 31, 857–863. doi:10.1093/bioinformatics/btu744

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* 10, 5416. doi:10.1038/s41467-019-13056-x

Kozlowski, L. P., and Bujnicki, J. M. (2012). MetaDisorder: A meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinforma.* 13, 111. doi:10.1186/1471-2105-13-111

Lange, J., Wyrwicz, L. S., and Vriend, G. (2016). Kmad: Knowledge-based multiple sequence alignment for intrinsically disordered proteins. *Bioinformatics* 32, 932–936. doi:10.1093/bioinformatics/btv663

Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 31, 3701–3708. doi:10.1093/nar/gkg519

Littmann, M., Bordin, N., Heinzinger, M., Schütze, K., Dallago, C., Orengo, C., et al. (2021a). Clustering FunFams using sequence embeddings improves EC purity. *Bioinformatics* 37, 3449–3455. doi:10.1093/bioinformatics/btab371

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. doi:10.1038/s41598-020-80786-0

Littmann, M., Heinzinger, M., Dallago, C., Weissenow, K., and Rost, B. (2021c). Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci. Rep.* 11, 23916. doi:10.1038/s41598-021-03431-4

Liu, J., Tan, H., and Rost, B. (2002). Loopy proteins appear conserved in evolution. *J. Mol. Biol.* 322, 53–64. doi:10.1016/s0022-2836(02)00736-2

Marot-Lassauzaie, V., Goldberg, T., Armenteros, J. J. A., Nielsen, H., and Rost, B. (2021). Spectrum of protein location in proteomes captures evolutionary relationship between species. *J. Mol. Evol.* 89, 544–553. doi:10.1007/s00239-021-10022-4

Marquet, C., Heinzinger, M., Olenyi, T., Dallago, C., Erckert, K., Bernhofer, M., et al. (2021). Embeddings from protein language models predict conservation and variant effects. *Hum. Genet.* doi:10.1007/s00439-021-02411-y

Marx, V. (2022). Method of the year: Protein structure prediction. *Nat. Methods* 19, 5–10. doi:10.1038/s41592-021-01359-1

Mirabello, C., and Wallner, B. (2019). rawMSA: End-to-end deep learning using raw multiple sequence alignments. *PLOS ONE* 14, e0220182. doi:10.1371/journal.pone.0220182

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 Res.* 45, D170–D176. doi:10.1093/nar/gkw1081

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 Disord. Proteins* 1, e24428. doi:10.4161/idp.24428

Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., and Fidelis, K. (2014). Assessment of protein disorder region predictions in CASP10. *Proteins.* 82, 127–137. doi:10.1002/prot.24391

Monzon, V., Haft, D. H., and Bateman, A. (2022). Folding the unfoldable: Using AlphaFold to explore spurious proteins. *Bioinforma. Adv.* 2, vbab043. doi:10.1093/bioadv/vbab043

Necci, M., Piovesan, D., Predictors, C., Disprot, C., and Tosatto, S. C. E. (2021). Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* 18, 472–481. doi:10.1038/s41592-021-01117-3

Nielsen, J. T., and Mulder, F. a. A. (2019). Quality and bias of protein disorder predictors. *Sci. Rep.* 9, 5137. doi:10.1038/s41598-019-41644-w

Nielsen, J. T., and Mulder, F. a. A. (2020). “Quantitative protein disorder assessment using NMR chemical shifts,” in *Intrinsically disordered proteins*. Editors B. B. Kragelund and K. Skriver (New York, NY: Springer US), 303–317.

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”. *Front. Mol. Biosci.* 3, 4. doi:10.3389/fmolb.2016.00004

Nwanochie, E., and Uversky, V. N. (2019). Structure determination by single-particle cryo-electron microscopy: Only the sky (and intrinsic disorder) is the limit. *Int. J. Mol. Sci.* 20, 4186. doi:10.3390/ijms20174186

Ofer, D., Brandes, N., and Linial, M. (2021). the language of proteins: NLP, machine learning & protein sequences. *Comput. Struct. Biotechnol. J.* 19, 1750–1758. doi:10.1016/j.csbj.2021.03.022

Oldfield, C. J., Xue, B., Van, Y.-Y., Ulrich, E. L., Markley, J. L., Dunker, A. K., et al. (2013). Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochimica Biophysica Acta - Proteins Proteomics* 1834, 487–498. doi:10.1016/j.bbapap.2012.12.003

Park, Y., and Marcotte, E. M. (2012). Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods* 9, 1134–1136. doi:10.1038/nmeth.2259

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). in *PyTorch: An imperative style, high-performance deep learning library*. Editors H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox, and R. Garnett (New York: Curran Associates, Inc).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (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. doi:10.1142/s0219720005000886

Pentony, M. M., and Jones, D. T. (2009). Modularity of intrinsic disorder in the human proteome. *Proteins.* 78, 212–221. doi:10.1002/prot.22504

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). *Deep contextualized word representations*. New Orleans, Louisiana: Association for Computational Linguistics.

Piovesan, D., Monzon, A. M., and Tosatto, S. C. E. (2022). Intrinsic protein disorder, conditional folding and AlphaFold2. *bioRxiv* [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/2022.03.03.482768v1.

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., et al. (2005). FoldIndex(C): A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 21, 3435–3438. doi:10.1093/bioinformatics/bti537

Quaglia, F., Mészáros, B., Salladini, E., Hatos, A., Pancsa, R., Chemes, L. B., et al. (2022). DisProt in 2022: Improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res.* 50, D480–D487. doi:10.1093/nar/gkab1082

R Core Team (2021). R: A language and environment for statistical computing. *MSOR Connect.* 1.

Radivojac, P., Obradovic, Z., Brown, C. J., and Dunker, A. K. (2002). “Improving sequence alignments for intrinsically disordered proteins,” in *Pacific symposium on biocomputing. Pacific symposium on biocomputing*. Lihue, HI, United States: WorldScientific, 589–600.

Radivojac, P., Obradovic, Z., Smith, D. K., Zhu, G., Vucetic, S., Brown, C. J., et al. (2004). Protein flexibility and intrinsic disorder. *Protein Sci.* 13, 71–80. doi:10.1110/ps.03128904

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer*. Journal of Machine Learning Research. *arXiv*.

Reddi, S. J., Kale, S., and Kumar, S. (2018). “On the convergence of Adam and beyond,” in The International Conference on Learning Representations, Vancouver, BC, Canada.

Redl, I., Fisicaro, C., Dutton, O., Hoffmann, F., Henderson, L., Owens, B. M. J., et al. (2022). Adopt: Intrinsic protein disorder prediction through deep bidirectional transformers. bioRxiv, 2022.2005.2025.493416.

Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9, 173–175. doi:10.1038/nmeth.1818

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2016239118. doi:10.1073/pnas.2016239118

Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., Garner, E., Guilliot, S., et al. (1998). “Thousands of proteins likely to have long disordered regions Pacific Symposium on Biocomputing,” in *Pacific symposium on biocomputing*. Kapalua, Maui, Hawaii: WorldScientific, 437–448.

Rost, B., and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins.* 20, 216–226. doi:10.1002/prot.340200303

Rost, B., and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232, 584–599. doi:10.1006/jmbi.1993.1413

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.* 12, 85–94. doi:10.1093/protein/12.2.85

Schlessinger, A., Liu, J., and Rost, B. (2007a). Natively unstructured loops differ from other loops. *PLoS Comput. Biol.* 3, e140. doi:10.1371/journal.pcbi.0030140.eor

Schlessinger, A., Punta, M., and Rost, B. (2007b). Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics* 23, 2376–2384. doi:10.1093/bioinformatics/btm349

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., and Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE* 4, e4433. doi:10.1371/journal.pone.0004433

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., and Rost, B. (2011). Protein disorder—A breakthrough invention of evolution? *Curr. Opin. Struct. Biol.* 21, 412–418. doi:10.1016/j.sbi.2011.03.014

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. *J. Mol. Biol.* 427, 982–996. doi:10.1016/j.jmb.2014.12.007

Steinegger, M., Mirdita, M., and Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* 16, 603–606. doi:10.1038/s41592-019-0437-4

Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nat. Commun.* 9, 2542. doi:10.1038/s41467-018-04964-5

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi:10.1038/nbt.3988

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. doi:10.1093/bioinformatics/btu739

Tantos, A., Friedrich, P., and Tompa, P. (2009). Cold stability of intrinsically disordered proteins. *FEBS Lett.* 583, 465–469. doi:10.1016/j.febslet.2008.12.054

Tompa, P., Dosztanyi, Z., and Simon, I. (2006). Prevalent structural disorder in *E. coli* and *S. cerevisiae* proteomes. *J. Proteome Res.* 5, 1996–2000. doi:10.1021/pr0600881

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. doi:10.1002/prot.21773

Tompa, P., Szasz, C., and Buday, L. (2005). Structural disorder throws new light on moonlighting. *Trends biochem. Sci.* 30, 484–489. doi:10.1016/j.tibs.2005.07.008

The UniProt ConsortiumBateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi:10.1093/nar/gkaa1100

Uversky, V. N., Oldfield, C. J., Midic, U., Xie, H., Xue, B., Vucetic, S., et al. (2009). Unfoldomics of human diseases: Linking protein intrinsic disorder with diseases. *BMC Genomics* 10, S7. doi:10.1186/1471-2164-10-s1-s7 van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (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. doi:10.12688/f1000research.7178.1

Vicedo, E., Schlessinger, A., and Rost, B. (2015b). Environmental pressure may change the composition protein disorder in prokaryotes. *PLoS One* 10, e0133990. doi:10.1371/journal.pone.0133990

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., et al. (2005). DisProt: A database of protein disorder. *Bioinformatics* 21, 137–140. doi:10.1093/bioinformatics/bth476

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. doi:10.1093/bioinformatics/btr682

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. doi:10.1093/bioinformatics/btw446

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. *J. Mol. Biol.* 337, 635–645. doi:10.1016/j.jmb.2004.02.002

Wilson, C. J., Choy, W.-Y., and Karttunen, M. (2022). AlphaFold2: A role for disordered protein/region prediction?. *Int. J. Mol. Sci.* 23 (9), 4591. doi:10.3390/ijms23094591

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemom. intelligent laboratory Syst.* 2, 37–52. doi:10.1016/0169-7439(87)80084-9

Wright, P. E., and Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 293, 321–331. doi:10.1006/jmbi.1999.3110

Wu, Z., Johnston, K. E., Arnold, F. H., and Yang, K. K. (2021). Protein sequence design with deep generative models. *Curr. Opin. Chem. Biol.* 65, 18–27. doi:10.1016/j.cbpa.2021.04.004

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. doi:10.1093/bioinformatics/bti534