Predicting single-point mutational effect on protein stability

Simon K.S. Chu
Biophysics Graduate Program
University of California, Davis
Davis, CA 95616
kschu@ucdavis.edu

Justin Siegel
Department of Chemistry
University of California, Davis
Davis, CA 95616
jbsiegel@ucdavis.edu

Abstract

Engineering a protein's stability improves its shelf life and expands its application environment. Current studies of protein stability often involve predicting stability change from single-point mutations. However, the prediction model must be able to resolve single-character difference in a protein sequence hundreds of amino acid long. In this study, we predicted single-point mutational effect on protein stability and compared sequence-only and geometric learning approaches using sequence embedding. We showed that sequence-only models suffice in predicting single-point mutational change. A simple MLP incorporating only the embedding at the mutation site achieves similar performance with geometric model. The observation is consistent across 28 single-point mutational datasets with a wide range of functional properties.

Protein stability enhancement has been a focus in protein engineering due to its role in extending a protein's shelf life and expanding its application environment. Typical protein stability studies in silico predict stability change of the mutated sequences and validate the predicted outcome experimentally. However, capturing a protein's change of stability upon single-point mutation has been challenging: first, it is unknown how far a mutation's effect will propagate to its neighbors in physical and sequence space. Second, the prediction model must be able to resolve the single-point mutation amid the length of protein sequence up to hundreds or even thousands of amino acids.

Traditional molecular modeling predicts protein stability by describing the physical basis behind the models. For instance, Frenz et al. performed mutation and structural relaxation in silico, and showed moderate performance on a multi-protein dataset [1, 2]. However, systematic evaluation on single-protein point mutations discovered only weak correlation between experimental data and predictions generated from molecular modeling or machine learning based on molecular modeling features [3]. Alongside structural modeling, evolutionary analysis incorporates only protein sequence information. Recently, protein language models were developed and have shown their ability to perform supervised and unsupervised functional predictions. [4–11].

In this study, we compare geometric and sequence-only models in protein stability prediction incorporating embeddings from protein language model. Following Gligorijević et al.'s work in annotating protein function from structures [12], we applied graph convolution to prediction of single-point mutational change and compared the performance to non-geometric baselines.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

1 Method

1.1 Datasets

Romero et al. measured the susceptibility of beta-glucosidase (bgl3) to heat challenge in the dataset used for this study [13]. We used the 33th layer of Evolutionary Scale Modeling (ESM-1b) as the sequence embedding [4], whereas a mutation is represented as the difference between mutant sequence embedding and that of wildtype, defined as a simple subtraction in this study. We denote the difference as *mutational embedding* which represents the mutational effect on the whole sequence.

Missing residues in the crystal structure (PDB code: 1gnx) [14] poses a challenge in geometric learning. As such, we generated protein graph from AlphaFold2 [15] on wildtype sequence, which has a root-mean-square deviation (RMSD) smaller than 1Å from the experimental structure (figure A12), and approximated the single-point mutant structure with that of wildtype. For geometric models, we defined C_α atom of each residue as a node and edges were drawn between nodes if they are within 6Å from each other. Mutational embedding corresponding to each residue was then assigned to the respective node on the protein graph. The resulted 3500 samples were then randomly split in 8-1-1 ratio as train-validation-test sets.

To broaden the scope of comparison between geometric and sequence-only models, we included additional datasets from Riesselman et al., from which datasets were selected to avoid redundancy, multi-protein, tRNA, multiple mutations and sequence length longer than 1024 for ESM-1b inference. The resultant 28 datasets span a wide range of functional properties from enzyme function to cell growth [16–35].

1.2 Model architecture

We tested one geometric and three sequence-only architectures, named as GCNModel, NoEdgeGCN, SeqPoolingMLP and SingleSiteMLP.

GCNModel. Graph-convolution-based model. Graph input is passed through three layers of GCN, global softmax aggregation [36], and finally a 3-layer MLP.

NoEdgeGCN. Sequence-only model. It has the same architecture as GCNModel but has no message passing. Alternatively, it can be viewed as a GCNModel ignoring all edges in the protein graph.

SeqPoolingMLP. Sequence-only model. A 3-layer MLP after pooling mutation embedding on all residues. This model is a special case of a single-layer GCNModel without message passing.

SingleSiteMLP. Sequence-only model specifically for single-site mutation prediction. A simple 3-layer MLP model which uses only the mutational embedding at the mutation site as the input.

Each convolutional layer is a submodule of GCN [37] followed by LeakyReLU. Each MLP layer is a submodule of Linear followed by LeakyReLU. All hidden layers have the same channel size in both GCN and MLP layers. All global softmax aggregation has an inverse temperature β of 1. LeakyReLU has a default slope of 0.2 for $x < 0$.

The experiment was implemented in PyTorch, PyTorch Geometric and PyTorch Lightning [38–40]. The code can be accessed on github.

1.3 Untrained protocols and unsupervised predictions

Five untrained protocols and unsupervised predictions were evaluated. First, we benchmarked traditional molecular modeling and sequence conservation methods, i.e. Rosetta Cartesian DDG and Position Specific Scoring Matrix. Rosetta Cartesian DDG (cart_ddg) is a molecular modeling protocol that mutates, relaxes, and scores the energetic difference between mutant and wildtype structures [1], whereas Position Specific Scoring Matrix (PSSM), aligns sequences and scores the likelihood of every amino acid type at each position on the query sequence, and was obtained through psiblast on Uniref90 [41, 42].

Three unsupervised protein sequence models were evaluated. We included EVmutation and DeepSequence which are protein sequence models accounting for pairwise interactions and higher order interactions through variational autoencoder (VAE) respectively [10, 11]. Evolutionary Scale Mod-

Table 1: Performance comparison on unsupervised and supervised predictions.

Model	PCC	SRC	R-square
Rosetta cart_ddg	NA	0.22	NA
PSSM	NA	0.69	NA
EVmutation	NA	0.74	NA
DeepSequence	NA	0.76	NA
ESM-1v	NA	0.61	NA
GCNModel	0.65±0.01	0.70±0.01	0.40±0.01
NoEdgeGCN	0.69±0.02	0.72±0.01	0.46±0.02
SeqPoolingMLP	0.71±0.01	0.73±0.02	0.49±0.01
SingleSiteMLP	**0.77**±0.00	**0.78**±0.00	**0.57**±0.01

eling (ESM-1v) is a BERT model trained on Uniref90 database and has been shown to perform zero-shot prediction on a variety of single-point mutational datasets [4].

2 Results

Among all the unsupervised predictions, DeepSequence ranks the first with a Spearman Rank Correlation (SRC) of 0.76 with mutational change of protein stability, as tabulated in table 1. Pairwise model EVmutation performs similarly (SRC 0.74) while attention-based ESM-1v scores 0.61 on this dataset without pretraining on homologous sequences. Interestingly, traditional sequence alignment method (PSSM) has a strong correlation of 0.69 despite its simplicity. Rosetta cart_ddg ranks the last. We omitted Pearson Correlation Coefficient (PCC) and R-square since the protocols are not designed nor trained for the experiment readings.

Interestingly, sequence-only models suffice in predicting protein stability change on this dataset. GCNModel has a slightly lower performance to that of NoEdgeGCN where no edge is drawn between nodes. Despite having fewer layers, SiteSiteMLP and SeqPoolingMLP outperform GCNModel. The result is insensitive to the choice of graph convolution, ESM weights and radius for edge definition (figure A1, A2, A3).

Surprisingly, SingleSiteMLP which takes the embedding only at the mutation site outperforms all other models on every metric. SingleSiteMLP surpasses SeqPoolingMLP, second in rank, by 0.06 in PCC, 0.05 in SRC and 0.08 in R-square. However, SingleSiteMLP is restricted to only single-point mutation whereas GCNModel, NoEdgeGCN and SeqPoolingMLP have an architecture for any number of mutations.

To test the generalizability of the observation, we expanded geometric and sequence-only model comparison to another 27 single-point mutational datasets. Illustrated in Figure 1, sequence-only model(s) perform equally or better than geometric model in all datasets except PAPB_singles in PCC assessed by Wilcoxon signed-rank test ($p < 0.05$) and similarly on SRC and R-square (figure A4, A5). In particular, SingleSiteMLP performs similarly or better than all other models in 25 out of 28 datasets in PCC. Switching from ESM-1b to onehot embedding results in significant performance drop, which highlights the importance of pretrained sequence embedding (figure A7, A8, A9).

3 Discussion

We discovered that ESM-1b can resolve single-point mutations and capture neighboring interactions. For example, the mutational embedding of R368G on beta-glucosidase reflects the removal of hydrogen bond to residue D420 by highlighting its hydrogen bonding partner on the adjacent alpha helix (figure 2). However, non-local interactions captured by the language model are relatively mild on average.

To conceptualize non-local interactions, we use norm (magnitude) of the mutational embedding to represent information contained in each residue. Residues with zero norm implies the language model predicts no mutational effect on them, i.e. no difference between wildtype and mutant sequence embedding. As such, prediction of mutational change shall not depend on these residues. On

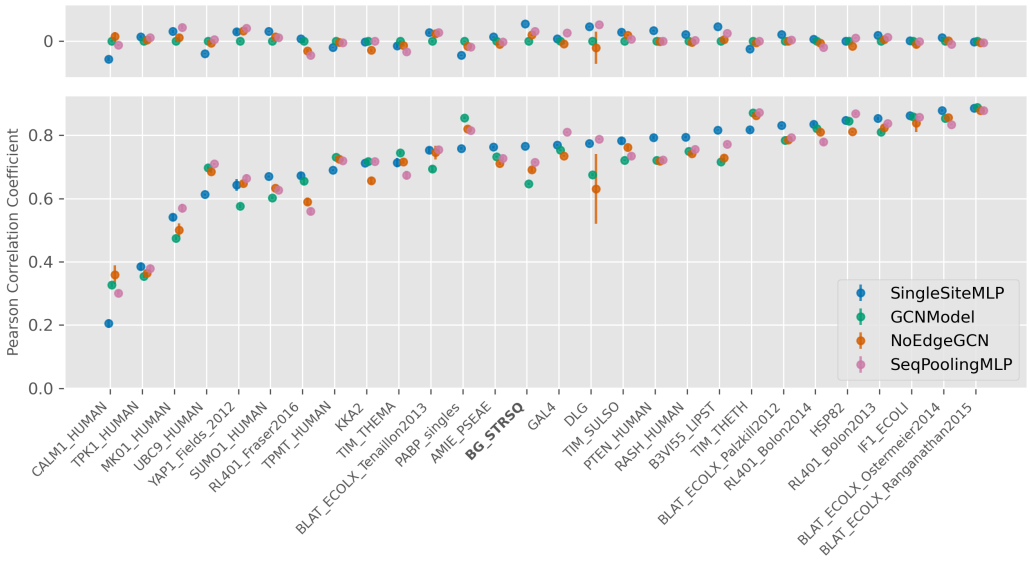

Figure 1: Performance comparison between geometric and sequence-only models on functional predictions of single-point mutation. Beta-glucosidase stability dataset (BG_STRSQ) is bolded. (Lower) PCCs evaluated on 28 datasets, ranked by SingleSiteMLP performance. (Upper) Differences in PCC relative to geometric model.

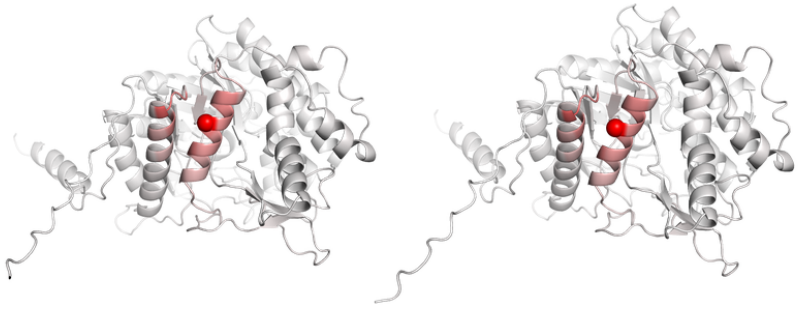

Figure 2: R368G mutational embedding (left) ESM-1b and (right) ESM-1v of beta-glucosidase (bgl3). The mutation site (residue 368) is represented in sphere. The embedding is colorized by the norm of embedding on each node, i.e. white represents smaller norm and red represents larger value.

average, mutation site has a norm twice as large as compared to those 3 Å away. The contrast is even stronger when focusing on residues geometrically close but non-adjacent in sequence (figure A11). This potentially explains the locality of single-point mutational embedding and why sequence-only model(s) performs similarly or better than geometric model.

4 Conclusion

We predicted single-point mutational effect on a protein's stability and compared sequence-only and geometric learning approaches. We showed that sequence-only models are sufficient for single-point mutational prediction, and the observation is consistent for a variety of functional datasets. With embedding only at the mutation site, a simple MLP model can predict single-point mutational change with similar performance. This finding could be attributed to the observation that mutational effect on average propagates weakly to residues geometrically close to mutation site but non-adjacent in the sequence.

5 Broader Impact

The authors have no ethical aspects nor future societal consequences to add.

Acknowledgments and Disclosure of Funding

We thank Fangzhou Li from Tagkopoulos lab, UC Davis for useful discussions, and Youtian Cui, Tim Coulther and Peishan Huang from Siegel lab, UC Davis for comments on manuscript.

References

[1] Brandon Frenz, Steven M Lewis, Indigo King, Frank DiMaio, Hahnbeom Park, and Yifan Song. Prediction of protein mutational free energy: benchmark and sampling improvements increase classification accuracy. *Frontiers in bioengineering and biotechnology*, 8:1175, 2020.

[2] Hahnbeom Park, Philip Bradley, Per Greisen Jr, Yuan Liu, Vikram Khipple Mulligan, David E Kim, David Baker, and Frank DiMaio. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of chemical theory and computation*, 12(12):6201–6212, 2016.

[3] Peishan Huang, Simon KS Chu, Henrique N Frizzo, Morgan P Connolly, Ryan W Caster, and Justin B Siegel. Evaluating protein engineering thermostability prediction tools using an independently generated dataset. *ACS omega*, 5(12):6487–6493, 2020.

[4] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021.

[5] Maxat Kulmanov, Mohammed Asif Khan, and Robert Hoehndorf. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, 2018.

[6] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-only deep representation learning. *bioRxiv*, page 589333, 2019.

[7] Maxwell L Bileschi, David Belanger, Drew Bryant, Theo Sanderson, Brandon Carter, D Sculley, Mark A DePristo, and Lucy J Colwell. Using deep learning to annotate the protein universe. *BioRxiv*, page 626507, 2019.

[8] Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

[9] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32:9689, 2019.

[10] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

[11] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

[12] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):1–14, 2021.

[13] Philip A Romero, Tuan M Tran, and Adam R Abate. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, 112(23):7159–7164, 2015.

[14] A. Guasch, M. Vallmitjana, R. Perez, J.A. Querol, E.and Perez-Pons, and M. Coll. Beta-glucosidase from streptomyces. To be published.

[15] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[16] Emily E Wrenbeck, Laura R Azouz, and Timothy A Whitehead. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nature communications*, 8(1):1–10, 2017.

[17] Justin R Klesmith, John-Paul Bacik, Ryszard Michalczyk, and Timothy A Whitehead. Comprehensive sequence-flux mapping of a levoglucosan utilization pathway in e. coli. *ACS synthetic biology*, 4(11):1235–1243, 2015.

[18] Elad Firnberg, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular biology and evolution*, 31(6): 1581–1592, 2014.

[19] Michael A Stiffler, Doeke R Hekstra, and Rama Ranganathan. Evolvability as a function of purifying selection in tem-1 β-lactamase. *Cell*, 160(5):882–892, 2015.

[20] Hervé Jacquier, André Birgy, Hervé Le Nagard, Yves Mechulam, Emmanuelle Schmitt, Jérémy Glodt, Beatrice Bercot, Emmanuelle Petit, Julie Poulain, Guilène Barnaud, et al. Capturing the mutational landscape of the beta-lactamase tem-1. *Proceedings of the National Academy of Sciences*, 110(32):13067–13072, 2013.

[21] Zhifeng Deng, Wanzhi Huang, Erol Bakkalbasi, Nicholas G Brown, Carolyn J Adamski, Kacie Rice, Donna Muzny, Richard A Gibbs, and Timothy Palzkill. Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *Journal of molecular biology*, 424(3-4):150–167, 2012.

[22] Jochen Weile, Song Sun, Atina G Cote, Jennifer Knapp, Marta Verby, Joseph C Mellor, Yingzhou Wu, Carles Pons, Cassandra Wong, Natascha van Lieshout, et al. A framework for exhaustively mapping functional missense variants. *Molecular systems biology*, 13(12):957, 2017.

[23] Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422): 138–142, 2012.

[24] Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields, and Jay Shendure. Massively parallel single-amino-acid mutagenesis. *Nature methods*, 12(3):203–206, 2015.

[25] Parul Mishra, Julia M Flynn, Tyler N Starr, and Daniel NA Bolon. Systematic mutant analyses elucidate general and client-specific aspects of hsp90 function. *Cell reports*, 15(3):588–598, 2016.

[26] Eric D Kelsic, Hattie Chung, Niv Cohen, Jimin Park, Harris H Wang, and Roy Kishony. Rna structural determinants of optimal codons revealed by mage-seq. *Cell systems*, 3(6):563–571, 2016.

[27] Alexandre Melnikov, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S Mikkelsen. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic acids research*, 42(14):e112–e112, 2014.

[28] Lisa Brenan, Aleksandr Andreev, Ofir Cohen, Sasha Pantel, Atanas Kamburov, Davide Cacchiarelli, Nicole S Persky, Cong Zhu, Mukta Bagul, Eva M Goetz, et al. Phenotypic characterization of a comprehensive set of mapk1/erk2 missense mutants. *Cell reports*, 17(4):1171–1183, 2016.

[29] Yvonne H Chan, Sergey V Venev, Konstantin B Zeldovich, and C Robert Matthews. Correlation of fitness landscapes from three orthologous tim barrels originates from sequence and structure constraints. *Nature communications*, 8(1):1–12, 2017.

[30] Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, and Stanley Fields. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein. *Rna*, 19(11):1537–1551, 2013.

[31] Kenneth A Matreyek, Lea M Starita, Jason J Stephany, Beth Martin, Melissa A Chiasson, Vanessa E Gray, Martin Kircher, Arineh Khechaduri, Jennifer N Dines, Ronald J Hause, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature genetics*, 50(6):874–882, 2018.

[32] Pradeep Bandaru, Neel H Shah, Moitrayee Bhattacharyya, John P Barton, Yasushi Kondo, Joshua C Cofsky, Christine L Gee, Arup K Chakraborty, Tanja Kortemme, Rama Ranganathan, et al. Deconstruction of the ras switching cycle through saturation mutagenesis. *Elife*, 6:e27810, 2017.

[33] Benjamin P Roscoe and Daniel NA Bolon. Systematic exploration of ubiquitin sequence, e1 activation efficiency, and experimental fitness in yeast. *Journal of molecular biology*, 426(15): 2854–2870, 2014.

[34] David Mavor, Kyle Barlow, Samuel Thompson, Benjamin A Barad, Alain R Bonny, Clinton L Cario, Garrett Gaskins, Zairan Liu, Laura Deming, Seth D Axen, et al. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *Elife*, 5: e15802, 2016.

[35] Carlos L Araya, Douglas M Fowler, Wentao Chen, Ike Muniez, Jeffery W Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences*, 109 (42):16858–16863, 2012.

[36] Guohao Li, Chenxin Xiong, Ali Thabet, and Bernard Ghanem. Deepergcn: All you need to train deeper gcns. *arXiv preprint arXiv:2006.07739*, 2020.

[37] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[39] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[40] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 3 2019. URL `https://github.com/PyTorchLightning/pytorch-lightning`.

[41] National Center for Biotechnology Information (US) and Christiam Camacho. *BLAST (r) Command Line Applications User Manual*. National Center for Biotechnology Information (US), 2008.

[42] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1): D204–D212, 2015.

A Appendix

A.1 Hyperparameter search and training

A hyperparameter search on {8, 16, 32} hidden channels and {0, 1e-6, 1e-4} weight decay was performed on beta-glucosidase dataset. All models were trained on Adam optimizer with learning rate of 5e-3, $\beta 1 = 0.9$, $\beta 2 = 0.999$ on a patience of 25 epochs and 256 batch size on RTX 2070. For each model architecture, the best hyperparamters were picked based on PCC on validation set. SingleSiteMLP has a weight decay of 1e-4 while the rest has no weight decay. All architectures have 32 hidden channels. We fixed the hyperparameters on the rest of 27 datasets. For each model type, the reported performance is on test set by the checkpoint of the lowest mean-square-error (MSE) on the validation set. All dataset-architecture pairs were repeated 5 times with random initialization except the ablation study of onehot-encoding embedding.

A.2 Performance evaluation

Untrained protocols and unsupervised models were evaluated on the whole dataset whereas supervised models were assessed on the test set only. For supervised models, all errors were estimated from 5 randomly initialized models by standard error of mean. Performance of ESM-1v were evaluated on the average performance of masked-marginals over esm1v_t33_650M_UR90S_[1-5].

A.3 License

Evolutionary Scale Modeling (ESM) is available on `https://github.com/facebookresearch/esm`. Both ESM and the code for this project are licensed under MIT license. The github to the project is `https://github.com/SimonKitSangChu/MLSB2021`.

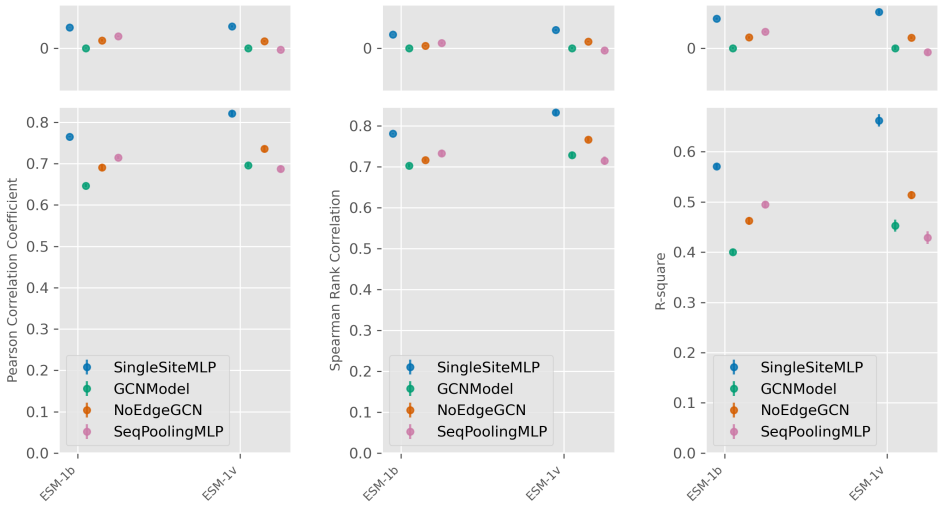

Figure A1: Performance comparison between using ESM-1b and ESM-1v embeddings on beta-glucosidase stability dataset. Model performance was evaluated on (left) PCC, (middle) SRC and (right) R-square.(Upper) Differences in performance metric compared to geometric model.

A.4 Embedding localization analysis

For each mutation in the dataset, we calculated the norm of mutational embedding on each residue, and normalized it across all residues. The statistics was then binned by geometric distance from mutation site, and averaged across all residues within that bin. We highlighted non-local mutational effect on residues geometrically close but non-adjacent on sequence, by filtering out nodes 1) closer than 16 residues from mutation site on sequence and 2) further from 25 Å geometric distance. The observation on beta-glucosidase dataset (figure A11) is consistent with those on all 28 datasets.

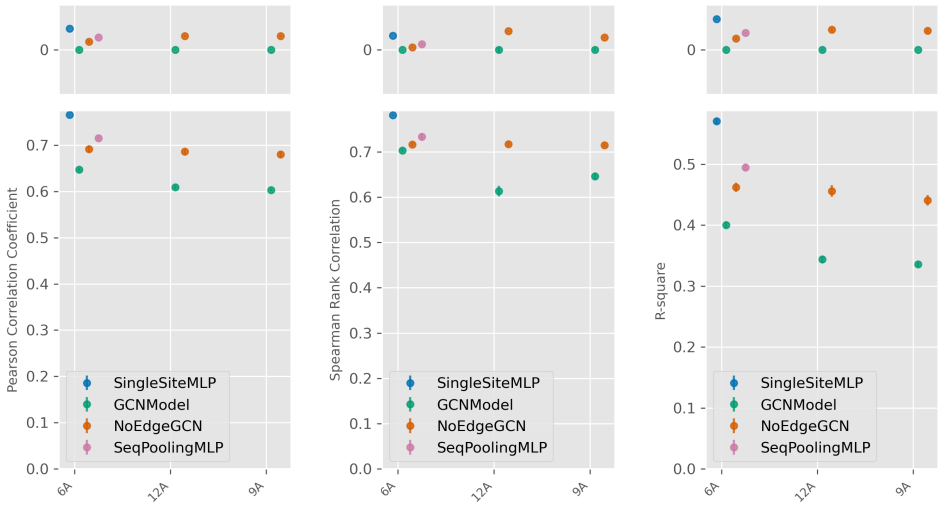

Figure A2: Performance comparison of radius choice in edge definition. All units are defined in angstrom. Model performance evaluated on (left) PCC, (middle) SRC and (right) R-square. (Upper) Differences in performance metric compared to geometric model.

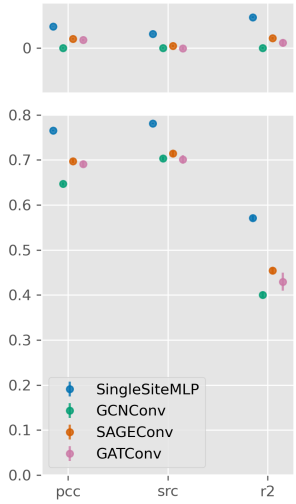

Figure A3: Performance comparison between SingleSiteMLP, GCNConv, SAGEConv and GATConv models on beta-glucosidase stability dataset.

A.5 Position Specific Scoring Matrix

The Position Specific Scoring Matrix (PSSM) was obtained through psiblast with the following command.

psiblast -query dataset.fasta -db uniref90 -num_iterations 3 -out_ascii_pssm dataset.pssm

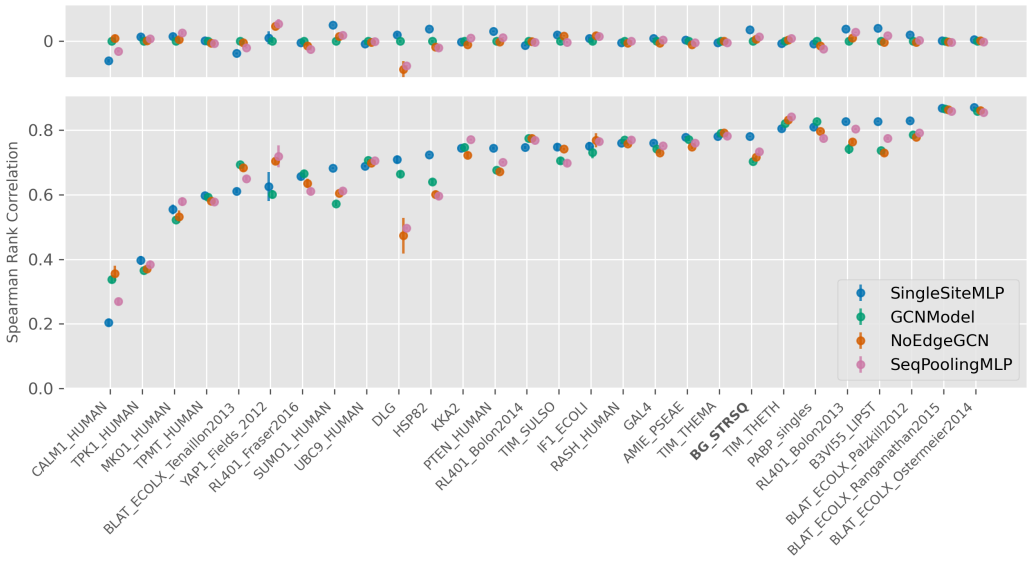

Figure A4: Performance comparison between geometric and sequence-only models on functional predictions of single-point mutations. (Lower) SRCs evaluated on 28 datasets ranked by SingleSiteMLP performance. (Upper) Differences in SRC relative to geometric model.

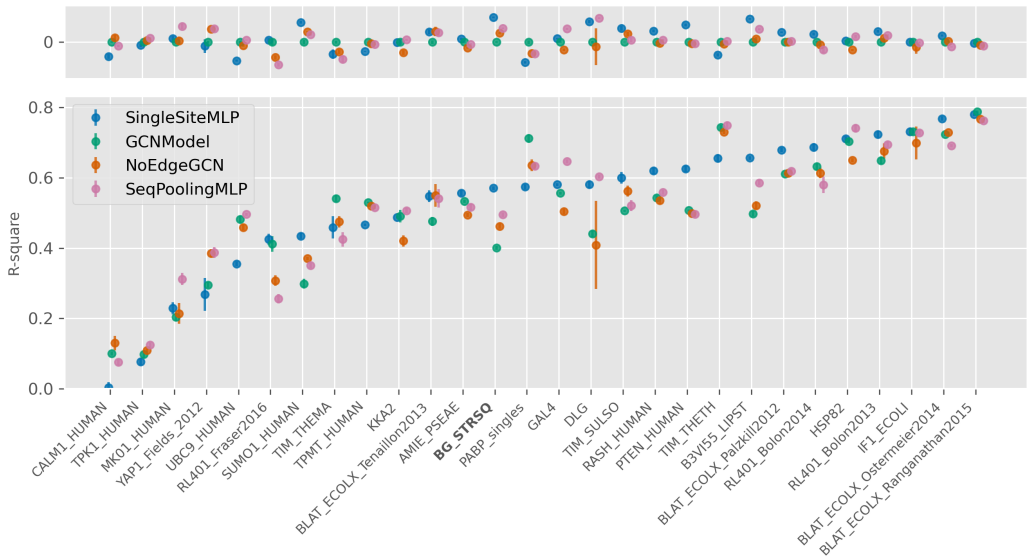

Figure A5: Performance comparison between geometric and sequence-only models on functional predictions of single-point mutations. (Lower) R-square evaluated on 28 datasets ranked by SingleSiteMLP performance. (Upper) Differences in R-square relative to geometric model.

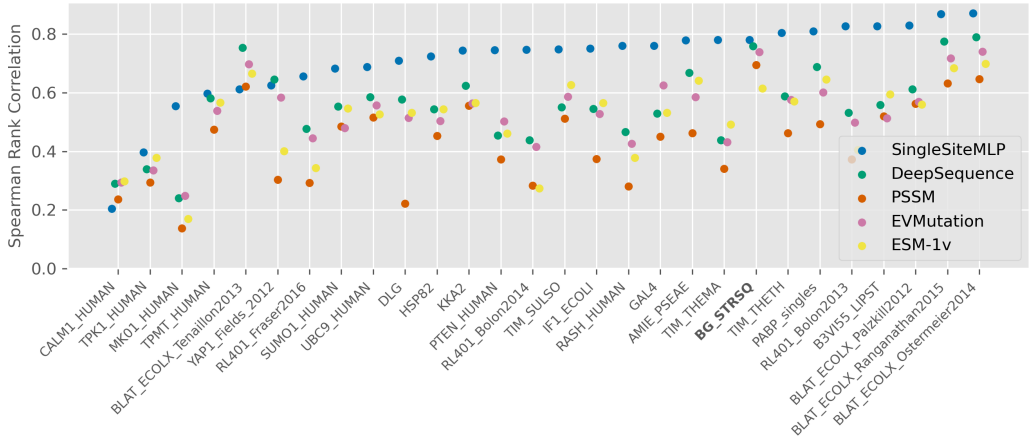

Figure A6: Performance comparison between supervised SingleSiteMLP and unsupervised models.

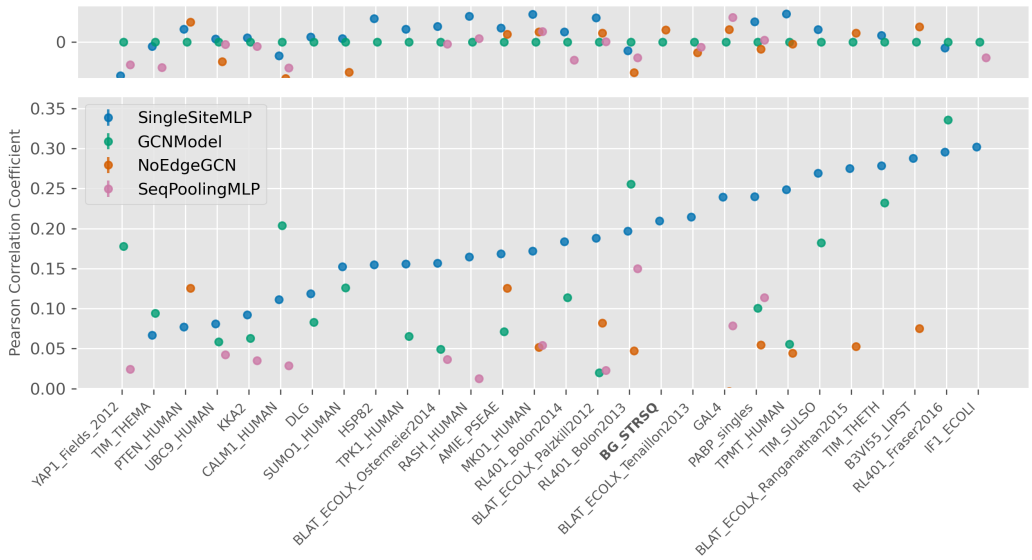

Figure A7: Performance evaluation when trained on onehot encoding. (Lower) PCC evaluated on 28 datasets ranked by SingleSiteMLP performance. (Upper) Differences in PCC relative to geometric model. Only samples with PCC > 0 were visualized.

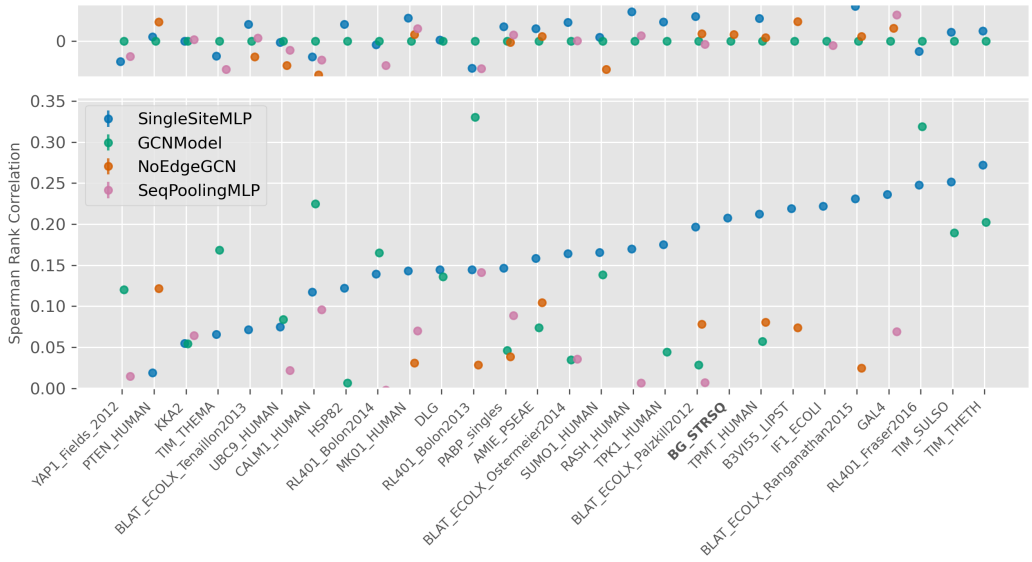

Figure A8: Performance evaluation when trained on onehot encoding. (Lower) SRC evaluated on 28 datasets ranked by SingleSiteMLP performance. (Upper) Differences in SRC relative to geometric model. Only samples with SRC > 0 were visualized.

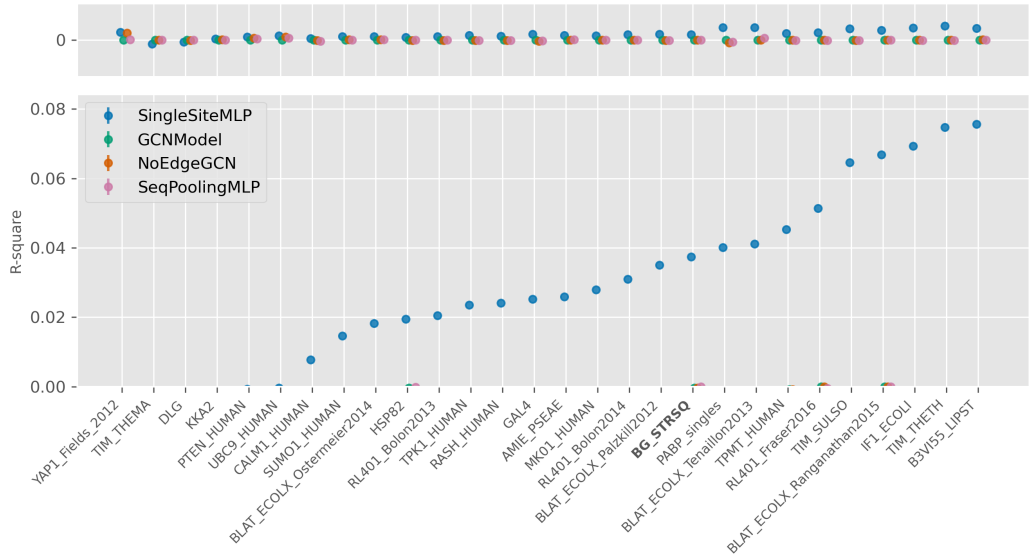

Figure A9: Performance evaluation when trained on onehot encoding. (Lower) R-square evaluated on 28 datasets ranked by SingleSiteMLP performance. (Upper) Differences in R-square relative to geometric model. Only samples with R-square > 0 were visualized.

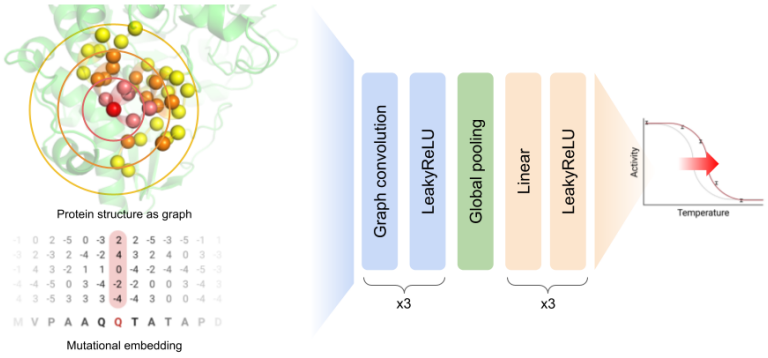

Figure A10: Illustration of GCNModel architecture. GCNModel takes protein graph with mutational embedding as input and passes it through three layers of GCN, global pooling, and three-layer MLP to predict the mutational change of stability.

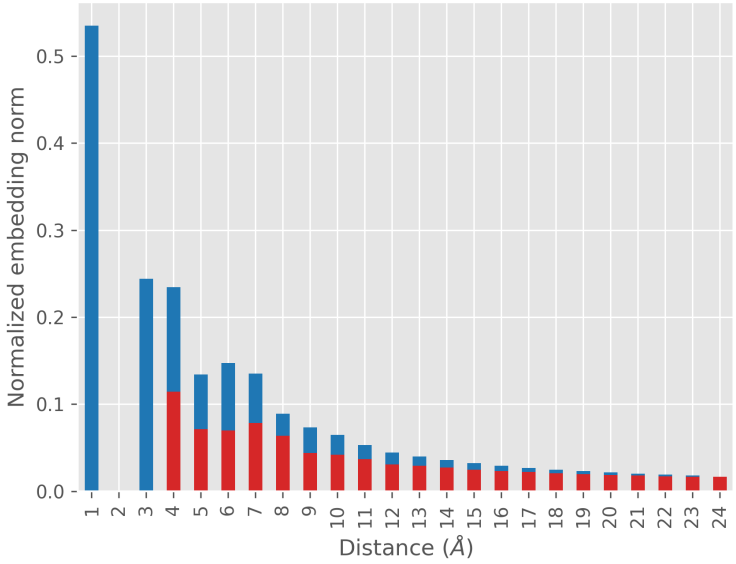

Figure A11: Average mutational embedding norm on each residue (normalized) versus geometric distance from mutation site. Blue bars represent statistics on all residues averaged over all mutations. Red bars include residues geometrically close but non-adjacent on sequence.

Table A1: Predicted lddt on AlphaFold2 model used in protein graph construction.

Dataset	plddt
AMIE_PSEAE	97.96
B3VI55_LIPST	95.18
BG_STRSQ	93.66
BLAT_ECOLX_Ostermeier2014	94.91
BLAT_ECOLX_Palzkill2012	94.79
BLAT_ECOLX_Ranganathan2015	94.69
BLAT_ECOLX_Tenaillon2013	94.89
CALM1_HUMAN	84.91
DLG	76.46
GAL4	67.13
HSP82	85.15
IF1_ECOLI	86.51
KKA2	93.83
MK01_HUMAN	90.00
PABP_singles	81.04
PTEN_HUMAN	82.90
RASH_HUMAN	92.18
RL401_Bolon2013	94.15
RL401_Bolon2014	94.00
RL401_Fraser2016	93.88
SUMO1_HUMAN	78.59
TIM_SULSO	95.84
TIM_THEMA	95.53
TIM_THETH	97.05
TPK1_HUMAN	97.96
TPMT_HUMAN	95.03
UBC9_HUMAN	97.00
YAP1_Fields_2012	56.78

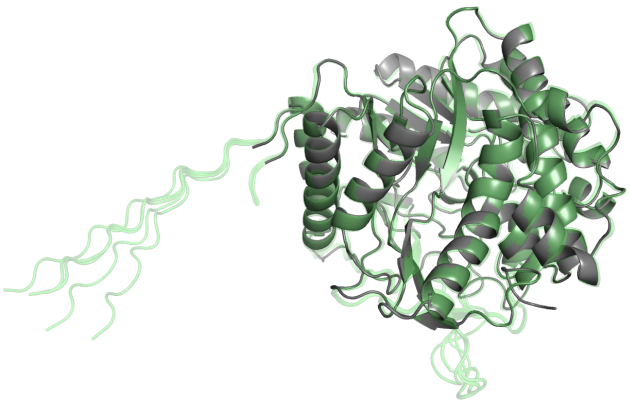

Figure A12: Alphafold2 models in green aligned to beta-glucosidase (bgl3) crystal structure in gray (PDB code: 1gnx). The average root-mean-square deviation (RMSD) from the model to crystal structure is 0.56 Å.