Embeddings from protein language models predict conservation and variant effects

Céline Marquet (celine.marquet@tum.de)
Technical University Munich: Technische Universitat Munchen https://orcid.org/0000-0002-8691-5791

Michael Heinzinger
Technical University Munich: Technische Universitat Munchen

Tobias Olenyi
Technical University Munich: Technische Universitat Munchen

Christian Dallago
Technical University Munich: Technische Universitat Munchen

Michael Bernhofer
Technical University Munich: Technische Universitat Munchen

Kyra Erckert
Technical University Munich: Technische Universitat Munchen

Burkhard Rost
Technical University Munich: Technische Universitat Munchen

Research Article

Keywords: Mutation effect prediction, conservation prediction, SAV analysis, language model

Posted Date: June 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-584804/v1

Embeddings from protein language models predict conservation and variant effects

Céline Marquet [1,2*†], Michael Heinzinger [1,2†], Tobias Olenyi[1,2], Christian Dallago[1,2], Kyra Erckert[1,2], Michael Bernhofer[1,2] & Burkhard Rost[1,3]

1 TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany
2 TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Boltzmannstr. 11, 85748 Garching, Germany
3 Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany & TUM School of Life Sciences Weihenstephan (TUM-WZW), Alte Akademie 8, Freising, Germany
* Corresponding author: celine.marquet@tum.de, http://www.rostlab.org/
† Céline Marquet and Michael Heinzinger contributed equally to this work
Tel: +49-289-17-811 (email rost: assistant@rostlab.org)

Abstract

The emergence of SARS-CoV-2 variants stressed the demand for tools allowing to interpret the effect of single amino acid variants (SAVs) on protein function. While Deep Mutational Scanning (DMS) sets continue to expand our understanding of the mutational landscape of single proteins, the results continue to challenge analyses. Protein Language Models (LMs) use the latest deep learning (DL) algorithms to leverage growing databases of protein sequences. These methods learn to predict missing or marked amino acids from the context of entire sequence regions. Here, we explored how to benefit from learned protein LM representations (embeddings) to predict SAV effects. Although we have failed so far to predict SAV effects directly from embeddings, this input alone predicted residue conservation almost as accurately from single sequences as using multiple sequence alignments (MSAs) with a two-state per-residue accuracy (conserved/not) of Q2=80% (embeddings) vs. 81% (ConSeq). Considering all SAVs at all residue positions predicted as conserved to affect function reached 68.6% (Q2: effect/neutral; for PMD) without optimization, compared to an expert solution such as SNAP2 (Q2=69.8). Combining predicted conservation with BLOSUM62 to obtain variant-specific binary predictions, DMS experiments of four human proteins were predicted better than by SNAP2, and better than by applying the same simplistic approach to conservation taken from ConSeq. Thus, embedding methods have become competitive with methods relying on MSAs for SAV effect prediction at a fraction of the costs in computing/energy. This allowed prediction of SAV effects for the entire human proteome (~20k proteins) within 17 minutes on a single GPU.

Key words: Mutation effect prediction, conservation prediction, SAV analysis, language model
Abbreviations used: AUC, Area under the curve; CNN, convolutional neural network; DL, Deep Learning; DMS, deep mutational scanning; FNN, feed forward neural network; GoF, gain of function SAV; LoF, loss of function SAV; LM, Language Model; LR, logistic regression; MAVE, Multiplexed Assays of Variant Effect; MCC, Matthews correlation coefficient; ML, machine learning; MSA,

multiple sequence alignments; LM, Language Model; NLP, natural language processing; PDB, Protein Data Bank; ROC, Receiver operating characteristic; SAV, single amino-acid variant (also known as SAAV or nsSNP, or missense mutation/variant); SSD, Solid State Drive;

Introduction

Many different resources capturing SAV effects. Mutations in the Spike (S) surface protein of SARS-CoV-2 have widened the attention to the complex issue of protein variant effects (Korber et al. 2020; Laha et al. 2020; Mercatelli and Giorgi 2020; O'Donoghue et al. 2020). The ability to distinguish between beneficial (=gain of function, GoF), deleterious (=loss of function, LoF) and neutral single amino acid variants (SAVs; also referred to as SAAV, missense mutations, or non-synonymous Single Nucleotide Variants: nsSNVs) is a key aspect of understanding how SAVs affect proteins (Adzhubei et al. 2010; Bromberg and Rost 2007, 2009; Ng and Henikoff 2003; Studer et al. 2013; Wang and Moult 2001). Recently, an unprecedented amount of *in vitro* data describing the quantitative effects of SAVs on protein function has been produced through Multiplexed Assays of Variant Effect (MAVEs), such as deep mutational scans (DMS) (Fowler and Fields 2014; Weile and Roth 2018). However, a comprehensive atlas of *in vitro* variant effects for the entire human proteome still remains out of reach (AVE Alliance Founding Members 2020). Yet, even with a complete picture, intrinsic problems will remain: (1) False positives (FP): *in vitro* data capture SAV effects upon molecular function much better than those upon biological processes, e.g., disease implications may be covered in databases such as OMIM (Amberger et al. 2019), but not in MaveDB (Esposito et al. 2019). (2) False negatives (FN): The vast majority of proteins have several structural domains (Liu and Rost 2004), hence most are likely to have several different molecular functions. However, each experimental assay tends to measure the impact upon only one of those functions. (3) FP & FN: *In vivo* protein function might be impacted in several ways not reproducible by *in vitro* assays.

Evolutionary information from MSAs most important to predict SAV effects. Many *in silico* methods try to narrow the gap between known sequences and unknown SAV effects; these include (by original publication date): MutationTaster (Schwarz et al. 2010), PolyPhen2 (Adzhubei et al. 2010), SIFT (Sim et al. 2012), SNAP2 (Hecht et al. 2015), Evolutionary Action (Katsonis and Lichtarge 2014), CADD (Kircher et al. 2014), Envision (Gray et al. 2018), DeepSequence (Riesselman et al. 2018), and methods predicting rheostat positions susceptible to gradual effects (Miller et al. 2017). Of these only Envision and DeepSequence used DMS for development. Most others trained on sparsely annotated data sets such as disease-causing SAVs from OMIM (Amberger et al. 2019; McKusick-Nathans Institute of Genetic Medicine 2021), or from databases such as PMD (Kawabata et al. 1999; Nishikawa et al. 1994). While only some methods use sophisticated machine learning (ML) algorithms, almost all rely on evolutionary information derived from multiple sequence alignments (MSAs) to predict variant effect. The combination of evolutionary information and machine learning has long been established as backbone of computational biology (Rost and Sander 1993), now even allowing AlphaFold2 to predict protein structure at unprecedented levels of accuracy (Callaway 2020; Jumper et al. 2021). Nevertheless, for almost no other task is evolutionary information as crucial as for SAV effect prediction. Although different sources of information matter as input, when MSAs are available, they trump all other features (Hecht et al. 2015). Even models building on the simplest evolutionary information, e.g., the BLOSUM62 matrix condensing both biophysical constraints into a 20x20 substitution matrix (Ng and Henikoff 2003), or a simple conservation weight (Reeb et al. 2020) reach amazingly close to much more advanced methods.

Embeddings capture language of life written in proteins. Every year algorithms improve natural language processing (NLP), in particular by feeding large text corpora into Deep Learning (DL) based

Language Models (LMs). These advances have been transferred to protein sequences by learning to predict masked or missing amino acids using large databases of raw protein sequences as input (Alley et al. 2019; Elnaggar et al. 2021; Heinzinger et al. 2019; Rao et al. 2020; Rives et al. 2021). Processing the information learned by such protein LMs, e.g. by constructing 1024-dimensional vectors of the last hidden layers, yields a representation of protein sequences referred to as embeddings (Fig. 1 in (Elnaggar et al. 2021)). Embeddings have been used successfully as exclusive input to predicting secondary structure and subcellular localization at performance levels almost reaching (Alley et al. 2019; Heinzinger et al. 2019; Rives et al. 2021) or even exceeding (Elnaggar et al. 2021; Stärk et al. 2021) state-of-the-art methods using evolutionary information from MSAs as input. Embeddings can even substitute sequence similarity for homology-based annotation transfer (Littmann et al. 2021a; Littmann et al. 2021b). The power of such embeddings has been increasing with the advance of algorithms (Elnaggar et al. 2021). Naturally, there will be some limit to such improvements. However, the advances over the last months prove that this limit had not been reached by the end of 2020.

Here, we analyzed ways of using embeddings from protein LMs to predict the effect of SAVs upon protein function with a focus on molecular function, using experimental data from DMS and PMD. Although these embeddings resulted from models trained on predicting the wild-type amino acid in a sequence, at least the simple versions that we exploited failed to meaningfully capture SAV effects. Nevertheless, embeddings captured the signal for which residues were observed to be conserved without using MSA information; this enabled predicting conservation from single sequences (through a CNN). The prediction of residue conservation, in turn, enabled to predict SAV effects (equating conserved with effect). The resulting novel method was competitive with more advanced solutions at a fraction of the computational/environmental costs.

Methods

Data sets. *ConSuf10k assessed conservation:* The method predicting residue conservation used *ConSurf-DB* (Ben Chorin et al. 2020). This resource provided sequences and conservation for 89,673 proteins. For all, experimental high-resolution three-dimensional (3D) structures were available in the Protein Data Bank (PDB) (Berman et al. 2000). The ConSurf conservation scores ranged from 1 (most variable) to 9 (most conserved). The PISCES server (Wang and Dunbrack 2003) was used to redundancy reduce the data set such that no pair of proteins had more than 25% pairwise sequence identity. We removed proteins with resolutions >2.5Å, those shorter than 40 residues, and those longer than 10,000 residues. The resulting data set (ConSurf10k) with 10,507 proteins (or domains) was randomly partitioned into training (9,392 sequences), cross-training/validation (555) and test (519) sets.

PMD4k assessed SAV effects. From the SNAP2 development set (Hecht et al. 2015), we extracted annotations from PMD ("no change" as "neutral"; annotations with any level of increase or decrease in function as "effect"). This yielded 51,817 binary annotated SAVs (neutral: 13,638, effect: 38,179) in 4,061 proteins.

DMS4 sampled large-scale DMS in vitro experiments annotating SAV effects. This set contained binary classifications (effect/non-effect) for four human proteins (corresponding genes: BRAC1, PTEN, TPMT, PPARG) taken from a recent analysis (Reeb 2020). These were selected as they were the first proteins with comprehensive DMS experiments including synonymous variants (needed to map from continuous effect scores to binary effect vs. neutral) resulting in 15,621 SAV annotations (Findlay et al. 2018; Majithia et al. 2016; Matreyek et al. 2018). SAVs with beneficial effect (=gain of function) were excluded because they disagree between experiments (Reeb et al. 2020). The continuous effect scores of the four DMS experiments were mapped to binary values (effect/neutral) by considering the 95% interval around the mean of all experimental measurements as neutral, and the 5% tails of the distribution as "effect", as described in more detail elsewhere (Reeb et al. 2020). In total, the set had 11,788 neutral SAVs and 3,516 deleterious effect SAVs. Additionally, we used two other thresholds: the 90% interval from mean (8,926 neutral vs. 4,545 effect) and the 99% interval from mean (13,506 neutral vs. 1,548 SAVs effect).

Input features. All newly developed prediction methods exclusively used embeddings from pre-trained protein LMs, namely from *ProtBert* (Elnaggar et al. 2021) based on the NLP (Natural Language Processing) algorithm BERT (Devlin et al. 2019) trained on the BFD database with over 2.3 million protein sequences (Steinegger and Söding 2018), and *ProtT5-XL-U50* (Elnaggar et al. 2021) (for simplicity referred to as *ProtT5*) based on the NLP method T5 (Raffel et al. 2020) trained on BFD and fine-tuned on Uniref50 (The UniProt Consortium 2021). All embeddings were obtained from the bio_embeddings pipeline (Dallago et al. 2021). The per-residue embeddings were extracted from the last hidden layer of the models with size 1024×L, where L is the length of the protein sequence and 1024 is the size of the embedding space of ProtBert and ProtT5.

Predict conservation (ProtT5cons, Fig. 1A). Using ProtBert/ProtT5 embeddings as input (Fig. 1a), we trained three supervised classifiers to distinguish between nine *conservation classes* taken from ConSurf-DB (early stop when optimum reached for ConSurf10k validation set). The models were: (1) standard logistic regression (LR) with 9,000 (9k) free parameters; (2) feedforward neural network (FNN; with two FNN layers connected through ReLU activations (Fukushima 1969); dropout rate

0.25; 33k free parameters); (3) standard convolutional neural network (CNN with two convolutional layers with a window-size of 7, connected through ReLU activations; dropout rate of 0.25; 231k free parameters). To put the number of free parameters into perspective: the ConSurf10k data set contained about 2.7 million samples, i.e., an order of magnitude more samples than free parameters of the largest model. On top of the 9-class prediction, we implemented a binary classifier (*conserved* / *non-conserved*; threshold for projecting nine to two classes optimized on validation set). The best performing model (CNN trained on ProtT5) was referred to as ProtT5cons.

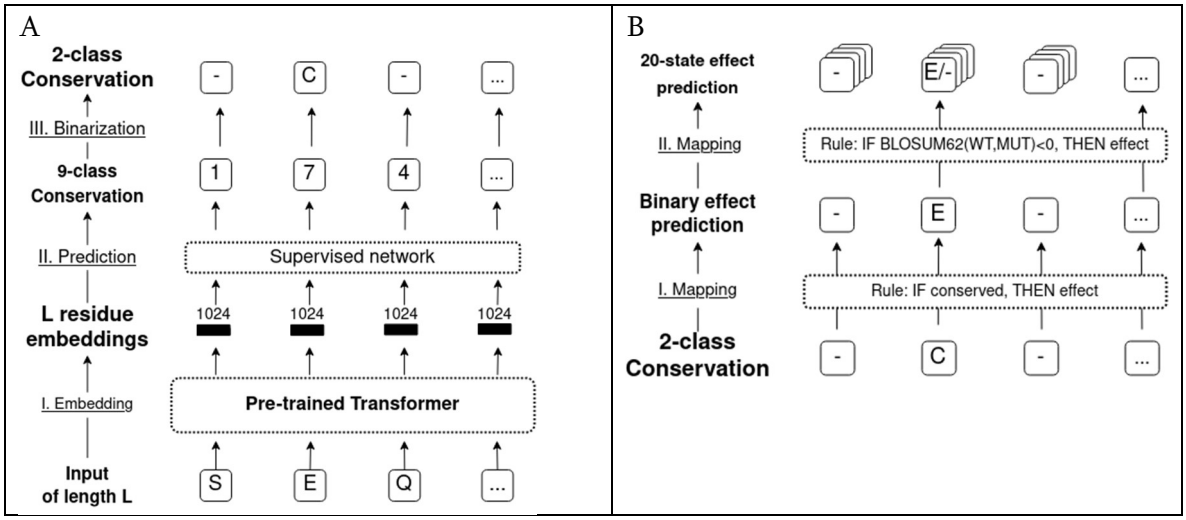

Fig. 1: Sketch of methods. Panel A sketches the conservation prediction pipeline: (I) embed protein sequence ("SEQ") using either ProtBERT or ProtT5 (Elnaggar et al. 2021). (II) Input embedding into supervised method (here: logistic regression, FNN or CNN) to predict conservation in 9-classes as defined by ConSurf (Ben Chorin et al. 2020). (III) Map nine-class predictions >5 to *conserved* (C), others to *non-conserved* (-). Panel B use binary conservation predictions to predict SAV effect prediction by (I) considering all residue-positions predicted as conserved (C) as effect (E), all others as neutral. (II) Residues predicted as conserved are further split into specific substitutions (SAVs) predicted to have an effect (E) or not (-) if the corresponding BLOSUM62 score is <0, all others are predicted as neutral (-).

Predict SAV effects (ProtT5beff, Fig. 1B). To predict SAV effects, we first ran the binary ProtT5cons method predicting conservation, and marked all residues predicted to be conserved as "effect", all others as "neutral". This first level treated all 19 non-native SAVs at one sequence position equally (referred to as "19equal" in tables and figures). To refine, we followed the lead of SIFT (Ng and Henikoff 2003) using the BLOSUM62 (Henikoff and Henikoff 1992) substitution scores. The method dubbed *BLOSUM62bin* proceeded as follows: SAVs less likely than expected (negative values in BLOSUM62) were classified as "effect", all others as "neutral". Next, we combined this with the conservation prediction and referred to the resulting method as *ProtT5beff* ("effect" if ProtT5cons predicts conserved, i.e., value >5, and BLOSUM62 negative, otherwise "neutral", Fig. 1b). This method predicted binary classifications of SAVs into effect/neutral without using any experimental data about SAV effects for optimization.

Evaluation. *Conservation - ProtT5cons:* The standard-of-truth for the conservation prediction were the values from ConSurfDB generated using HHMER (Mistry et al. 2013), CD-HIT (Fu et al. 2012), and MAFFT-LINSi (Katoh and Standley 2013) to align proteins in the PDB (Burley et al. 2019). For proteins with >50 sequences in the resulting MSA, an evolutionary rate at each residue position is computed and used with the MSA to reconstruct a phylogenetic tree. To put the performance of ProtT5cons into perspective, we generated ConSeq (Berezin et al. 2004) estimates for conservation with PredictProtein (Bernhofer et al. 2021) using MMseqs2 (Steinegger and Söding 2018) and PSI-BLAST (Altschuh et al. 1988) to generate MSAs. Furthermore, we computed a random baseline by computing metrics for our predictions against randomly shuffled ConSurfDB values.

Effect prediction - ProtT5beff: To assess to which extent the prediction of SAV effects from ProtT5cons predictions of conservation could be attributed to mistakes in ProtT5cons, we applied ConSeq using the two approaches explained above (*ConSeq 19equal*: conserved predictions at one sequence position were considered "effect" for all 19 non-native SAVs and *ConSeq blosum62*: only negative BLOSUM62 scores at residues predicted as conserved were considered "effect"; all others considered "neutral" with both using the same threshold in conservation as for our method, i.e. conservation >5 for effect) for PMD4k and DMS4. This failed for 122 proteins on PMD4k (3% of PMD4k) because the MSAs were deemed too small. We also compared ProtT5beff to the baseline based only on BLOSUM62 with the same thresholds as above (BLOSUM62bin). Furthermore, we compared to SNAP2 (at default binary threshold of effect: SNAP2>-0.05). SNAP2 failed for four of the PMD4k proteins (0.1% of PMD4k). For the random baseline, we randomly shuffled ground truth values for each PMD4k and DMS4.

Performance measures. We applied the following measures.

$$Q2 = 100 \cdot \frac{\#residues\ predicted\ correctly\ in\ 2\ states}{\#all\ residues} \quad \text{(Eqn. 1)}$$

Q2 scores (Eqn. 1) described both binary predictions (conservation and SAV effect). The same held for F1-scores (Eqn. 6, 7) and MCC (Matthews Correlation Coefficient, Eqn. 8). We defined conserved/effect as the positive class and non-conserved/neutral as the negative class (indices "+" for positive, "-" for negative) and used the standard abbreviations of TP (true positives: number of residues predicted and observed as conserved/effect), TN (true negatives: predicted and observed as non-conserved/neutral), FP (false positives: predicted conserved/effect, observed non-conserved/neutral), FN (false negatives: predicted non-conserved/neutral, observed conserved/effect).

$$Accuracy_+ = Precision_+ = Positive\ Predicted\ Value = \frac{TP}{TP+FP} \quad \text{(Eqn. 2)}$$

$$Accuracy_- = Precision_- = Negative\ Predicted\ Value = \frac{TN}{TN+FN} \quad \text{(Eqn. 3)}$$

$$Coverage_+ = Recall_+ = Sensitivity = \frac{TP}{TP+FN} \quad \text{(Eqn. 4)}$$

$$Coverage_- = Recall_- = Specificity = \frac{TN}{TN+FP} \quad \text{(Eqn. 5)}$$

$$F1_+ = 100 \cdot 2 \cdot \frac{Precision_+ \cdot Recal_+}{Precision_+ + Recal_+} \quad \text{(Eqn. 6)}$$

$$F1_- = 100 \cdot 2 \cdot \frac{Precisio_- \cdot Recall_-}{Precision_- + Recall_-} \quad \text{(Eqn. 7)}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+\quad)\cdot(TP+\quad)\cdot(TN+FP)\cdot(TN+F\quad)}} \quad \text{(Eqn. 8)}$$

$$Q9 = 100 \cdot \frac{\#residues\ predicted\ correctly\ in\ 9\ states}{\#all\ residues} \quad \text{(Eqn. 9)}$$

Q9 is exclusively used to measure performance for the prediction of nine classes of conservation taken from ConSurf.

Error estimates. We estimated symmetric 95% confidence intervals (CI, Eqn. 10) for all metrics using bootstrapping (Efron et al. 1996) by computing 1.96* standard deviation (SD) of randomly selected variants from all test sets with replacement over n = 1000 bootstrap sets:

$$CI = 1.96 \cdot SD = 1.96 \cdot \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} \quad \text{(Eqn. 10)}$$

with y_i being the metric for each bootstrap sample and $\bar{y}$ the mean over all bootstrap samples. We considered differences in performance significant if two CIs did not overlap.

Results

Embeddings predict conservation. Inputting only embeddings from ProtBert and ProtT5 into relatively simple machine learning models (Fig. 1) predicted residue conservation scores compiled from multiple sequence alignments (MSAs) by ConSurf (Ben Chorin et al. 2020): even the simplistic linear regression (LR) reached levels of performance within about 20% of what could be achieved by using the fast alignment method MMseqs2 (Steinegger and Söding 2017) for ConSeq (Berezin et al. 2004) (Fig. 2). The top prediction used ProtT5 embeddings. For both embeddings, the CNN outperformed FNN and these outperformed the LR. Either model trained on ProtT5 embeddings outperformed the respective ProtBert counterpart. Differences between ProtBert and ProtT5 were statistically significant (at the 95% confidence interval, Eqn. 10). Although the nine-state Q9 values (Eqn. 9: nine classes of conservation taken from ConSurf) were substantially smaller than the two-state Q2 values (Eqn. 1) and the improvement over randomly shuffled ConSurfDB values was higher for Q2 than for Q9 for all methods (Fig. 2, Table S1). Other performance measures ($F1_{effect}$, $F1_{neutral}$, MCC) did not change the numerical ranking and were therefore confined to the SOM (Table S1).

ConSurfDB (Ben Chorin et al. 2020) simplifies the degree of conservation by a single digit integer (9: highly conserved, 1: highly variable). The optimal threshold for a binary conservation prediction (conserved/non-conserved) was 5 (>5 conserved, Fig. S1). However, performance was stable across a wide range of choices: between values from 4 to 7, MCC (Eqn. 8) changed between 0.60 and 0.58, i.e. performance varied by 3.3% for 44.4% of all possible threshold (Fig. S1). This can be explained by the nine- and two-class confusion matrices (Fig. S2 and S3) for *ProtT5cons*, which shows that most mistakes are made between neighboring classes of similar conservation, or between the least conserved classes 1 and 2.

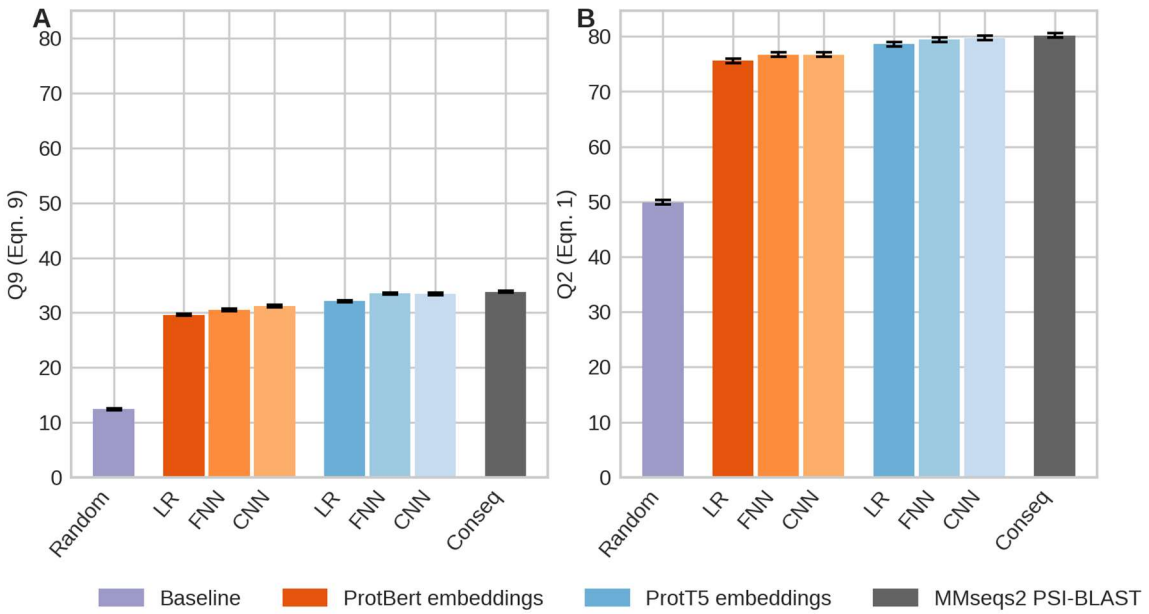

Fig. 2: Conservation predicted accurately from embeddings. Data: hold-out test set of ConSurf10k (519 sequences); panel A: nine-state per-residue accuracy (Q9, Eqn. 9) in predicting conservation; panel B: two-state per-residue accuracy (Q2, Eqn. 1). Methods: LR: logistic regression (9,000=9k free parameters), FNN: feed-forward network (33k parameters), and CNN: convolutional neural network

(231k parameters with 0.25 dropout rate, effectively giving 231/4k parameters); ConSeq: computation of conservation weight through fast multiple sequence alignments (MSAs) (Berezin et al. 2004); Random: random label swap. Model inputs are differentiated by color (red: ProtBert embeddings (Elnaggar et al. 2021), blue: ProtT5 embeddings (Elnaggar et al. 2021), grey: MSAs (MMseqs2 (Steinegger and Söding 2017) and PSI-BLAST (Altschuh et al. 1988)). Black bars mark the 95% confidence interval (±1.96 SD; Eqn. 10). Best were predictions using the ProtT5 embeddings (Elnaggar et al. 2021) and of those closest to ConSeq came the CNN-based predictions (Table S1 for more detail).

Conservation-based prediction of SAV effect seemingly better for DMS4 than for PMD4k. In using predicted conservation as proxy for SAV effect, we chose the method best in conservation prediction, namely the CNN using ProtT5 embeddings (method dubbed *ProtT5cons*, Fig. 1B). The over-simplistic approach of considering any residue predicted as conserved to have an effect on any SAV (meaning: treat all 19 non-native SAVs alike) was referred to as "19equal" in the tables. We "refined" by combining conservation prediction with a binary BLOSUM62 score (effect: if ProtT5cons predicted conserved AND BLOSUM62<0, neutral otherwise), which we referred to as *ProtT5beff*. Both solutions outperformed the simple BLOSUM62 matrix with respect to MCC on the PMD4k data set dominated by SAVs affecting molecular function (ProtT5cons 19equal and ProtT5beff vs. BLOSUM62bin in Fig. 3A, Table 1). Alternative performance measure no longer clearly correlated as well as for conservation (Table 1, Table S2). The following results were common to all measures carving aspects such as precision and recall into a single number (Q2, $F1_{effect}$, $F1_{neutral}$ and MCC). First, the expert method SNAP2 trained on some version of PMD4k achieved numerically higher values than all simplistic methods introduced here. Second, for PMD4k only, using the same SAV effect prediction for all 19 non-native SAVs consistently reached higher values than using the BLOSUM62 values (Fig. 3 and Table 1: *19equal* higher than *blosum62*). For some measures (Q2, $F1_{effect}$) values obtained by using ConSeq for conservation (i.e. a method using MSAs) were higher than those for the ProtT5cons prediction (without using MSAs), while for others (MCC, $F1_{neutral}$) this was reversed (Fig. 3, Table 1).

Table 1: Performance in binary SAV effect prediction *

Data set	*PMD4k*		*DMS4*	
Method/ Metric	Q2 (Eqn. 1)	*MCC (Eqn. 8)*	*Q2 (Eqn. 1)*	*MCC (Eqn. 8)*
Random	61.08% ± 0.41	-0.002 ± 0.016	64.27% ± 0.76	-0.001 ± 0.018
BLOSUM62bin	56.17% ± 0.43	0.049 ± 0.010	44.47% ± 0.84	0.169 ± 0.014
ProtT5cons 19equal	68.58% ± 0.41	0.227 ± 0.010	62.20% ± 0.82	0.322 ± 0.014
ProtT5beff	52.26% ± 0.43	0.160 ± 0.016	**71.47% ± 0.75**	**0.369 ± 0.016**
ConSeq 19equal	**71.51% ± 0.39**	0.206 ± 0.010	50.70% ± 0.84	0.267 ± 0.012
ConSeq blosum62	54.32% ± 0.43	0.138 ± 0.016	63.81% ± 0.8	0.318 ± 0.014
SNAP2bin	70.66% ± 0.39	**0.280 ± 0.010**	41.55% ± 0.82	0.204 ± 0.012

* **Data sets:** The *PMD4k* data set contained 4k proteins from the PMD (Kawabata et al. 1999); 74% of the SAVs were deemed effect in a binary classification. DMS4 marks the first four human proteins (BRAC1, PTEN, TPMT, PPARG) for which we obtained comprehensive experimental DMS measurements along with a means of converting experimental scores into a binary version (effect/neutral) using synonyms. DMS4 results are shown for a threshold of 95%: the continuous effect scores were binarized by assigning the middle 95% of effect scores as neutral variants and SAVs resulting in effect scores outside this range as effect variants (Reeb et al. 2020). **Methods**: *BLOSUM62bin*: negative BLOSUM62 scores predicted as effect, others as neutral; *ProtT5cons|ConSeq 19equal*: all 19 non-native SAVs predicted equally: effect if ProtT5cons|ConSeq predicted conserved, otherwise neutral; *ProtT5beff|ConSeq blosum62*: effect if ProtT5cons|ConSeq predicts conserved and BLOSUM62 negative, otherwise neutral; *SNAP2bin*: effect SNAP2-score>-0.05, otherwise neutral. ± values mark the 95% confidence interval (Eqn. 10). For each measure and data set, significantly best results are highlighted in bold.

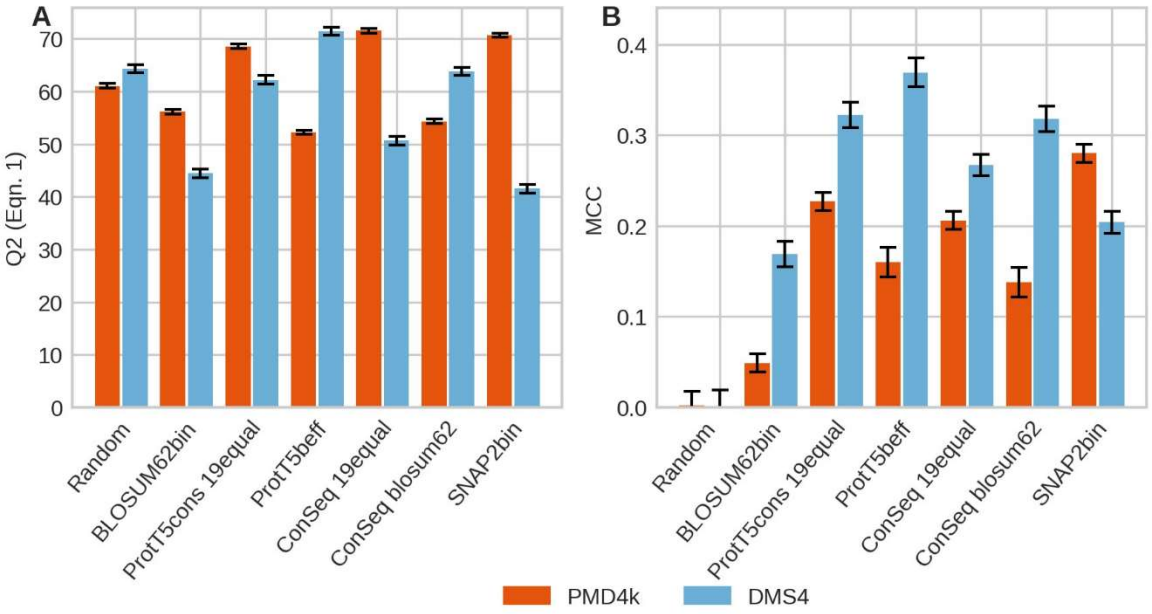

Fig. 3: Embedding-based SAV effect prediction seemingly competitive. Data: PMD4k (red bars; 4k proteins from PMD (Kawabata et al. 1999)); DMS4 (blue bars) first four proteins (BRAC1, PTEN, TPMT, PPARG) with comprehensive experimental DMS measurements including synonyms (here 95% threshold) (Reeb et al. 2020). Methods: BLOSUM62bin: negative BLOSUM62 scores predicted as effect, others as neutral; ProtT5cons|ConSeq 19equal: all 19 non-native SAVs predicted equally: effect if ProtT5cons|ConSeq predicted conserved, otherwise neutral; ProtT5beff|ConSeq blosum62: effect if ProtT5cons|ConSeq predicts conserved and BLOSUM62 negative, otherwise neutral; SNAP2bin: effect SNAP2-score>-0.05, otherwise neutral; Random: random shuffle of experimental labels. All values computed for binary mapping of experimental values (effect/neutral) with panel A giving the two-state per-residue accuracy (Q2, Eqn. 1) and panel B giving the Matthews Correlation Coefficient (MCC, Eqn. 8). Error bars: Black bars mark the 95% confidence interval (±1.96 SD, Eqn. 10). For all methods, the differences between the two data sets PMD4k and DMS4 were statistically significant (exception: random). The surprising result that the best embedding-based method

(ProtT5beff) not using MSAs reached a higher MCC (panel B) than the SAV effect prediction based on the more accurate conservation "prediction" from ConSeq.

Most relative values differed substantially between PMD4k and DMS4, i.e. the first four proteins (BRAC1, PTEN, TPMT, PPARG) for which we had obtained large-scale experimental DMS measures that could be converted into a binary scale (effect/neutral). First, all results using BLOSUM62 to convert ProtT5cons into SAV-specific predictions outperformed both the MSA-based conservation lookup from ConSeq and the expert SNAP2 trained on PMD4k (Table 1: ProtT5cons blosum62 highest). Second, combining the BLOSUM62 matrix with conservation also improved ConSeq (Table 1: ConSeq: *19equal* lower than *blosum62*). Third, different performance measures rank correlated much better than for PMD4k (Tables S1-S5). Results for DMS4 at intervals of 90% (Table S3) and 99% (Table S5) around the mean showed similar trends.

No straightforward prediction of SAV effect from LM mask probabilities. A more direct approach to predict SAV effects from embeddings is to make use of the probabilities that transformers trained on masked language modeling (e.g. ProtBert) assign to each of the 20 amino acids for masked input sequences. What we describe in the following uses only reconstruction probabilities computed from embeddings. For instance, when corrupting and reconstructing all residues in ConSurf4k (one residue at a time), ProtBert assigns a probability to the native and to each of the 19 non-native (SAVs) amino acids for each position in the protein. This way, ProtBert correctly predicted the native amino acid in 45.3% of all cases compared to a random baseline of 9.4% predicting the most frequent amino acid (Fig. S4). When analyzing the entropy of these probability distributions, we found a modest correlation with conservation (Fig. 4, Spearman's R=-0.374). In contrast, preliminary results for directly using the same reconstruction probabilities to discriminate neutral (expected: high probability) from effect SAVs (expected: low probability) did not improve over *ProtT5beff*.

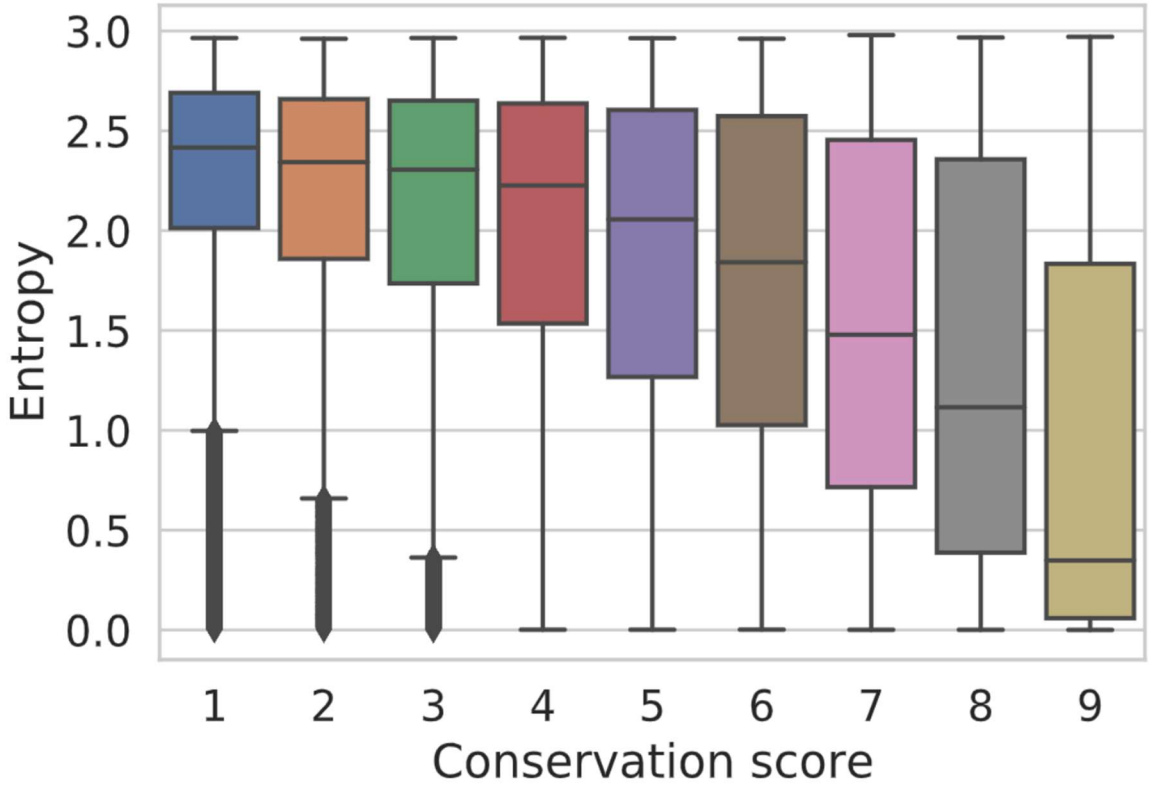

Fig. 4: Embeddings capture conservation without supervised training. Originally, ProtBert was optimized on reconstructing corrupted input tokens from non-corrupted sequence context (masked language modeling). Here, we corrupted and reconstructed, one residue at a time in the entire ConSurf10k dataset. For each residue position, ProtBert returns 20 values describing the probability for observing each of the 20 amino acids at that position. The higher the value, the probability, i.e. the lower the corresponding entropy, the more certain the protein LM predicts the corresponding amino acid. ProtBert's sequence reconstructions are more confident for more conserved residues as shown by a modest (negative) correlation between ConSurf's residue conservation and the entropy/uncertainty of ProtBert's reconstruction probabilities (Spearman's R of –0.374). Thus, ProtBert learned some aspect correlated with residue conservation during pre-training without having ever seen MSAs.

SAV effect predictions blazingly fast. One important advantage of predicting SAV effects without using MSAs is the computational efficiency. For instance, to predict the effects for all 19 non-native SAVs in the entire human proteome (all residues in all human proteins) took 20 minutes on one Nvidia Quadro RTX 8000. Although this was more than using BLOSUM62bin alone as only a simple lookup is required (nearly instantaneous), the embedding-based predictions were also much better (Table 1). In contrast, running methods such as SNAP2 (or ConSeq) requires first to generate MSAs. Even the blazingly fast MMseqs2 (Steinegger and Söding 2017) needed about 90 minutes using batch-processing on an Intel Skylake Gold 6248 processor with 40 threads, SSD and 377GB main memory. However, when deploying tools as a webserver, batch-processing might not be readily available as users submit single queries, in turn increasing runtime substantially.

Discussion

No direct prediction of SAV effect from generic embeddings. Setting out to use reconstruction probabilities from protein language models to predict the effects of SAVs upon protein function we encountered several surprises and could not fully explain the puzzle pictured by the results. On the one hand, embeddings can predict the native amino acid in a protein sequence about half of all the times (Fig. S4). Although there is no single high value threshold to apply to distinguish native from non-native (Fig. S6), this finding implied that embeddings contain information allowing to identify residue positions at which the native amino acid stands out from the 19 non-native SAVs. However, whatever constraint marked those positions in the embeddings did not suffice to predict SAV effects.

This might be related to a problem well-known in machine learning: the softmax function, which maps raw logits (output of the LM) to class probabilities, tends to assign high probabilities to a small subset of possible classes (Hinton et al. 2015). As a result, probabilities (and thus the difference between them) become extremely small for most other classes. In our case, this means that the LM is overly confident in predicting just one or a few amino acids while assigning extremely small probabilities to all others. This renders differences among the 20 values uninformative (Fig. S6). One NLP solution is to introduce a "temperature factor" in the softmax function which "softens" class probabilities (Caccia et al. 2020). This distributes the probability mass more evenly and renders difference more meaningful. A different, yet much harder to solve problem is ProtBert's tendency to over-predict frequent amino acids (Fig. S5). Like all other machine-learning devices, LMs pick up class imbalance during training (here the fact that not all 20 amino acids occur equally often). Taken together, these problems make it difficult to use simple residue masking and the resulting reconstruction probabilities without further processing or fine-tuning for SAV effect prediction.

On the other hand, it was already shown that methods using embeddings from protein LMs as input get very close or even outperform methods that use evolutionary constraints defined by MSAs as input (Elnaggar et al. 2021; Rao et al. 2020; Stärk et al. 2021). Therefore, we can currently only conclude that the simplest approach of using reconstruction probabilities from ProtBert without further processing are not readily suitable to scan the mutational landscape of proteins. Overall, it might not appear too surprising because the embeddings from the protein LMs used here (ProtBert & ProtT5 (Elnaggar et al. 2021)) could not learn evolutionary constraints directly because they never saw MSAs. This is very different for methods trained directly to extract family-specific embeddings from MSAs that apparently excel at predicting SAV effects for DMS data (Riesselman et al. 2018).

Embeddings predict conservation. Even a simple Linear Regression sufficed to predict per-residue conservation values directly from generic embeddings (Fig. 2, Table S1). Furthermore, even relatively simple CNNs (with almost 100-times fewer free parameters than samples despite early stopping) reached levels in predicting conservation values within a few percent of the result from ConSeq explicitly using MSAs (Fig. 2, Table S1). If generic embeddings have extracted enough information from raw sequence databases to predict conservation, did that imply that the protein LMs did extract evolutionary information? Unfortunately, we could not answer this question. Clearly, the methodology applied to predict conservation never encountered any explicit information about MSAs, i.e., never had an explicit opportunity to pick up evolutionary information.

Predicted conservation informative about SAV effects. DMS data sets with comprehensive experimental probing of the mutability landscape (Hecht et al. 2013) as, e.g., collected by MaveDB (Esposito et al. 2019) continue to pose problems for analysis, possibly due to a diversity of assays and protocols (Livesey and Marsh 2020; Reeb et al. 2020). Nevertheless, it seems reasonable to assume that such data sets capture some aspects about the susceptibility to change better than traditional data sets with carefully selected SAVs such as PMD4k – which incidentally have been used to optimize most methods predicting SAV effects thereby creating some circularity (Grimm et al. 2015) and overlap (Reeb et al. 2016). If true, carefully tuned studies as contained in the DMS4 data set tested here, might contain largely reliable information for experimental probing of SAV effects. In turn, our results clearly suggested that embedding-based predictions of SAV effects (by simply assigning "effect" to those SAVs where ProtT5cons predicted conserved and the corresponding BLOSUM62 value was negative) outperformed ConSeq with MSAs using the same idea, and even the expert effect prediction method SNAP2 (Fig. 3, Table 1). On top, these results were obtained on a data set never used for optimizing a single free parameter.

Strictly speaking, this might not be considered fully correct. Instead, it might be argued that one single free parameter was optimized using the data set, because for the PMD4k data set the version that predicted the same effect for all 19-SAVs appeared to outperform the SAV-specific prediction using BLOSUM62 (*19equal* vs *blosum62* in Fig. 3 and Table 1). However, not even the values computed for PMD4k could distract from the simple fact that not all SAVs are equal, i.e., that regardless of model performance, 19equal is agnostic to the sequence context in which a SAV appears. In fact, the concept of combining predictions with BLOSUM62 values has been shown to succeed for function prediction before (Bromberg and Rost 2008; Schelling et al. 2018) in that sense it was arguably not an optimizable hyperparameter.

Embeddings predicted conservation (Fig. 2); conservation predicted SAV effects (Fig. 3). Did this imply that embeddings did capture evolutionary information? Once again, we could not answer this question either way directly. To repeat: our procedure/method never used information from MSAs in any direct way. Could it have implicitly learned this? Here we enter a circle: if so, why could mask reconstruction probabilities from raw embeddings not directly predict SAV effects? Although we could not clearly answer the question, we see only one hypothesis that appears compatible with all the findings, namely that embeddings capture biophysical constraints shaping proteins and determining their function that are also reflected in profiles of evolutionary information captured in MSAs. In other words, embeddings capture a reality that constraints what can be observed in evolution, and this reality is exactly what is used for the part of the SAV effect prediction that succeeds. If so, we would argue that our simplified method did not succeed because it predicted conservation without using MSAs, but that it captured positions biophysically "marked by constraints". This assumption would explain how a partially incorrect prediction of conservation (ProtT5cons) not using evolutionary information could predict SAV effects better than a more correct prediction (ConSeq) using MSAs to extract evolutionary information (Fig. 3: ProtT5cons vs. ConSeq). In this interpretation, it is not "residue conservation" that enables the small fraction of correct SAV effect predictions but some other feature.

Fast predictions climate-change friendly. Clearly, our simple protocol introduced here will not be a contender for the best tool to predict SAV effects from sequence, not even amongst all the methods not using MSAs. However, our solution allows for extremely efficient, speedy predictions. While the pre-training of protein LMs consumes immense amounts of energy (Elnaggar et al. 2021), this has to

be accomplished only once. Applying embeddings such as from ProtT5 to predict residue conservation (ProtT5cons) or SAV effects after that consumes very little additional resources. When running prediction servers such as PredictProtein (Bernhofer et al. 2021) queried over 3,000 times every month, such investments could be recovered rapidly at seemingly small prices to pay in terms of slightly reduced performance. This might be as good a moment to consider the balance between gain in energy and loss in performance more carefully.

Conclusions

Embeddings extracted from protein Language Models (LMs, Fig. 1), namely from ProtBert and ProtT5 (Elnaggar et al. 2021), contain information that suffices to predict residue conservation in protein families without using multiple sequence alignments (Fig. 2). Such predictions combined with BLOSUM62 scores suffice to predict the effects of sequence variation (single amino acid variants, or SAVs) without optimizing any additional free parameter. Such predictions are blazingly fast, thereby they save computing, and ultimately energy resources when applied to daily sequence analysis. On top, our simple, fast solutions appeared to reach levels of performance not sufficiently worse than that for expert methods predicting variant effects (Fig. 3). This was remarkable because implicitly the probabilities computed from the same raw embeddings used to predict conservation did not contain any information relevant for predicting SAV/variant effects. In combination, our results suggested the major signal captured by variant effect predictions originates from some biophysical constraint revealed by raw protein sequences.

Acknowledgements

Thanks to Tim Karl and Inga Weise (both TUM) for invaluable help with technical and administrative aspects of this work. Thanks to Nir Ben-Tal (Tel Aviv U) and his team for the excellent services around ConSurf and ConSeq, to Yana Bromberg (Rutgers U), Max Hecht (Amazon), Jonas Reeb and Dmitirii Nechaev (both TUM) for advancing SNAP, the Dunbrack lab for Pisces, and most importantly to Martin Steinegger and his team for MMseqs2 and BFD. Last, but not least, thanks to all who deposit their experimental data in public databases, and to those who maintain these databases.

Author contributions

C.M. implemented and evaluated the methods, and took the lead in writing the manuscript. M.H. conceived, trained, and evaluated the neural networks on conservation prediction, contributed ideas and proofread the manuscript. T.O. and C.D contributed crucial ideas and provided valuable comments. M.B. helped in generating the evaluation methods ConSeq and SNAP2. K.E. supported the work with coding advice and created the original ConSurf10k data set. B.R. supervised and guided the work and co-wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the DFG grant RO 1320/4-1, Software Campus Funding (BMBF 01IS17049) and the KONWIHR Program.

Competing interests

No author declares any competing interest.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nature Methods 7: 248-249. doi: 10.1038/nmeth0410-248

Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM (2019) Unified rational protein engineering with sequence-based deep representation learning. Nature Methods 16: 1315-1322. doi: 10.1038/s41592-019-0598-1

Altschuh D, Vernet T, Berti P, Moras D, Nagai K (1988) Coordinated amino acid changes in homologous protein families*. Protein Engineering, Design and Selection 2: 193-199. doi: 10.1093/protein/2.3.193

Amberger JS, Bocchini CA, Scott AF, Hamosh A (2019) OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res 47: D1038-D1043. doi: 10.1093/nar/gky1151

AVE Alliance Founding Members (2020) Atlas of Variant Effect Alliance.

Ben Chorin A, Masrati G, Kessel A, Narunsky A, Sprinzak J, Lahav S, Ashkenazy H, Ben - Tal N (2020) ConSurf - DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. Protein Science 29: 258-267. doi: 10.1002/pro.3779

Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics (Oxford, England) 20: 1322-1324. doi: 10.1093/bioinformatics/bth070

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Research 28: 235-242. doi: 10.1093/nar/28.1.235

Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M, Olenyi T, Qiu J, Schutze K, Yachdav G, Ashkenazy H, Ben-Tal N, Bromberg Y, Goldberg T, Kajan L, O'Donoghue S, Sander C, Schafferhans A, Schlessinger A, Vriend G, Mirdita M, Gawron P, Gu W, Jarosz Y, Trefois C, Steinegger M, Schneider R, Rost B (2021) PredictProtein - Predicting Protein Structure and Function for 29 Years. Nucleic Acids Res. doi: 10.1093/nar/gkab354

Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research 35: 3823-3835.

Bromberg Y, Rost B (2008) Comprehensive in silico mutagenesis highlights functionally important residues in proteins. Bioinformatics 24: i207-i212.

Bromberg Y, Rost B (2009) Correlating protein function and stability through the analysis of single amino acid substitutions. BMC Bioinformatics 10: S8. doi: 10.1186/1471-2105-10-S8-S8

Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C, Dalenberg K, Duarte JM, Dutta S, Feng Z, Ghosh S, Goodsell DS, Green RK, Guranovic V, Guzenko D, Hudson BP, Kalro T, Liang Y, Lowe R, Namkoong H, Peisach E, Periskova I, Prlic A, Randle C, Rose A, Rose P, Sala R, Sekharan M, Shao C, Tan L, Tao YP, Valasatava Y, Voigt M, Westbrook J, Woo J, Yang H, Young J, Zhuravleva M, Zardecki C (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Research 47: D464-D474. doi: 10.1093/nar/gky1004

Caccia M, Caccia L, Fedus W, Larochelle H, Pineau J, Charlin L (2020) Language GANs Falling Short. arXiv:1811.02549 [cs].

Callaway E (2020) 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature 588: 203-204. doi: 10.1038/d41586-020-03348-4

Dallago C, Schütze K, Heinzinger M, Olenyi T, Littmann M, Lu AX, Yang KK, Min S, Yoon S, Morton JT, Rost B (2021) Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. Current Protocols 1: e113. doi: https://doi.org/10.1002/cpz1.113

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs].

Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences USA 93: 13429-13434.

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B (2021) ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. MACHINE INTELLIGENCE 14: 30.

Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM, Rubin AF (2019) MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol 20: 223. doi: 10.1186/s13059-019-1845-6

Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, Janizek JD, Huang X, Starita LM, Shendure J (2018) Accurate classification of BRCA1 variants with saturation genome editing. Nature 562: 217-222. doi: 10.1038/s41586-018-0461-z

Fowler DM, Fields S (2014) Deep mutational scanning: a new style of protein science. Nature Methods 11: 801-807. doi: 10.1038/nmeth.3027

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150-2. doi: 10.1093/bioinformatics/bts565

Fukushima K (1969) Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements. IEEE Transactions on Systems Science and Cybernetics 5: 322-333. doi: 10.1109/TSSC.1969.300225

Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM (2018) Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. Cell Systems 6: 116-124.e3. doi: 10.1016/j.cels.2017.11.003

Grimm DG, Azencott CA, Aicheler F, Gieraths U, Macarthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Human Mutation 36: 513-523. doi: 10.1002/humu.22768

Hecht M, Bromberg Y, Rost B (2013) News from the protein mutability landscape. Journal of Molecular Biology 425: 3937-3948. doi: 10.1016/j.jmb.2013.07.028

Hecht M, Bromberg Y, Rost B (2015) Better prediction of functional effects for sequence variants. BMC Genomics 16: S1. doi: 10.1186/1471-2164-16-S8-S1

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

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences 89: 10915-10919. doi: 10.1073/pnas.89.22.10915

Hinton G, Vinyals O, Dean J (2015) Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [cs, stat].

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Tunyasuvunakool K, Ronneberger O, Bates R, Zídek A, Bridgland A, Meyer C, Kohl SAA, Potapenko A, Ballard AJ, Cowie A, Romera-

Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Steinegger M, Pacholska M, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) High accuracy protein structure prediction using deep learning. Fourteenth Critical Assessment of Protein Structure Prediction (CASP14)

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-80. doi: 10.1093/molbev/mst010

Katsonis P, Lichtarge O (2014) A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. Genome Research 24: 2050-2058. doi: 10.1101/gr.176214.114

Kawabata T, Ota M, Nishikawa K (1999) The Protein Mutant Database. Nucleic Acids Research 27: 355-357. doi: 10.1093/nar/27.1.355

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics 46: 310-315. doi: 10.1038/ng.2892

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, Angyal A, Brown RL, Carrilero L, Green LR, Groves DC, Johnson KJ, Keeley AJ, Lindsey BB, Parsons PJ, Raza M, Rowland-Jones S, Smith N, Tucker RM, Wang D, Wyles MD, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC (2020) Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 182: 812-827.e19. doi: 10.1016/j.cell.2020.06.043

Laha S, Chakraborty J, Das S, Manna SK, Biswas S, Chatterjee R (2020) Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. Infection, Genetics and Evolution 85: 104445. doi: 10.1016/j.meegid.2020.104445

Littmann M, Bordin N, Heinzinger M, Schütze K, Dallago C, Orengo C, Rost B (2021a) Clustering FunFams using sequence embeddings improves EC purity Bioinformatics. doi: https://doi.org/10.1093/bioinformatics/btab371

Littmann M, Heinzinger M, Dallago C, Olenyi T, Rost B (2021b) Embeddings from deep learning transfer GO annotations beyond homology. Scientific Reports 11: 1160. doi: 10.1038/s41598-020-80786-0

Liu J, Rost B (2004) Sequence-based prediction of protein domains. Nucleic Acids Research 32: 3522-3530.

Livesey BJ, Marsh JA (2020) Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol 16: e9380. doi: 10.15252/msb.20199380

Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA, Zhang X, Broekema MF, Patterson N, Duby M, Sharpe T, Kalkhoven E, Rosen ED, Barroso I, Ellard S, Consortium UKMD, Kathiresan S, Myocardial Infarction Genetics C, O'Rahilly S, Consortium UKCL, Chatterjee K, Florez JC, Mikkelsen T, Savage DB, Altshuler D (2016) Prospective functional classification of all possible missense variants in PPARG. Nat Genet 48: 1570-1575. doi: 10.1038/ng.3700

Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, Bhatia S, Evans WE, Relling MV, Yang W, Shendure J, Fowler DM (2018)

Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet 50: 874-882. doi: 10.1038/s41588-018-0122-z

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) (2021) Online Mendelian Inheritance in Man, OMIM®.

Mercatelli D, Giorgi FM (2020) Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Frontiers in Microbiology 11. doi: 10.3389/fmicb.2020.01800

Miller M, Bromberg Y, Swint-Kruse L (2017) Computational predictors fail to identify amino acid substitution effects at rheostat positions. Sci Rep 7: 41329. doi: 10.1038/srep41329

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res 41: e121. doi: 10.1093/nar/gkt263

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research 31: 3812-3814.

Nishikawa K, Ishino S, Takenaka H, Norioka N, Hirai T, Yao T, Seto Y (1994) Constructing a protein mutant database. Protein Engineering 7: 733. doi: 10.1093/protein/7.5.733

O'Donoghue SI, Schafferhans A, Sikta N, Stolte C, Kaur S, Ho BK, Anderson S, Procter J, Dallago C, Bordin N, Adcock M, Rost B (2020) SARS-CoV-2 structural coverage map reveals state changes that disrupt host immunity. bioRxiv: 2020.07.16.207308. doi: 10.1101/2020.07.16.207308

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs, stat].

Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A (2020) Transformer protein language models are unsupervised structure learners. bioRxiv: 2020.12.15.422761. doi: 10.1101/2020.12.15.422761

Reeb J (2020) Data for: Variant effect predictions capture some aspects of deep mutational scanning experiments. 1. doi: 10.17632/2rwrkp7mfk.1

Reeb J, Hecht M, Mahlich Y, Bromberg Y, Rost B (2016) Predicted molecular effects of sequence variants link to system level of disease. PLoS Computational Biology 12: e1005047. doi: 10.1371/journal.pcbi.1005047. doi: 10.1371/journal.pcbi.1005047

Reeb J, Wirth T, Rost B (2020) Variant effect predictions capture some aspects of deep mutational scanning experiments. BMC Bioinformatics 21: 107. doi: 10.1186/s12859-020-3439-4

Riesselman AJ, Ingraham JB, Marks DS (2018) Deep generative models of genetic variation capture the effects of mutations. Nat Methods 15: 816-822. doi: 10.1038/s41592-018-0138-4

Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118. doi: 10.1073/pnas.2016239118

Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology 232: 584-599. doi: 10.1006/jmbi.1993.1413

Schelling M, Hopf TA, Rost B (2018) Evolutionary couplings and sequence variation effect predict protein binding sites. Proteins 86: 1064-1074. doi: 10.1002/prot.25585

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7: 575-6. doi: 10.1038/nmeth0810-575

Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research 40: W452-W457. doi: 10.1093/nar/gks539

Stärk H, Dallago C, Heinzinger M, Rost B (2021) Light Attention Predicts Protein Location from the Language of Life. bioRxiv: 2021.04.25.441334. doi: 10.1101/2021.04.25.441334

Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology 35: 1026.

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

Studer RA, Dessailly BH, Orengo CA (2013) Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem J 449: 581-94. doi: 10.1042/BJ20121221

The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49: D480-D489. doi: 10.1093/nar/gkaa1100

Wang G, Dunbrack RL, Jr (2003) PISCES: a protein sequence culling server. Bioinformatics 19: 1589-1591. doi: 10.1093/bioinformatics/btg224

Wang Z, Moult J (2001) SNPs, protein structure, and disease. Human Mutation 17: 263-270. doi: https://doi.org/10.1002/humu.22

Weile J, Roth FP (2018) Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. Human Genetics 137: 665-678. doi: 10.1007/s00439-018-1916-x

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ConsEmbSOM.pdf