Accurate Mutation Effect Prediction 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, WA 98195, USA.
[2.] Institute for Protein Design, University of Washington, Seattle, WA 98195, USA.
[3.] Molecular Engineering Graduate Program, University of Washington, WA 98195, USA.
[4.] School of Biological Sciences, Seoul National University, Seoul, 08826, Republic of Korea.
[5.] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.

Abstract

Predicting the effects of mutations on protein function is an outstanding challenge. Here we assess the performance of the deep learning based RoseTTAFold structure prediction and design method for unsupervised mutation effect prediction. Using RoseTTAFold in inference mode, without any additional training, we obtain state of the art accuracy on predicting mutation effects for a set of diverse protein families. Thus, although the architecture of RoseTTAFold was developed to address the protein structure prediction problem, during model training RoseTTAFold acquired an understanding of the mutational landscapes of proteins comparable to that of large recently developed language models. The ability to reason over structure as well as sequence could enable even more precise mutation effect predictions following supervised training.

Main Text

Accurate and unsupervised prediction of single 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 excellent performance 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 multiple sequence alignments (MSA Transformer) [4] or single sequences [5] also performed very well at mutation effect prediction, and have the advantage over DeepSequence of not requiring specific training on the protein of interest. RoseTTAFold was originally developed for protein structure prediction [6], but during training we included a masked token recovery task, and a recently developed version, RoseTTAFold Joint (RF_{joint}) was further trained to solve ‘inpainting’ problems in which substantial portions of both sequence and structure are rebuilt [7]. To assess RF_{joint}'s understanding of protein sequence-structure relationships, we set out to investigate whether it could predict experimental mutational data from published deep mutational scanning (DMS) sets [8] with no further training (using a zero-shot approach). We compared RoseTTAFold performance on this task to that of the state of the art MSA Transformer; both are MSA based methods requiring no further training.

RF_{joint} was evaluated on a set of 38 deep mutational scans curated by Riesselman et al. [3]. Each of the mutational scans recorded a different protein function with varying measurements. 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 RF_{joint} and MSA Transformer predictions. For RF_{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.

We found that 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). RF_{joint} also slightly outperformed MSA Transformer (Figure 2). RoseTTAFold 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 (data not shown; this may be in part because RoseTTAFold generates 3D models from sequence with reasonable accuracy). We also found little dependency of prediction accuracy on MSA depth (Supplementary Figure 1).

Conclusion

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. Just as large language models like the 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 more directed training, it should be possible to further improve performance by better utilizing protein structural information, which can be readily input into RoseTTAFold but not into pure sequence based models, and by fine-tuning specifically for the mutational effect prediction task. More generally, our results demonstrate that RoseTTAFold has quite a broad understanding of protein mutational landscapes, which should be very useful for protein design and other challenges involving inference over both sequence and structure.

Materials and Methods

We used the published RF_{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. Up to 256 sequences were considered from the input MSA of a target protein with an additional 1024 extra sequences passed into the model. All default parameters from RF_{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. Inference code for predicting the effect of single mutations through this pipeline is available here: https://github.com/RosettaCommons/RFDesign/tree/main/inpainting

Acknowledgements and Disclosure of Funding

We would like to thank Justas Dauparas, Ivan Anishchanka, Doug Tischer, Hahnbeom Park, Sergey Ovchinnikov and Eric Horvitz for helpful comments and suggestions. This work was supported by Microsoft (S.M. M.B., D.B., and generous gifts of Azure computing time), Eric and Wendy Schmidt by recommendation of the Schmidt Futures (D.J.), the EMBO Postdoctoral Fellowship (ALTF 292-2022) (J.L.W.), the Audacious Project at the Institute for Protein Design (D.B.), a gift from Amgen (J.LW. and D.B.) and the Howard Hughes Medical Institute (D.B.).

References

1. Shin, Jung-Eun et al. “Protein design and variant prediction using autoregressive generative models.” Nature communications vol. 12,1 2403. 23 Apr. 2021, doi:10.1038/s41467-021-22732-w
2. Hopf, Thomas A et al. “Mutation effects predicted from sequence co-variation.” Nature biotechnology vol. 35,2 (2017): 128-135. doi:10.1038/nbt.3769
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. Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, Alexander Rives Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8844-8856, 2021.
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. 1–28.
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*, 871–876. https://predictioncenter.org/casp14/
7. Wang, Jue et al. "Scaffolding Protein Functional Sites Using Deep Learning". Science, vol 377, no. 6604, 2022, pp. 387-394. American Association For The Advancement Of Science (AAAS), https://doi.org/10.1126/science.abn2100.
8. Starita, L. M., & Fields, S. (2015). Deep mutational scanning: A highly parallel method to measure the effects of mutation on protein function. *Cold Spring Harbor Protocols*, *2015*(8), 711–714. https://doi.org/10.1101/pdb.top077503

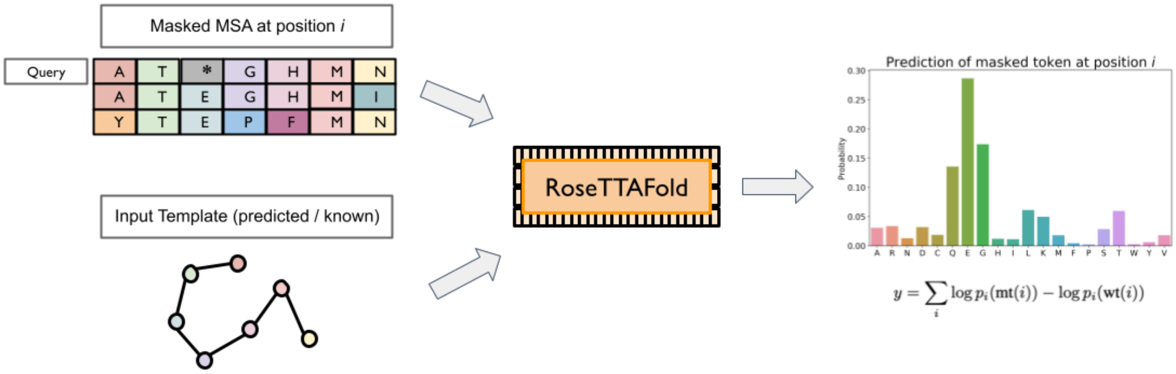

Masked MSA at position *i*
Query
A T * G H M N
A T E G H M I
Y T E P F M N
Input Template (predicted / known)
RoseTTAFold
Prediction of masked token at position *i*
Probability
0.30
0.25
0.20
0.15
0.10
0.05
0.00
A R N D C Q E G H I L K M F P S T W Y V
$y = \sum_i \log p_i(\mathrm{mt}(i)) - \log p_i(\mathrm{wt}(i))$

Figure 1. Overall pipeline for zero-shot prediction of mutation effect using RoseTTAFold. A MSA is generated and masked at the mutation position in the query sequence, and structural templates are fed into pre-trained RoseTTAFold. 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.

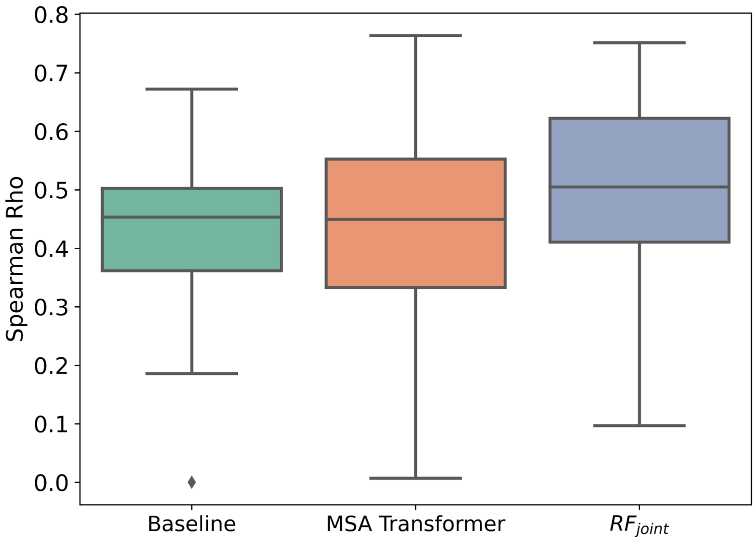
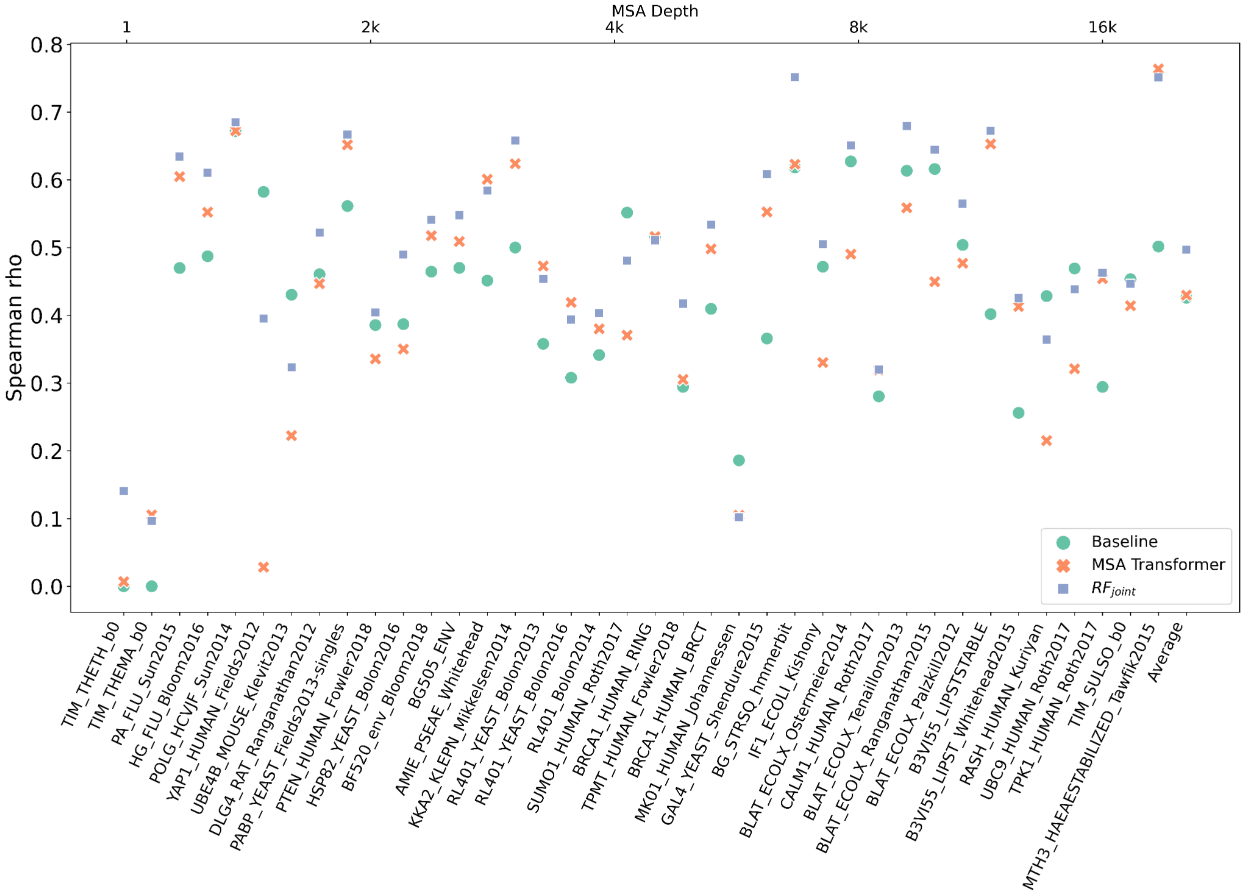


Figure 2. Boxplots of spearman rho correlations on deep mutation scanning datasets. Baseline refers to the non-ML MSA baseline. $\mathrm{RF}_{\mathrm{joint}}$ refers to the model trained on a joint

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.



Supplementary Figure 1. Spearman rho correlations for all deep mutational scanning datasets evaluated. Each point corresponds to a different protein. The points are arranged according to increasing MSA depth for RF_{joint} and MSA Transformer.