nature communications

Article

https://doi.org/10.1038/s41467-024-53982-z

SSEmb: A joint embedding of protein sequence and structure enables robust variant effect predictions

Received: 21 January 2024

Accepted: 28 October 2024

Published online: 07 November 2024

Check for updates

Lasse M. Blaabjerg[1], Nicolas Jonsson[1], Wouter Boomsma[2] ✉, Amelie Stein[1] ✉ & Kresten Lindorff-Larsen[1] ✉

The ability to predict how amino acid changes affect proteins has a wide range of applications including in disease variant classification and protein engineering. Many existing methods focus on learning from patterns found in either protein sequences or protein structures. Here, we present a method for integrating information from sequence and structure in a single model that we term SSEmb (Sequence Structure Embedding). SSEmb combines a graph representation for the protein structure with a transformer model for processing multiple sequence alignments. We show that by integrating both types of information we obtain a variant effect prediction model that is robust when sequence information is scarce. We also show that SSEmb learns embeddings of the sequence and structure that are useful for other downstream tasks such as to predict protein-protein binding sites. We envisage that SSEmb may be useful both for variant effect predictions and as a representation for learning to predict protein properties that depend on sequence and structure.

Small changes in the amino acid sequence of a protein can have a wide range of effects on its molecular structure, stability and function. Discerning the magnitude and consequences of such effects is central to understanding the molecular mechanisms of evolution and human disease[1]. Furthermore, the ability to manipulate sequences to change or optimize function is fundamental to the field of protein engineering and design[2].

The decreased cost of DNA sequencing has enabled the development of experiments that can generate biological data at a large scale. An example of this is the development of high-throughput assays that can provide a quantitative readout of changes in activity, stability, or abundance for thousands of protein variants in a single experiment. Such high-throughput assays, often called Multiplexed Assays of Variant Effects (MAVE) or Deep Mutational Scanning experiments, have enabled a substantial increase in the available data mapping the relationship between protein sequence and function[1,3–6].

Many different types of MAVEs have been developed to probe different aspects of the protein sequence-function relationship. In particular, changes in protein abundance have been shown to be an important driver of change in protein activity[7]. Thus, a specific type of MAVE called Variant Abundance by Massively Parallel Sequencing (VAMP-seq) has been developed to quantify variant effects on cellular protein abundance[8]. By combining data generated by multiple MAVEs that probe different effects for each substitution—for example, on abundance, activity, or binding—it is possible to obtain mechanistic insights into how and why particular amino acid substitutions affect protein function[9–12].

Although MAVEs, in principle, can provide a complete mapping of the protein sequence-function relationship, the assays themselves are often costly and time-consuming. Furthermore, it is difficult to design assays that probe all relevant functions of a protein, and it may not always be possible to probe all possible variants in a single assay. The combinatorial explosion of multiple variants presents an additional challenge in experimental screening. In contrast, computational predictors of variant effects are able to make predictions for variants that have not been assayed before at a close-to-zero marginal cost and can

[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen N, Denmark. [2]Center for Basic Machine Learning Research in Life Science, Department of Computer Science, University of Copenhagen, Copenhagen N, Denmark. ✉e-mail: wb@di.ku.dk; amelie.stein@bio.ku.dk; lindorff@bio.ku.dk

serve as an additional source of information in the protein sequence-function mapping[1].

Machine learning-based methods have proven to be a useful tool for predicting complex relationships in biology. Often, machine learning models are trained in a supervised manner, where the algorithm is trained to learn the mapping between a set of related input and target values. Although generally effective, such supervised learning algorithms do sometimes not accurately predict the variant effects probed by MAVEs, possibly due to experimental differences between assays that make it difficult to compare and standardize read-outs. In contrast, self-supervised algorithms that are trained directly on large amounts of input data only have emerged as a compelling alternative in the field of variant effect prediction[13–20].

The majority of self-supervised predictors of variant effects rely on a single type of protein data that is used to learn the implicit correlations within the data. The output of a self-supervised predictor is often a probability distribution over the possible amino acid substitutions at a particular position in the protein. Examples of the types of data used as input include the wild-type amino acid sequence[21,22], a multiple sequence alignment (MSA)[13,14,23–27], or the protein structure[28–30]. Some methods have combined predictions from multiple protein data types at an aggregate level[7,12,31,32], although some results suggest that a richer representation might be learned by combining multiple data types at the input level[19,33–40].

Here, we present the SSEmb (Sequence Structure Embedding) model (Fig. 1). The idea behind our model is that mapping multiple sources of protein information − in this case, an MSA and the three-dimensional structure − into a single learned embedding should yield a model that is able to make more robust predictions, i.e., predictions that are less sensitive to lack of information in a single input. Specifically, we base our model on the MSA Transformer model[41], which can be used to make predictions of variant effects using a subsampled MSA as input[16]. Although the MSA Transformer performs well in variant effect prediction, the accuracy of the predictions has been shown to be sensitive to the depth of the MSA used as input[6]. By constraining the MSA Transformer with structure information and combining the learned embeddings with a structure-based graph neural network (GNN)[42], we show that we can obtain improved variant effect predictions when the input MSA is shallow. In contrast to other recent methods[37], SSEmb is trained fully end-to-end, with a focus on integrating and aligning sequence- and structure-based information throughout the entire model. We show that the resulting dual embedding of sequence and structure in SSEmb is information-rich, leads to accurate predictions of variant effects probed by MAVEs, and can be used for other downstream tasks besides variant effect prediction. We exemplify this by using the SSEmb embeddings to predict protein-protein binding sites with results comparable to specialized state-of-the-art methods.

Results

Development of the SSEmb model

The SSEmb model was trained in a self-supervised manner using a combination of MSAs and protein structures (Fig. 1). The protein structures were taken from the previously compiled CATH 4.2 data set[43] that contains 18,204 training proteins with 40% non-redundancy partitioned by CATH class. For each of these protein structures, we also generated an MSA[44]. Before training, we removed proteins from the data set that were also present in the MAVE validation set or in the ProteinGym[6] test set using a 95% sequence identity cut-off. During training, we mask a random subset of amino acid positions in the wild-type amino acid sequence, and the SSEmb model was trained to predict the masked amino acid type. We constrain the MSA Transformer with structure information by only allowing attention between positions in the MSA that are proximal in the three-dimensional protein structure. In order to combine information from the MSA and the protein structure at the model input level, features from the structure-constrained MSA Transformer were input to the nodes of the protein graph processed by the GNN module. Specifically, we extracted the last-layer embeddings from the structure-constrained MSA Transformer for each of the MSA query sequence positions and concatenated these embedding features to the nodes of the GNN. The structure-constrained MSA Transformer model was initialized using weights from the original pre-trained MSA Transformer, while the GNN was trained from scratch with an architecture similar to the Geometric Vector Perceptron (GVP) model[42]. Further details on SSEmb model architecture and training can be found in the Methods section.

Validation using multiplexed assays of variant effects

During training, we validated the SSEmb model on the results from ten MAVEs probing the effects of individual substitutions. Various model design choices, along with relevant hyperparameters, were selected based on SSEmb's performance on these validation assays, which we selected based on three main criteria. First, the validation set should contain a mix of assays probing protein activity and abundance, with a majority of assays probing activity. Because protein activity and abundance are known to be correlated but distinct measures[9], this allowed us to obtain better insights into what the model was learning during training. Second, the validation set should contain a mix of assays considered either difficult or relatively easier to predict with methods based either on protein structure or sequence alignments. Because these methods are known to capture some aspects of the data, this criterion enabled us to investigate whether the model was

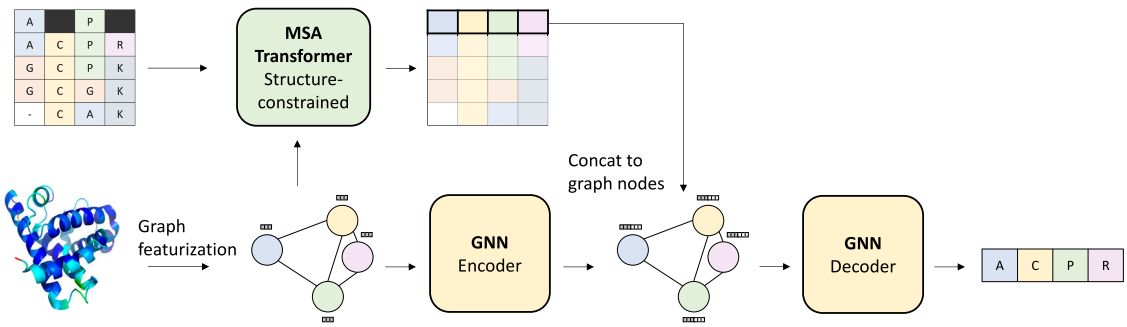

Fig. 1 | Overview of the SSEmb model and how it is trained. The model takes as input a subsampled MSA with a partially masked query sequence and a complete protein structure. The protein structure graph is used to mask (constrain) the row attention (i.e., attention across MSA columns) in the MSA Transformer. The MSA query sequence embeddings from the structure-constrained MSA Transformer are concatenated to the protein graph nodes. During training, SSEmb tries to predict the amino acid type at the masked positions. The model is optimized using the cross-entropy loss between the predicted and the true amino acid tokens at the masked positions. Variant effect prediction is made from these predictions as described in Methods.

Table 1 | Overview of SSEmb results on the MAVE validation set after model training

Protein	MAVE reference	MAVE type	Spearman $\|\rho_S\|$ (↑)		
			SSEmb	GEMME	Rosetta
NUD15	Suiter et al. 2020	Abundance	**0.584**	0.543	0.437
TPMT	Matreyek et al. 2018	Abundance	0.523	**0.529**	0.489
CP2C9	Amorosi et al. 2021	Abundance	**0.609**	0.423	0.519
P53	Kotler et al. 2018	Competitive growth	0.577	**0.655**	0.488
PABP	Melamed et al. 2013	Competitive growth	**0.595**	0.569	0.384
SUMO1	Weile et al. 2017	Competitive growth	**0.481**	0.406	0.433
RL401	Roscoe & Bolon 2014	E1 reactivity	**0.438**	0.390	0.366
PTEN	Mighell et al. 2018	Competitive growth	0.422	**0.532**	0.423
MAPK	Brenan et al. 2016	Competitive growth	0.395	**0.445**	0.307
LDLRAP1	Jiang et al. 2019	Two-hybrid assay	**0.411**	0.348	0.377
Mean	–	–	**0.503**	0.484	0.422

We use the Spearman correlation coefficient to quantify the agreement between the data generated by the MAVEs and the predictions from SSEmb, GEMME, and Rosetta. In this validation, only single-mutant variant effects were considered. The following protein structures were used in the SSEmb and Rosetta input: NUD15: 5BON_A, TPMT: 2H11_A, CP2C9: 1R90_A, P53: 4QO1_B, PABP: 1CVJ_G, SUMO1: 1WYW_B, RL401: 6NYO_E, PTEN: 1D5R_A, MAPK: 4QTA_A, LDLRAP1: 3SO6_A. The following UniProt IDs were used as input to construct multiple sequence alignments for GEMME: NUD15: Q9NV35, TPMT: P51580, CP2C9: P11712, P53: P04637, PABP: P11940, SUMO1: P63165, RL401: P0CH08, PTEN: P60484, MAPK1: P28482, LDLRAP1: Q5SW96. Some assay data points were removed during the merging of predictions in order to facilitate fair comparison between models.

[1] Bold values correspond to the best-performing model for each MAVE.

learning the same aspects as well as correlations not picked up by these methods. As an example, structure-based variant effect prediction methods have been shown to correlate better with measurements of changes in protein stability and abundance than with measurements of changes in protein activity, such as those probed via competitive growth assays[7,9,12,45,46]. By including both types of assays in the validation set, we could better gauge how well the model captured different mechanistic aspects during model development. Third, as the total number of published data sets generated by MAVEs is still relatively small, we sought to develop a validation set that was as small as possible while still enabling informative feedback during model development. The assay data for the validation set was taken from ProteinGym[6], with the exception of the LDLRAP1 assay, where we used data from Jiang and Roth[47,48], and the MAPK assay, where we used data from Brenan et al.[49,50].

To benchmark SSEmb performance during development, we compare our results to two non-machine-learning methods that use either only the MSA or only the protein structure as input. For the MSA-based model, we selected GEMME[27], which has been shown to produce state-of-the-art results for protein activity prediction outperforming most current machine learning methods using a relatively simple evolutionary model[6]. For the structure-based model, we selected stability predictions using Rosetta[51], which are commonly used within protein engineering and have been shown to make useful predictions of protein stability and abundance[7,52,53].

After training, the SSEmb model achieves a higher Spearman correlation on the MAVE validation set than both GEMME and Rosetta (Table 1). In particular, the SSEmb model performs approximately the same as GEMME on the activity-based MAVEs that probe protein functions, but performs better overall on the abundance assays, suggesting that the added structural information in the SSEmb model helps it make better predictions for this problem. Overall, these results indicate that SSEmb is able to make accurate variant effect predictions that correlate well with measures of both activity and abundance.

Testing SSEMb on ProteinGym benchmark

ProteinGym is a large collection of data generated by MAVEs that has been used for benchmarking variant effect prediction models[6]. The ProteinGym substitution benchmark, as originally collected in ref. 6, contains a total of 87 datasets on 72 different proteins. Of the 87 datasets in ProteinGym, 76 contain only single substitution effects, whereas the remaining 11 assays include variant effects from multiple substitutions. When testing the SSEmb model on the ProteinGym substitution set, we excluded data generated by the nine MAVEs, which were part of our validation set. Furthermore, when multiple assays are present for a single UniProt ID, we report the mean correlation over the assays in order to remove potential biases from correlated measurements.

Table 2 | SSEmb performance on the originally released ProteinGym substitution benchmark compared to other variant effect prediction models grouped by UniProt ID and segmented by MSA depth

Model	Spearman ρ_S by MSA depth (↑)			
	Low	Medium	High	All
TranceptEVE L	**0.451**	**0.462**	0.502	**0.468**
GEMME	0.429	0.448	0.495	0.453
SSEmb (ours)	0.449	0.439	0.501	0.453
Tranception L	0.438	0.438	0.467	0.444
EVE (ensemble)	0.412	0.438	0.493	0.443
VESPA	0.411	0.422	**0.514**	0.438
EVE (single)	0.405	0.431	0.488	0.437
MSA Transformer (ensemble)	0.385	0.426	0.470	0.426
ESM2 (15B)	0.342	0.368	0.433	0.375
ProteinMPNN	0.189	0.151	0.237	0.175

Assays from the SSEmb validation set have been excluded from the original data set. Low: $N_{eff}/L < 1$, Medium: $N_{eff}/L < 100$, High: $N_{eff}/L > 100$[6].

[2] Bold values correspond to the best-performing model for each MSA class.

We first compared the prediction accuracy of the trained SSEmb model to the original MSA Transformer model on the ProteinGym set and found improved accuracy ($\langle\rho_s\rangle$ of 0.45 vs. 0.43) (Table 2 and Supplementary Fig. 1). In line with the design of SSEmb, this difference is greatest for proteins with the shallowest MSAs ($\langle\rho_s\rangle$ of 0.45 vs. 0.39) (Table 2 and Fig. 2). We also compared SSEmb to other high-accuracy variant effect prediction methods, and find that it generally compares favorably when benchmarked using ProteinGym, in particular for the low-MSA-depth proteins (Table 2).

https://doi.org/10.1038/s41467-024-53982-z

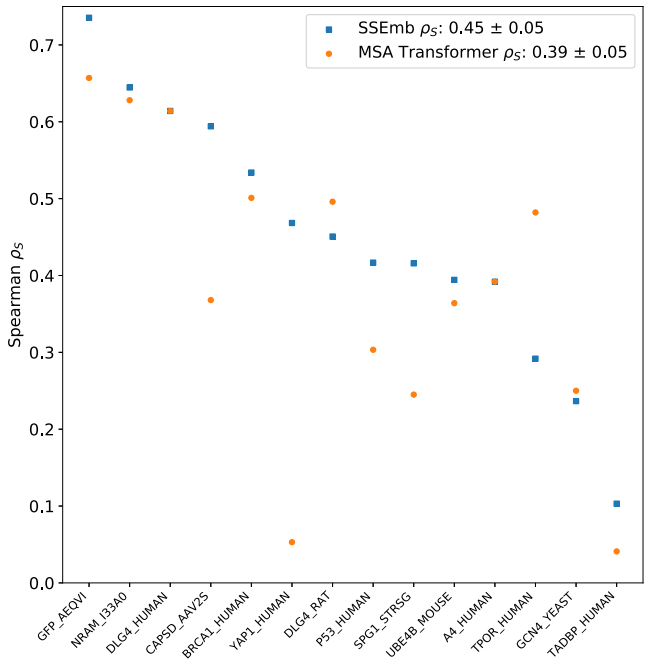

Fig. 2 | Overview of SSEmb results on the ProteinGym low-MSA ($N_{eff}/L < 1$) substitution benchmark subset grouped by UniProt ID. Spearman correlations are plotted for both SSEmb (blue) and the MSA Transformer ensemble (orange). The mean and standard error of the mean of the set of all ProteinGym Spearman correlations are presented in the legend. Assays from the SSEmb validation set have been excluded from the original data set. Source data are provided as a Source Data file.

As the increased accuracy of SSEmb at low MSA depth relies on the protein structure, we investigated the robustness of SSEmb to protein structure quality by comparing results obtained with experimental structures with those from AlphaFold (Supplementary Fig. 4). We find that SSEmb is robust to protein structure quality with only a weak correlation between performance and TM-scores of the protein structures used as input. We hypothesize that this robustness can be attributed to the backbone-based representation of the protein structure in the model as well as the complementary information contained within the MSAs.

Prediction of protein stability

Because of the tight interplay between protein stability, sequence conservation, and cellular protein abundance[8,9,54–58], we hypothesized that SSEmb would also be useful as a predictor of protein stability. We, therefore, tested the zero-shot performance of SSEmb on the recently described mega-scale measurements of protein stability[59]. We find that SSEmb achieves an absolute Spearman correlation coefficient of 0.61 (Supplementary Fig. 2), comparable to dedicated methods for protein stability predictions[60]. This result demonstrates that SSEmb can be used as a zero-shot predictor of protein stability, and additional accuracy could possibly be achieved by further supervised learning[61].

Classification of disease-causing variants

Loss of protein function and stability have been shown to correlate with expert-curated annotations of disease-causing variants in humans[15,62–66]. With some method-to-method variation, it has previously been shown that methods that perform well in predicting variant effects when compared to data from MAVEs are also more accurate in predicting variant pathogenicity[48,67]. Since we find that SSEmb performs well as a zero-shot predictor of changes in protein function and protein stability, we, therefore, hypothesized that it could also be useful as a zero-shot predictor of variant pathogenicity. We therefore evaluated the performance of SSEmb on a large set of variants with clinical annotations[67] and found that SSEmb overall performs relatively well compared to other variant effect prediction methods (Table 3). We also note subtle differences in the ranking between evaluating MAVEs (Table 2) and pathogenicity (Table 3), suggesting differences between the two tasks.

Table 3 | SSEmb performance on the ProteinGym clinical substitution benchmark compared to other variant effect prediction models[67]

Model	Avg. AUC (↑)
TranceptEVE L	0.920
GEMME	0.919
EVE	0.917
SSEmb	0.893
ESM-1b	0.892

Avg. AUC is computed as the Area Under the ROC Curve averaged across genes.

Prediction of protein-protein binding sites using embeddings

Protein-protein interactions (PPIs) are essential for cellular signaling and function. As such, mutations in PPI sites can often lead to disease[68–70] and targeting protein-protein binding sites via pharmacological drugs is an ongoing area of research[71,72]. Often, the protein-protein binding site consists of a small set of evolutionarily conserved surface residues that are essential for binding affinity[73]. Because these binding site residues are characterized by a combination of structural features (surface proximity) and sequence features (evolutionary conservation), we hypothesized that the SSEmb model should contain some information relevant to the identification of these binding sites in its embeddings. Indeed, we have recently shown that a broader class of functional sites in proteins could be identified by a combined analysis of protein stability and conservation[12].

In order to test this hypothesis, we train a small supervised downstream model to predict protein binding sites from the SSEmb embeddings using PPI training and test data[74]. Specifically, the downstream model takes as input the embeddings from the last layer in SSEmb and was trained to classify each residue as either belonging to a binding site or not using an attention mechanism (see "Methods" for further details). We compare our results to the state-of-the-art ScanNet model, which has been specifically developed for this classification task, as well as an xgboost baseline model with handcrafted structure- and sequence-based features[74]. We evaluate the models using the area under the precision-recall curves (PR-AUC) and find that the SSEmb downstream model performs in between the problem-specific ScanNet and baseline models across five different test sets with varying degrees of similarity to the training data (Table 4). These results provide proof of the principle that the SSEmb embeddings contain a rich mix of structure- and sequence-based information, which may serve as useful features for binding site prediction and other downstream tasks.

Ablation study

We performed an ablation study to investigate the importance of selected design choices (Supplementary Table 1). We thus analyzed the performance of various models using the originally released ProteinGym MAVE substitution benchmark, and focused on three components of SSEmb, which we used to inject structure information into the MSA Transformer model: (i) the GVP-GNN module after the MSA Transformer, (ii) structure-based row attention masking in the MSA Transformer, and (iii) fine-tuning of the MSA Transformer with column masking, which reduces reliance on conservation-based

Table 4 | Using the SSEmb embeddings to study protein-protein interactions

Model	PR-AUC (↑)				
	Test set (70%)	Test set (homology)	Test set (topology)	Test set (none)	Test set (all)
SSEmb downstream	0.684	0.651	0.672	0.571	0.642
Handcrafted features baseline	0.596	0.567	0.568	0.432	0.537
ScanNet	0.732	0.712	0.735	0.605	0.694

The results show the PR-AUC for our supervised downstream model compared to ScanNet and a baseline model across five different test sets. All training and test sets as well as performance metrics for ScanNet and the handcrafted-features baseline model are from[74].

signals. In particular, we assessed which components of the model gave rise to good agreement with the MAVE data at both low and high MSA depths (small values of $\Delta_{\text{High-Low}}$ in Supplementary Table 1). We find that, in particular, structure-based masking of the MSA Transformer and fine-tuning of the MSA Transformer with column masking contribute to decreasing the sensitivity of the final model to MSA depth. We also find that while the fine-tuned MSA transformer without a structural component performs best overall, it does so at the cost of accuracy for low-MSA-depth proteins. We note that the fully ablated SSEmb model outperforms the original MSA Transformer as implemented in the ProteinGym benchmark (comparing Table 2 and Supplementary Fig. S1). We hypothesize that this difference can be explained by the MSA-generation protocol used in SSEmb as well as our use of ensembling over MSA subsamples during inference, which mimics similar subsampling strategies in other well-performing variant effect prediction models[27].

Perspectives and limitations

The SSEmb design is based on the general idea of augmenting a protein language model with structure information. At a high level, our design consists of two main components: (i) a structure-constrained protein language model, which generates sequence-based embeddings, and (ii) a graph model, which processes the combined sequence- and structure-based information. In our implementation, we selected the MSA Transformer[41] as the protein language model and a modified version of GVP-GNN[42] as the graph model. However, alternative implementations could have been made, for example, by using ESM-1b[75] as the protein language model as a replacement for the MSA Transformer. Indeed, other studies have recently described implementing variations of this general idea with the goal of combining sequence- and structure-based information for variant effect prediction or other protein-based tasks[39,40,76,77]. Our results are in line with the general trends, which together indicate that augmenting protein language models with structure information can be a useful strategy for improving predictions across a wide range of tasks.

While our work overall demonstrates improved accuracy and the strength of integrating structure and sequence, some limitations remain. First, although we have shown that SSEmb performs well on variant effect prediction tasks relative to similar methods, the absolute value of correlations obtained is still relatively modest. We believe that better use of supervised methods, perhaps used in combination with self-supervised methods[78], could improve this in the future. Second, our model relies on the input of both a (subsampled) MSA and a protein structure. Although we have shown that our model works well even using shallow MSAs and predicted protein structures, we expect our model to be sensitive to these inputs. This limits the ability of our model to deal with cases where these inputs can not be guaranteed to be reliable such as for intrinsically disordered proteins or for protein complexes without experimentally resolved structures. Third, even though our model has been trained on a relatively large data set of protein sequences and structures, this training data represents only a small fraction of sequence-structure space. As such, we expect our model to suffer from degrading performance when making predictions for proteins that are very different from those found in our training data, such as certain types of de-novo-designed proteins. Overall, we show that our model is robust and useful across several types of problems but also that it may not always achieve a state of accuracy compared to models developed specifically for individual purposes.

Discussion

We have here presented a method for integrating information about protein sequence, conservation and structure in a single computational model. SSEmb uses a graph featurization of the protein structure both to constrain and integrate information from the corresponding MSA. Our results show that adding structural information to a pre-trained MSA-based model increases the ability of the model to predict variant effects in cases where the MSA is either lacking or shallow. We find that the embeddings learned by SSEmb during training contain information useful for downstream models. As an example, we show how a relatively simple downstream model trained with SSEmb embeddings as input is able to predict protein-protein binding sites. We hope that SSEMb will serve as a useful tool for studying how the integration of sequence- and structure-based protein information can improve computational predictions of variant effects and could, for example, be used to disentangle mechanistic aspects of variant effects[66].

Methods

Multiple sequence alignments

MSAs were generated using MMSeqs2[44] in combination with filtering via sequence identity buckets as implemented in ColabFold[79], which aims to maximize the diversity of sequences in the final alignment. This protocol to generate MSAs has been shown to work well for variation effect prediction using the GEMME model[80]. We use the original ColabFold Search protocol with the parameters `–diff = 512`, `–filter-min-enable = 64`, `–max-seq-id = 0.90`. Furthermore, we add the parameter `–cov = 0.75` to each sequence identity bucket in order to ensure that we only retrieve high-coverage sequences for the generated MSAs.

Subsampling of multiple sequence alignments

We randomly subsample the full MSA before using it as input to SSEmb in order to make the model train with GPU memory constraints. During training, the number of subsampled MSA sequences is set to 16. We explored how variants affect prediction performance on the validation set scales with MSA subsampling depth and ensembling. We find SSEmb is more robust to MSA sequence depth compared to the original MSA Transformer[16]. Furthermore, we find that ensembles of shallow MSAs outperform single-use or ensembles of deeper MSAs (Supplementary Fig. 3). In our case, ensembles are created by subsampling multiple times and taking the mean over the final variant effect predictions.

Based on these results, we used 16 subsampled MSA sequences in an ensemble of 5 during model inference.

Structure-constrained MSA Transformer

The structure-constrained MSA Transformer model used in SSEmb is based on the original architecture[41]. At initialization, we use the pre-trained model weights (https://github.com/facebookresearch/esm/tree/main/esm). We modify the original MSA Transformer by applying a binary contact mask to the attention maps going across MSA columns (i.e., row attention) before normalization. The contact mask corresponds to the 20 nearest neighbor graph structures used in the SSEmb GNN module, ensuring that row attention values are only propagated for positions that are spatially close in the three-dimensional protein structure. During training, we only fine-tune the row attention layers of the structure-constrained MSA Transformer in order to conserve the phylogenetic information encoded in the column attention layers[81].

GNN module

The SSEmb GNN module follows the architecture of the GVP model[42], with a few important adjustments: (i) Graph edges were defined for the 20 closest node neighbors instead of 30 in the original implementation, (ii) Node embedding dimensions were increased to 256 and 64 dimensions for the scalar and vector channels respectively. Edge embedding dimensions were kept at 32 and 1 as in the original implementation, (iii)The number of encoder and decoder layers was increased to four, (iv) we used vector gating[29], (v) the MSA query sequence embeddings from the structure-constrained MSA Transformer were concatenated to the node embeddings of the GNN decoder and passed through a dense layer to reduce dimensionality, and (vi) the models prediction task was changed from auto-regressive sequence prediction to a masked token prediction task. Further details are described below in the 'Model training' section.

Model training

During SSEmb training, we randomly mask amino acids in the wild-type sequence following a modified BERT masking scheme[82]. Before each forward pass, 15% of all wild-type sequence residues are selected for optimization. In this set of residues, 60% are masked, 20% are masked together with the corresponding MSA columns, 10% are replaced by a random residue type, and 10% are left unchanged. After masking, the SSEmb model is tasked to predict the amino acid types of the masked residues, given the protein structure and a subsampled MSA input. The masked prediction task is optimized using the cross-entropy loss between the selected 15% wild-type amino acid types and the corresponding predicted amino acid types. We trained SSEmb using the gradual unfreezing method in two steps[83]. First, we trained the GNN module until the model was close to convergence according to the training loss while keeping the parameters in the structure-constrained MSA Transformer frozen. Second, we unfreeze the row attention parameters in the structure-constrained MSA Transformer and fine-tune both the GNN module and the structure-constrained MSA Transformer using early stopping as assessed by mean correlation performance on the MAVE validation set. The training was performed using the Adam optimizer[84] with a learning rate of 10^{-3} for the GNN module and 10^{-6} for the structure-constrained MSA Transformer, respectively. Batch sizes were fixed at 128 and 2048 proteins for the two training stages, respectively.

Variant effect prediction

At inference, we randomly subsample 16 sequences from the full MSA five times with replacement in order to generate an ensemble of model predictions. The final SSEmb score is computed as the mean of the scores from this ensemble. Protein variant scores were computed according to the masked marginal method[16] from:

$$\sum_{i \in M} \log p(x_i = x_i^{\text{var}} | x_{\text{-M}}) - \log p(x_i = x_i^{\text{wt}} | x_{\text{-M}}) \quad (1)$$

where x^{var} and x^{wt} represent the variant (mutant) and wild-type sequences, and $x_{\text{-M}}$ represents a sequence where the set of substituted (mutated) positions M have been filled with mask tokens. The above model represents an additive variant effect model.

Predicted protein structures for ProteinGym benchmarks

We used AlphaFold[85] to predict structures for the proteins in the ProteinGym DMS and clinical substitution benchmarks. In practice, we used the ColabFold implementation[79] with default settings. Due to compute constraints, predicted structures were used without relaxation. For each sequence, we selected the predicted protein structure with the highest rank as input to SSEmb.

All protein structures used as input to the SSEmb model during testing were pre-processed using the OpenMM PDBFixer package[86]. The protein structures in the training set were used without modifications.

Rosetta protocol

Rosetta $\Delta\Delta G$ values were computed using the Cartesian $\Delta\Delta G$ protocol[51] and the Rosetta version with GitHub SHA1 224ebc0d2d0677ccdd9af42a54461d09367b65b3. Thermodynamic stability changes in Rosetta Energy Units were converted to a scale corresponding to kcal/mol by dividing with 2.9[51].

GEMME protocol

GEMME scores were computed using default settings[27]. MSAs for the GEMME input were generated as previously described[7] using HHblits version 2.0.15[87] to search the UniRef30 database with settings: `-e 1e-10 -i 1 -p 40 -b 1 -B 20000`. We applied two additional filters to the HHblits output MSA before using them as input to GEMME. The first filter removes all the positions (columns) that are not present in the query sequence and the second filter removes all the sequences (rows) where the number of total gaps exceeds 50%. We note that other ways of constructing MSAs may improve the accuracy of GEMME[80]; in the comparison to the ProteinGym benchmark (Table 2), we therefore used data directly from ProteinGym.

Filtering of mega-scale protein stability data set

We tested the accuracy of of SSEmb in zero-shot predictions of changes in protein stability using data set 3 from[59], which consists of experimentally well-defined $\Delta\Delta G$ measurements for a total of 607,839 protein sequences. The data set was used with minor filtering. First, our model is focused on predicting the effects of non-synonymous mutations, and we, therefore, removed all synonymous, insertion, and deletion mutations. Second, we discarded a total of 75 protein domains for which no corresponding AlphaFold model was included in the original data set.

Downstream model for protein-protein binding site prediction

We used the PPBS data set[74] for protein-protein binding site predictions. The original data set contains a total of 20,025 protein chains with binary residue-level binding site labels. We filtered the data set to exclude protein chains that were marked as obsolete in the RCSB Protein Data Bank, where the protein chain had missing binding site labels or where the amino acid sequence in the structure did not match the sequence in the label data. Lastly, we removed protein chains that were longer than the 1024 amino acid sequence limit imposed by the MSA Transformer. The total number of protein chains in the modified PPBS data set is 19,264.

The SSEmb downstream model consisted of a small transformer encoder model, which follows the original encoder implementation[88]. The downstream model takes as input the last-layer 256-dimensional residue-level embedding vector from the SSEmb model as the only input. In order to compute all embeddings in a reasonable time frame, we did not mask any positions in the amino acid sequence inspired by the wild-type marginal method as previously described[16]. In this method, the SSEmb embeddings for a sequence of length L are generated using the wild-type sequence as input to the SSEmb model and extracting the last-layer SSEmb representation of the sequence corresponding to a matrix with dimensions $L \times 256$.

This method is much faster than the masked marginal method as it only needs a single forward pass to compute embeddings for all positions in the sequence, and it has been shown to approximate the masked marginal method well for variant effect prediction[16]. Positional encodings are added using sine and cosine functions and attention is applied across the entire amino acid sequence. The number of hidden dimensions is kept fixed at 256 and the sequence is processed using three attention layers with three attention heads each. The downstream model is optimized in a supervised manner on the training data[74] using a binary cross entropy loss and the Adam optimizer[84] with a learning rate of 10^{-4} and a batch size of ten proteins. The accuracy of the final model was evaluated using the area under the precision-recall curves (PR-AUC) (Table 4 and Supplementary Fig. 5)

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data to repeat our analyses are available via https://github.com/KULL-Centre/_2023_Blaabjerg_SSEmb and https://doi.org/10.5281/zenodo.12798018. Source data are provided with this paper.

Code availability

Scripts and data to repeat our analyses are available via https://github.com/KULL-Centre/_2023_Blaabjerg_SSEmb and https://doi.org/10.5281/zenodo.12798018, where we also provide access to a version of SSEmb via Google Colab.

References

1. Fowler, D. M. et al. An atlas of variant effects to understand the genome at nucleotide resolution. *Genome Biol.* **24**, 147 (2023).
2. Freschlin, C. R., Fahlberg, S. A. & Romero, P. A. Machine learning to navigate fitness landscapes for protein engineering. *Curr. Opin. Biotechnol.* **75**, 102713 (2022).
3. Kinney, J. B. & McCandlish, D. M. Massively parallel assays and quantitative sequence-function relationships. *Annu. Rev. Genom. Hum. Genet.* **20**, 99–127 (2019).
4. Rubin, A. F. et al. Mavedb v2: a curated community database with over three million variant effects from multiplexed functional assays. Preprint at https://doi.org/10.1101/2021.11.29.470445 (2021).
5. Tabet, D., Parikh, V., Mali, P., Roth, F. P. & Claussnitzer, M. Scalable functional assays for the interpretation of human genetic variation. *Annu. Rev. Genet.* **56**, 441–465 (2022).
6. Notin, P. et al. Tranception: protein fitness prediction with autoregressive transformers and inferencetime retrieval. In *International Conference on Machine Learning* 16990–17017 (PMLR, 2022).
7. Høie, M. H., Cagiada, M., Frederiksen, A. H. B., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep.* **38**, 110207 (2022).
8. Matreyek, K. A. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
9. Cagiada, M. et al. Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. *Mol. Biol. Evol.* **38**, 3235–3246 (2021).
10. Chiasson, M. A. et al. Multiplexed measurement of variant abundance and activity reveals vkor topology, active site and human variant impact. *Elife* **9**, e58026 (2020).
11. Faure, A. J. et al. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
12. Cagiada, M. et al. Discovering functionally important sites in proteins. *Nat. Commun.* **14**, 4175 (2023).
13. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
14. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
15. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
16. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv. Neural Inf. Process. Syst.* **35**, 29287–29303 (2021).
17. Pucci, F., Schwersensky, M. & Rooman, M. Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Curr. Opin. Struct. Biol.* **72**, 161–168 (2022).
18. Notin, P. et al. TranceptEVE: combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.07.519495 (2022).
19. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science* **381**, eadg7492 (2023).
20. Diaz, D. J., Kulikova, A. V., Ellington, A. D. & Wilke, C. O. Using machine learning to predict the effects and consequences of mutations in proteins. *Curr. Opin. Struct. Biol.***78**, 102518 (2023).
21. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
22. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
23. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
24. Lui, S. & Tiana, G. The network of stabilizing contacts in proteins studied by coevolutionary data. *J. Chem. Phys.* **139**, 155103 (2013).
25. Nielsen, S. V. et al. Predicting the impact of lynch syndrome-causing missense mutations from structural calculations. *Plos Genet.* **13**, e1006739 (2017).
26. Hopf, T. A. et al. Mutation effects predicted from sequence covariation. *Nat. Biotechnol.* **35**, 128–135 (2017).
27. Laine, E., Karami, Y. & Carbone, A. Gemme: A simple and fast global epistatic model predicting mutational effects. *Mol. Biol. Evol.* **36**, 2604–2619 (2019).
28. Boomsma, W. & Frellsen, J. Spherical convolutions and their application in molecular modelling. *Advances in Neural Information Processing Systems*, (eds Guyon, I. et al.) Vol. 30 (Curran Associates, Inc., 2017).
29. Jing, B., Eismann, S., Soni, P. N. & Dror, R. O. Equivariant graph neural networks for 3d macromolecular structure. Preprint at https://doi.org/10.48550/arXiv.2106.03843 (2021).
30. Hsu, C. et al. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning* 8946–8970 (PMLR, 2022).
31. Strokach, A., Lu, T. Y. & Kim, P. M. Elaspic2 (el2): Combining contextualized language models and graph neural networks to predict effects of mutations. *J. Mol. Biol.* **433**, 166810 (2021).

32. Nguyen, V. T. D. & Hy, T. S. Multimodal pretraining for unsupervised protein representation learning. *Biol Methods Protoc.* **9**, bpae043 (2024).
33. Mansoor, S., Baek, M., Madan, U. & Horvitz, E. Toward more general embeddings for protein design: Harnessing joint representations of sequence and structure. Prepeint at *bioRxiv* https://doi.org/10.1101/2021.09.01.458592 (2021).
34. Wu, F., Radev, D. & Xu, J. When geometric deep learning meets pretrained protein language models. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.05.522958 (2023).
35. Wang, Z. et al. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Sci. Rep.* **12**, 6832 (2022).
36. Yang, k. K., Zanichelli, N. & Yeh, H. Masked inverse folding with sequence transfer for protein representation learning. *Protein Eng Des Sel.* **36**, gzad015 (2023).
37. Chen, L. et al. Learning protein fitness landscapes with deep mutational scanning data from multiple sources. *Cell Syst.* **14**, 706–721.e5 (2023).
38. Zhang, Z. et al. A systematic study of joint representation learning on protein sequences and structures. Preprint at https://doi.org/10.48550/arXiv.2303.06275 (2023).
39. Boadu, F., Cao, H. & Cheng, J. Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function. *Bioinformatics* **39**, i318–i325 (2023).
40. Li, G., Yao, S. & Fan, L. Prostage: Predicting effects of mutations on protein stability by using protein embeddings and graph convolutional networks. *J. Chem. Inf. Model.* **64**, 340–347 (2024).
41. Rao, R. et al. Msa transformer. *International Conference on Machine Learning,* **139**, (2021).
42. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L. & Dror, R. Learning from protein structure with geometric vector perceptions. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2009.01411 (2021).
43. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. *Deep Generative Models for Highly Structured Data, Dgs@iclr 2019 Workshop* (2019).
44. Steinegger, M. & Søding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
45. Paul, S., Kollasch, A., Notin, P. & Marks, D. Combining structure and sequence for superior fitness prediction. https://openreview.net/forum?id=8PbTU4exnV (2023).
46. Kulikova, A. V. et al. Two sequence- and two structure-based ml models have learned different aspects of protein biochemistry. *Sci. Rep.* **13**, 13280 (2023).
47. Jiang, R. J. *Exhaustive Mapping of Missense Variation in Coronary Heart Disease-Related Genes*. MSc thesis, University of Toronto (Canada) (2019).
48. Livesey, B. J. & Marsh, J. A. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol. Syst. Biol.* **19**, e11474 (2023).
49. Brenan, L. et al. Phenotypic characterization of a comprehensive set of mapk1/erk2 missense mutants. *Cell Rep.* **17**, 1171–1183 (2016).
50. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020).
51. Park, H. et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
52. Frenz, B. et al. Prediction of protein mutational free energy: Benchmark and sampling improvements increase classification accuracy. *Front. Bioeng. Biotechnol.* **8**, 558247 (2020).
53. Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. *Protein Sci.* **32**, e4688 (2023).
54. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci.* **102**, 14338–14343 (2005).
55. Yen, H.-C. S., Xu, Q., Chou, D. M., Zhao, Z. & Elledge, S. J. Global protein stability profiling in mammalian cells. *Science* **322**, 918–923 (2008).
56. Serohijos, A. W., Rimas, Z. & Shakhnovich, E. I. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* **2**, 249–256 (2012).
57. Suiter, C. C. et al. Massively parallel variant characterization identifies nudt15 alleles associated with thiopurine toxicity. *Proc. Natl. Acad. Sci. USA* **117**, 5394–5401 (2020).
58. Bédard, C., Cisneros, A. F., Jordan, D. & Landry, C. R. Correlation between protein abundance and sequence conservation: what do recent experiments say? *Curr. Opin. Genet. Dev.* **77**, 101984 (2022).
59. Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
60. Blaabjerg, L. M. et al. Rapid protein stability prediction using deep learning representations. *Elife* **12**, e82593 (2023).
61. Dieckhaus, H., Brocidiacono, M., Randolph, N. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc Natl Acad Sci.* **121**, e2314853121 (2024).
62. Landrum, M. J. et al. Clinvar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
63. Stein, A., Fowler, D. M., Hartmann-Petersen, R. & Lindorff-Larsen, K. Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem. Sci.* **44**, 575–588 (2019).
64. Jepsen, M. M., Fowler, D. M., Hartmann-Petersen, R., Stein, A. & Lindorff-Larsen, K. Classifying disease-associated variants using measures of protein activity and stability. *Protein Homeostasis Diseases: Mechanisms and Novel Therapies* 91–107 (2020).
65. Backwell, L. & Marsh, J. A. Diverse molecular mechanisms underlying pathogenic protein mutations: Beyond the loss-of-function paradigm. *Annu. Rev. Genom. Hum. Genet.* **23**, 475–498 (2022).
66. Cagiada, M., Jonsson, N. & Lindorff-Larsen, K. Decoding molecular mechanisms for loss of function variants in the human proteome. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.21.595203 (2024).
67. Notin, P. et al. Proteingym: large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems* 36 (NIPS, 2024).
68. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
69. Mosca, R. et al. dsysmap: exploring the edgetic role of disease mutations. *Nat. Methods* **12**, 167–168 (2015).
70. Cheng, F. et al. Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat. Genet.* **53**, 342–353 (2021).
71. Winter, A. et al. Biophysical and computational fragment-based approaches to targeting protein-protein interactions: applications in structure-guided drug discovery. *Q. Rev. Biophys.* **45**, 383–426 (2012).
72. Scott, D. E., Bayly, A. R., Abell, C. & Skidmore, J. Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. *Nat. Rev. Drug Discov.* **15**, 533–50 (2016).
73. Teppa, E., Zea, D. J. & Marino Buslje, C. Protein-protein interactions leave evolutionary footprints: High molecular coevolution at the core of interfaces. *Protein Sci.* **26**, 2438–2444 (2017).
74. Tubiana, J., Schneidman-Duhovny, D. & Wolfson, H. J. Scannet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat. Methods* **19**, 730–739 (2022).

75. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA* **118**, e2016239118 (2021).
76. Jha, K., Saha, S. & Singh, H. Prediction of protein-protein interaction using graph neural networks. *Sci. Rep.* **12**, 8360 (2022).
77. Ceccarelli, F., Giusti, L., Holden, S. B. & Liò, P. Neural embeddings for protein graphs. Preprint at https://doi.org/10.48550/arXiv.2306.04667 (2023).
78. Blaabjerg, L. M. et al. Rapid protein stability prediction using deep learning representations. *Elife* **12**, https://doi.org/10.7554/elife.82593 (2023).
79. Mirdita, M. et al. Colabfold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
80. Abakarova, M., Marquet, C., Rera, M., Rost, B. & Laine, E. Alignment-based protein mutational landscape prediction: doing more with less. *Genome Biol Evol.* **15**, evad201 (2023).
81. Lupo, U., Sgarbossa, D. & Bitbol, A.-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nat. Commun.* **13**, 6298 (2022).
82. Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at https://doi.org/10.48550/arXiv.1810.04805 (2019).
83. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification. Preprint at https://doi.org/10.48550/arXiv.1801.06146 (2018).
84. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at https://doi.org/10.48550/arXiv.1412.6980 (2014).
85. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
86. Eastman, P. et al. Openmm 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory. Comput.* **9**, 461–469 (2013).
87. Remmert, M., Biegert, A., Hauser, A. & Söding, J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat. Methods* **9**, 173–175 (2011).
88. Vaswani, A. et al. *Attention is all you need. Advances in Neural Information Processing Systems*, (eds Guyon, I. et al.) Vol. 30, 5998–6008 (Curran Associates, Inc., 2017).

Acknowledgements

Our research is supported by the PRISM (Protein Interactions and Stability in Medicine and Genomics) center funded by the Novo Nordisk Foundation (NNF18OC0033950 to A.S. and K.L.L.), and by grants from the Carlsberg Foundation (CF21-0392 to K.L.L.), Novo Nordisk Foundation (NNF20OC0062606 and NNF18OC0052719 to W.B.) and the Lundbeck Foundation (R272-2017-4528 to A.S.).

Author contributions

L.M.B., W.B., A.S., and K.L.L. conceived the overall study. N.J. performed and analyzed results from variant effect prediction benchmarks. L.M.B. wrote the first draft of the manuscript with input from W.B., A.S., and K.L.L. All authors contributed to the writing of the manuscript.

Competing interests

K.L.L. holds stock options and is a consultant for Peptone Ltd. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53982-z.

Correspondence and requests for materials should be addressed to Wouter Boomsma, Amelie Stein or Kresten Lindorff-Larsen.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.