Mansoor Sanaa (Orcid ID: 0009-0004-5992-060X)
Baker David (Orcid ID: 0000-0001-7896-6217)

RESEARCH NOTE

Protein Science
DOI10.1002/pro.4780

Full title: Zero-shot Mutation Effect Prediction on Protein Stability and Function using RoseTTAFold

Short title: Mutation Effect Prediction of Proteins using RoseTTAFold

Sanaa Mansoor[1,2,3], Minkyung Baek[1,2,4], David Juergens[1,2,3], Joseph L. Watson[1,2], David Baker[1,2,5,*]

[1] Department of Biochemistry, University of Washington, Seattle, Washington, United States of America

[2] Institute for Protein Design, University of Washington, Seattle, Washington, United States of America

[3] Molecular Engineering Graduate Program, University of Washington, Washington, United States of America

[4] School of Biological Sciences, Seoul National University, Seoul, Republic of Korea

[5] Howard Hughes Medical Institute, University of Washington, Seattle, Washington, United States of America

* Corresponding author

Corresponding author information:

Email address: dabaker@uw.edu

Mailing address: 3963 Stevens Way NE, Seattle WA 98105 USA.

Phone number: +1 (206) 310-7409

Received: Apr 12, 2023;Revised: Aug 07, 2023;Accepted: Sep 07, 2023

Manuscript information:

1. Total number of manuscript pages: 18
2. Supplementary pages: 3
3. Tables and figures: 2 figures.
4. Description of supplementary files: Supplementary material section contains 2 figures. Supplementary Figure 1 compares the effect of addition of RoseTTAFold-predicted structural templates as input for the task of single mutation effect prediction, arranged according to increasing MSA input depth. Supplementary Figure 2 shows the spearman rho correlations of all proteins evaluated, arranged according to increasing MSA input depth.

Abstract

Predicting the effects of mutations on protein function and stability is an outstanding challenge. Here, we assess the performance of a variant of RoseTTAFold jointly trained for sequence and structure recovery, RF_{joint}, for mutation effect prediction. Without any further training, we achieve comparable accuracy in predicting mutation effects for a diverse set of protein families using RFjoint to both another zero-shot model (MSA Transformer) and a model which requires specific training on a particular protein family for mutation effect prediction (DeepSequence). Thus, although the architecture of RF_{joint} was developed to address the protein design problem of scaffolding functional motifs, RF_{joint} acquired an understanding of the mutational landscapes of proteins during model training that is equivalent to that of recently developed large protein language models. The ability to simultaneously reason over protein structure and sequence could enable even more precise mutation effect predictions following supervised training on the task. These results suggest that RF_{joint} has a quite broad understanding of protein sequence-structure landscapes, and can be viewed as a joint model for protein sequence and structure which could be broadly useful for protein modeling.

Brief Statement: The RoseTTAFold deep neural network was trained to predict protein structures from amino acid sequences. RoseTTAFold was further modified to scaffold a given functional motif by training it for joint sequence and structure recovery of input proteins, resulting in RoseTTAFold Joint (RF_{joint}). Here we show that during the training, RF_{joint} acquired an understanding of protein sequence-structure relationships that enable zero-shot prediction of the effects of mutations on protein stability and function. Thus, RF_{joint} could be useful for distinguishing deleterious and neutral alleles in genome-wide association studies and designing proteins with higher stability and activity. Inference code for predicting the effect of single mutations on protein function or stability through this pipeline is available here: https://github.com/RosettaCommons/RFDesign/tree/main/inpainting. All input

data (target MSAs, structural templates), experimental and predicted values of all methods compared are available on Zenodo at link: https://doi.org/10.5281/zenodo.8106250

Keywords: deep learning, protein design, mutation effect prediction, zero-shot prediction, language models.

Introduction

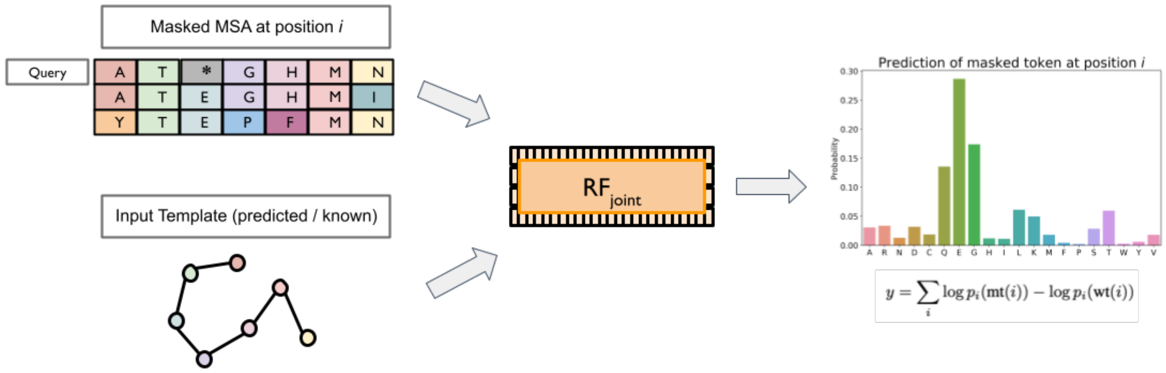
Accurate prediction of single point mutation effects using sequence information alone would help relate observed sequence polymorphisms to human disease [1, 2] and contribute to the design of proteins with higher functional activities. Deep learning methods have recently shown considerable promise for mutation effect prediction. DeepSequence [3], a probabilistic model for sequence families, obtained high accuracy in mutation effect prediction using latent variables for capturing higher-order interactions between residues in proteins through training on multiple sequence alignments (MSAs) for the target protein of interest. Large protein language models trained on MSAs (MSA Transformer) [4] or single sequences [5] also perform well at mutation effect prediction using an unsupervised or zero-shot approach. These language models have the advantage over DeepSequence of not requiring specific training on the protein family of interest.

RoseTTAFold was originally developed for protein structure prediction [6] and more recently RoseTTAFold Joint (RF_{joint}) was further trained to solve protein 'inpainting' problems [7]. During the inpainting process using the specifically trained RoseTTAFold network, a pass through the network starts from the functional site and fills in missing sequence and structure, resulting in the creation of a complete and viable protein scaffold. Included in RF_{joint} training was a masked MSA token recovery task for sequence prediction: predicting the correct amino acid sequence at specific masked positions within the alignment.

To assess RF_{joint}'s understanding of protein mutational landscapes, we set out to investigate whether it could predict experimental mutational data from published deep mutational scanning (DMS) sets [8] with no further training (i.e., using a "zero-shot" approach). We compared the performance of RoseTTAFold Joint on this task to that of MSA Transformer and DeepSequence. All three are MSA based methods, RF_{Joint} and MSA Transformer require no further training, while DeepSequence is trained on data from the family of interest. While not developed specifically for this task, we found that the performance in predicting the effects of single mutations on a set of diverse proteins was slightly better for $\mathrm{RF}_{\mathrm{joint}}$ than MSA Transformer and comparable to the specifically trained DeepSequence.

Results

$\mathrm{RF}_{\mathrm{joint}}$ was evaluated on a set of 38 deep mutational scans curated by Riesselman et al. [3] (The original dataset consisted of 42, we excluded the tRNA (TRNA_YEAST), the toxin-antitoxin complex (PARE_PARD), HIS7_YEAST_Kondrashov2017 and the PABP-doubles datasets to focus on single mutations made to monomeric proteins). Each of the mutational scans recorded a different protein function with varying measurements. Given that only 2 out of the 38 DMS datasets pertain specifically to stability, the evidence for the stability change prediction is weaker compared to that for the functional effect prediction. Each dataset was treated as a separate prediction task, and each variant was scored individually. For each target protein, we generated MSAs using iterative sequence search against the UniClust30 database as described in Baek *et al.* [6] and used it for both $\mathrm{RF}_{\mathrm{joint}}$ and MSA Transformer predictions. For $\mathrm{RF}_{\mathrm{joint}}$, the variants were scored by masking out the mutation site in the query sequence in the MSA, and the MSA token recovery head was used to predict the distribution over the masked position. The predicted effect of the mutation was calculated as the log odds ratio of the mutant amino acid and the wild-type amino acid (Figure 1). The performance on each dataset was assessed based on the spearman correlation of the predictions to the observed experimental values. For DeepSequence, we compared the results of MSA Transformer and $\mathrm{RF}_{\mathrm{joint}}$ to the published spearman rho values [3], which are from an ensemble of models trained on a different set of MSAs than those used for MSA Transformer or $\mathrm{RF}_{\mathrm{joint}}$ for each target protein.


Masked MSA at position *i*
Query
Input Template (predicted / known)
RF joint
Prediction of masked token at position *i*
Probability
0.30
0.25
0.20
0.15
0.10
0.05
0.00
A R N D C Q E G H I L K M F P S T W Y V

A	T	*	G	H	M	N
A	T	E	G	H	M	I
Y	T	E	P	F	M	N

$$y = \sum_i \log p_i(\mathrm{mt}(i)) - \log p_i(\mathrm{wt}(i))$$

Figure 1. Overall pipeline for zero-shot prediction of mutation effect using $\mathrm{RF_{joint}}$. A MSA is generated and masked at the mutation position in the query sequence, and structural templates are fed into pre-trained $\mathrm{RF_{joint}}$. Using the masked token prediction head, the emitted probability distribution of the 20 amino acids over the mutation site is used to calculate the effect of a mutation as the log odds ratio of the wild-type and mutation amino acid.

We found that $\mathrm{RF_{joint}}$ predicts mutational effects considerably better than a baseline calculated as the log odds ratio of the frequency of the mutant amino acid and of the wild-type amino acid in the MSA (Figure 2). $\mathrm{RF_{joint}}$ also slightly outperformed MSA Transformer and is comparable to the protein family-specific DeepSequence (Figure 2). $\mathrm{RF_{joint}}$ has the advantage in principle over the purely sequence based models of also being able to utilize structural template information, but we did not observe a significant improvement with incorporation of template structure information (Supplementary Figure 1; this may be in part because RoseTTAFold generates 3D models from MSA with reasonable accuracy). We also found little dependency of prediction accuracy on MSA depth (Supplementary Figure 2).

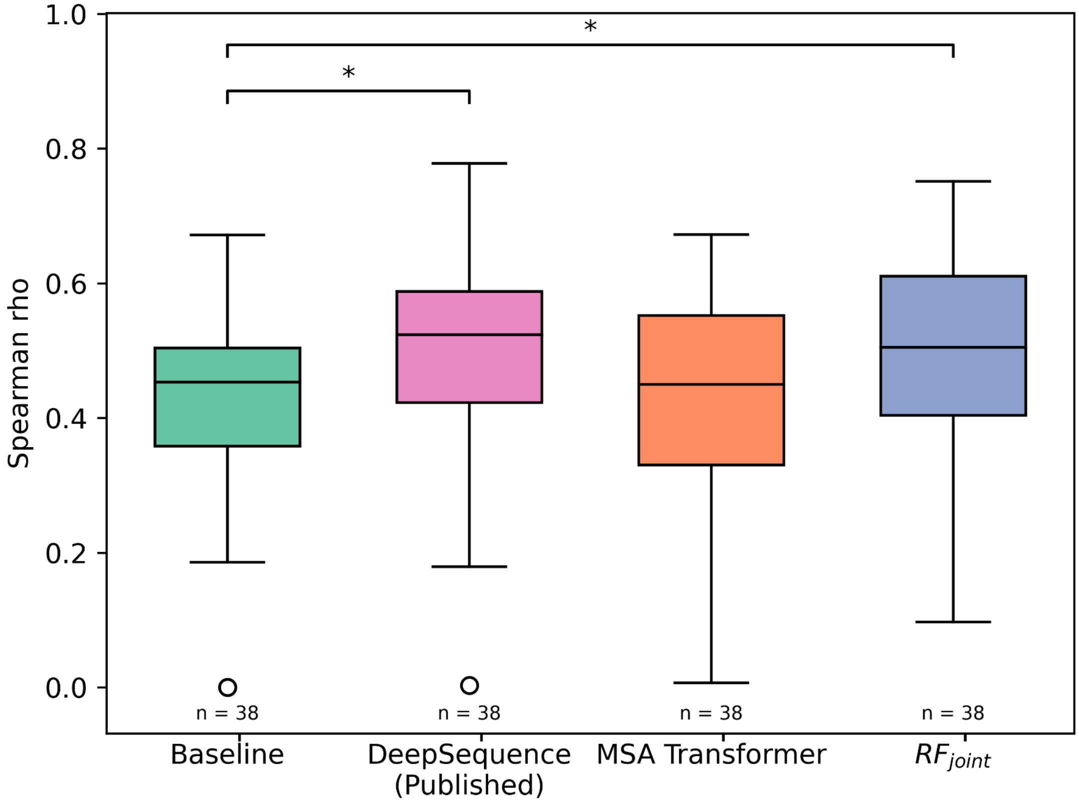


Figure 2. Boxplots of spearman rho correlations on deep mutation scanning datasets. Baseline refers to the non-ML MSA baseline. RF_{joint} refers to the model trained on a joint sequence and structure recovery task [7]. Box plots show the median (center line), interquartile range (hinges), and 1.5 times the interquartile range (whiskers); outliers are plotted as individual points. An asterisk above bars indicates significant differences: Baseline-DeepSequence ($p < 0.05$) and Baseline-RF_{joint} ($p < 0.05$), signifying p-values below the threshold. The average spearman rho correlation is 0.426 for the baseline, 0.502 for DeepSequence, 0.430 for MSA Transformer and 0.497 for RF_{joint}.

Discussion

We find that the RoseTTAFold network, developed originally for structure prediction and then extended to protein design, is also able to predict the effect of single mutations with quite high accuracy.

DeepSequence has a slightly higher average spearman rho correlation than RF_{joint}, but requires training for each protein family individually. Just as large protein language models, like MSA Transformer, provide general models of protein sequence, RoseTTAFold Joint may be viewed as a general joint model of protein sequence and structure. With further directed training, it should be possible to further improve mutation effect prediction performance by better utilizing protein structural information, which can be readily input into RoseTTAFold Joint but not into pure sequence based models, and by fine-tuning specifically on the mutant prediction task. As an additional future direction, exploring ensemble predictions using RF_{Joint} could further improve prediction accuracy and robustness. In conclusion, the predictive capabilities of RoseTTAFold Joint for protein structure and mutation effects, along with its potential for further enhancements through directed training and utilization of structural information, highlights its promising role as a general, joint model for protein sequence and structure.

Materials and Methods

1. **Deep Mutational Scanning Datasets:**

   RoseTTAFold was evaluated on a subset of 38 deep mutational scans collected by Riesselman et al. [3]. The proteins evaluated perform a wide range of functions and the experimental measures performed are different for each protein. We treat each deep mutational scanning dataset as a separate prediction task. Performance on each task is evaluated by spearman rho correlations of the calculated (baseline), published (DeepSequence) or predicted (RF_{joint} and MSA Transformer) scores to the experimental values.

2. **MSA Generation:**

   The same MSA inputs are used for both RoseTTAFold Joint and MSA Transformer at inference time. The protocol for generating MSAs is adopted from RoseTTAFold [6], where for each protein, sequences are found by iterative search against UniRef30 [9] and BFD [10] using HHblits [11]. Sequences are then filtered at 90% sequence identity cutoff. The E-value cutoff for sequence search is gradually relaxed (from 1e-10 to 1e-3) until the generated MSA has at least 2000 sequences with 75% coverage or 5000 sequences with 50% coverage. For the proteins that failed to get 5000 sequences (with E-value of 1e-3 and 50% sequence coverage cutoff), as many sequences as the protocol can find are used as an input MSA.

3. **Non-ML Baseline Setup:**

For establishing the non-ML baseline, we used the input MSA for each protein and calculated the log odds ratio of the frequency of the wild-type amino acid and mutant amino acid for each position (Equation 1). All sequences of the input MSA were used in this calculation.

$$\square_{\square\square\square\square\square\square\square\square\square,\square} = \square\square\square\,(\square\square\square\square_{\square\square,\square}) - \square\square\square(\square\square\square\square_{\square\square,\square}) \qquad \textit{Equation 1}$$

4. **RF$_{\mathrm{joint}}$ Inference Setup:**

We used the published RF$_{\mathrm{joint}}$ model [7] in inference mode for the task of single mutation effect prediction. All weights of the model were frozen and no further training was done. As described in the RF$_{\mathrm{joint}}$ paper [7], we split the input MSA into two groups, a small seed MSA and an extra MSA, to reduce the memory cost for all sequence-to-all sequence attention map calculation in the original RoseTTAFold. Up to 256 sequences were considered as a seed MSA (the input for RF$_{\mathrm{joint}}$'s main three-track blocks) from the input MSA of a target protein with an additional 1024 extra sequences (the input for RF$_{\mathrm{joint}}$'s ExtraMSAStack) passed into the model. All default parameters from RF$_{\mathrm{joint}}$ were used and the number of recycles was set to 1. RoseTTAFold [6] predicted structures for a target protein were used as structural templates for mutation effect prediction. The mutation site of interest was masked in the query sequence of the input MSA and the masked MSA token recovery head was used to predict the probability of all 20 amino acids over that masked position. The predicted effect of a mutation at position *i* was calculated as the log odds ratio of the probability of the wild-type amino acid to the mutant amino acid (Equation 2). This scoring is zero-shot i.e. the model requires no further training.

$$\square_{\square\square\square\square\square\square\square\square,\square} = \square\square\square\,(\square_{\square\square,\square}) - \square\square\square(\square_{\square\square,\square}) \qquad \textit{Equation 2}$$

5. MSA Transformer Inference Setup:

We used the published MSA Transformer [4,5] loaded with pre-trained weights (annotated as `esm_msa1b_t12_100M_UR50S` on the public ESM github). The default arguments were used, where 400 sequences were randomly sampled from the MSA for inference. We used the masked marginals scoring strategy for scoring mutants from MSA Transformer, which is done by introducing masks at the mutated positions and computing the score for a mutation by considering its probability relative to the wildtype amino acid [5]. This is similar to the setup that we used for predicting the effect of a mutation through RF_{joint} (Equation 2).

Acknowledgements

We would like to thank Justas Dauparas, Ivan Anishchanka, Doug Tischer, Hahnbeom Park, Sergey Ovchinnikov and Eric Horvitz for helpful comments and suggestions.

References

1. Shin JE, Riesselman AJ, Kollasch AW, McMahon C, Simon E, Sander C, et al. Protein design and variant prediction using autoregressive generative models. Nat Commun. 2021;12(1).

2. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35(2).

3. Riesselman, A. J., Ingraham, J. B., & Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, *15*(10), 816–822. https://doi.org/10.1038/s41592-018-0138-4.

4. Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T., & Rives, A. (2021). MSA Transformer. *BioRxiv*, 2021.02.12.430858. https://doi.org/10.1101/2021.02.12.430858.

5. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., & Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *BioRxiv*, 2021.07.09.450648. https://doi.org/10.1101/2021.07.09.450648.

6. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. v, van Dijk, A. A., Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. https://doi.org/10.1126/science.abj8754.

7. Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J. L., Castro, K. M., Ragotte, R., Saragovi, A., Milles, L. F., Baek, M., Anishchenko, I., Yang, W., Hicks, D. R., Expòsit, M., Schlichthaerle, T., Chun, J.-H., Dauparas, J., Bennett, N., Wicky, B. I. M., … Baker, D. (2022). Scaffolding protein functional sites using deep learning. *Science*, *377*(6604), 387–394. https://doi.org/10.1126/science.abn2100

8. Starita LM, Fields S. Deep mutational scanning: A highly parallel method to measure the effects of mutation on protein function. Cold Spring Harb Protoc. 2015 Aug 1;2015(8):711–4.

9. Mirdita M, von Den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017 Jan 1;45(D1):D170–6.

10. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods. 2019 Jul 1;16(7):603–6.

11. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019 Sep 14;20(1).

Figure Legends:

1. **Figure 1.** Overall pipeline for zero-shot prediction of mutation effect using RF_{joint}. A MSA is generated and masked at the mutation position in the query sequence, and structural templates are fed into pre-trained RF_{joint}. Using the masked token prediction head, the emitted probability distribution of the 20 amino acids over the mutation site is used to calculate the effect of a mutation as the log odds ratio of the wild-type and mutation amino acid.
2. **Figure 2.** Boxplots of spearman rho correlations on deep mutation scanning datasets. Baseline refers to the non-ML MSA baseline. RF_{joint} refers to the model trained on a joint sequence and structure recovery task [7]. Box plots show the median (center line), interquartile range (hinges), and 1.5 times the interquartile range (whiskers); outliers are plotted as individual points. The average spearman rho correlation is 0.426 for the baseline, 0.502 for DeepSequence, 0.430 for MSA Transformer and 0.497 for RF_{joint}.

Conflict of Interest Statement:

The authors declare no conflict of interest.