Predicting protein variants with equivariant graph neural networks

Antonia Boca [1]   **Simon Mathis** [1]

Abstract

Pre-trained models have been successful in many protein engineering tasks. Most notably, sequence-based models have achieved state-of-the-art performance on protein fitness prediction while structure-based models have been used experimentally to develop proteins with enhanced functions. However, there is a research gap in comparing structure- and sequence-based methods for predicting protein variants that are better than the wildtype protein. This paper aims to address this gap by conducting a comparative study between the abilities of equivariant graph neural networks (EGNNs) and sequence-based approaches to identify promising amino-acid mutations. The results show that our proposed structural approach achieves a competitive performance to sequence-based methods while being trained on significantly fewer molecules. Additionally, we find that combining assay labelled data with structure pre-trained models yields similar trends as with sequence pre-trained models.

Our code and trained models can be found at: https://github.com/semiluna/partIII-amino-acid-prediction.

1. Introduction

In recent years, pre-trained models have garnered significant attention in the field of protein representation. Notably, models have been developed to deal with both the sequence and structure modalities of proteins (Rives et al., 2021; Elnaggar et al., 2022; Zhang et al., 2023). These models have demonstrated their potential in various applications such as protein fitness prediction (Meier et al., 2021; Notin et al., 2022) while being employed in a "zero-shot" manner, without the need for additional training data. Their success has

also shown promising experimental results in protein engineering (Shroff et al., 2020; Lu et al., 2022). Additionally, Hsu et al. (2021) have observed that augmenting simple models for assay labelled data with fitness predictions extracted from pre-trained sequence models can enhance their performance.

Despite the experimental success of pre-trained structural methods for protein engineering, particularly those based on predicting residues given local atom environments (Torng & Altman, 2017; Lu et al., 2022), several crucial aspects remain unexplored. Firstly, these methods have not been systematically compared with sequence-based approaches using the same datasets. Secondly, their potential to augment assay labelled data, when available, has not been evaluated.

This paper aims to fill this research gap by conducting a study of the comparative performance of structure-based and sequence-based methods on predicting variants *that are better than the wildtype protein*. We compare representatives of the most successful equivariant graph neural networks (EGNNs) on the task of residue identity prediction, namely GVP (Jing et al., 2021) and EQGAT (Le et al., 2022), with representatives of the most successful sequence-based approaches: Tranception (Notin et al., 2022), ESM-1v (Meier et al., 2021) and the MSA Transformer (Rao et al., 2021).

By undertaking this comparative analysis, we aim to provide insights into the performance and suitability of geometric GNNs in protein engineering, specifically in the context of predictions based on the local atomic environment. Our contributions are as follows:

- We apply the most successful pre-training approach for structural methods (Shroff et al., 2020) to equivariant GNNs by using the ATOM3D RES dataset (Townshend et al., 2022) for residue identity prediction (Table 1);

- We benchmark the resulting structure-based pre-trained models with the most successful zero-shot sequence-based approaches (Table 2). We observe that structure does not trump sequence in downstream tasks when used in this way, although the amount of available structures used during pre-training is significantly lower than the number of sequences used in training large language models;

---

[1]University of Cambridge. Correspondence to: Antonia Boca <bocaantonia@gmail.com>, Simon Mathis <simon.mathis@cl.cam.ac.uk>.

Table 1. Classification accuracies on the ATOM3D RES dataset.

MODEL	REPORTED TEST ACCURACY	OUR TEST ACCURACY
EQGAT	0.540	0.524
GVP	0.527	**0.580**

- We extend the simple combination approach for assay labelled data and pre-trained model outputs (Hsu et al., 2021) to the structure pre-trained domain. We find the same general trends as with sequence pre-trained models, as assay-labelled data quickly allows us to surpass zero-shot pre-trained sequence-based models with at few as 100 datapoints (Figure 2).

2. Methodology

We pre-train two equivariant graph neural networks on the task of residue identity prediction, also known as the RES task (Townshend et al., 2022). We choose the Geometric Vector Perceptron (Jing et al., 2021) and the Equivariant Graph Attention Network (Le et al., 2022). While Lu et al. (2022) used 3D-CNNs to engineer plastic enzymes, Jing et al. (2021) benchmark 3D-CNNs on the RES task and show that the GVP outperforms them, so we choose to focus on this structural method instead.

Table 1 shows a comparison between the reported accuracies of the two models and the accuracies achieved in this paper. We achieve a higher performance on the GVP model than originally reported in Jing et al. (2021). This jump in performance can be explained by the fact that Jing et al. (2021) only use a third of the original training dataset to train the GVP, possibly due to computational constraints. More details on our training parameters can be found in A.9.

2.1. RES task formalism

We formalise the RES classification task as follows. For a given point-cloud atomic graph $G = (V, E)$ with nodes $i, j \in V$ and edges $(i \to j) \in E$. Given a node $t \in V$ representing the C_α of a residue in the atomic graph, we can define the *node classification function* RES $: \mathcal{V} \times \mathcal{G} \to \mathbb{R}^{20}$ that takes as input node t and a *masked* atomic graph G_t from which we have removed the side-chain atoms of node t and returns the likelihood scores of each of the 20 naturally occurring amino-acids to be part of the side-chain of node t. A more extended version of this formalism can be found in A.1.

2.2. The scoring function

We now formalise the function we use to score each amino-acid mutation in a sequence. For a wildtype protein se-quence $x_1 \ldots x_n$ with $x_i \in \mathcal{A} = \{1, 2, \ldots, 20\}$ we associate the point-cloud atomic graph $G = (V, E)$ corresponding to the protein's structure. Edges are drawn between any two atoms that are less than 4.5Å apart.

Then, using the formalism defined in 2.1, the score associated with the presence of amino-acid $a \in \mathcal{A}$ at position i can be defined as:

$$S(i, a) = [\text{RES}(g(i), G)]_a \qquad (1)$$

Where $g : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, |V|\}$ is a mapping function from positions to the index of the node representing the central C_α of the amino-acid residue present at each position. Here, $G_{g(i)}$ denotes the masked graph from which we removed the side-chain attached to node $g(i)$.

Equation 1 essentially represents the score of amino-acid a for target position i, associated with node $g(i)$ in the atomic graph. Note that the true amino-acid at the same position is denoted by x_i.

2.3. Mutation generation

Once the equivariant models have been trained on the RES task, we use them to inform the generation of single-point mutations in monomers and homo-oligomers from the Pro-teinGym substitutions dataset (Notin et al., 2022). For each wildtype sequence we recover its structure, mask each amino-acid residue in turn, and retrieve the scores generated by the EGNN model for each of the 20 naturally occuring amino-acids. These scores are then ranked according to two strategies to determine the most promising single-point mutations. Figure 1 illustrates this approach visually.
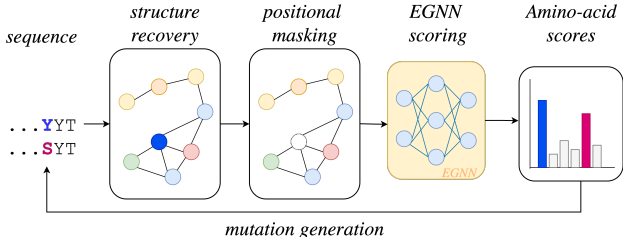

Figure 1. For every sequence, we recover the structure from the PDB and mask each amino-acid in turn. We pass the masked graph through a pre-trained EGNN model to recover the score associated with each amino-acid, which we then rank. The key idea is that this pre-training allows the model to identify amino acids which seem "unusual" given their local environment and propose better fitting candidates instead.

Structure recovery. The ProteinGym substitutions dataset contains 87 molecular sequences; for each of these sequences, a number of experimentally tested mutations are scored according to their *fitness*. We evaluate our methods

on a subset of the original dataset for which we could find either monomeric or homo-oligomeric structures. For each wildtype sequence, we recover the corresponding biological assembly from the Protein Data Bank (Berman et al., 2000). When multiple assemblies are available, we choose one at random. When assemblies are incomplete, we instead use the monomeric AlphaFold prediction (Jumper et al., 2021) if available. Otherwise, we discard the sequence.

2.4. Mutation ranking

Our approach allows us to score every possible residue mutation for each position in a sequence. Our goal is to generate meaningful mutations that have a higher chance of being bio-physically relevant, so we discard positions where the equivariant model makes the wrong prediction. A more detailed analysis of this design choice can be found in Appendix A.4.

Global ranking. We rank the remaining mutations according to two strategies: *global* and *positional*. When performing global ranking, we sort mutations in descending order of their score, regardless of their position. If we denote the single-point mutation to amino-acid a at position i by $\mathbf{m}_i^a$, then $\forall i, j$ and $\forall a, b \in \mathcal{A}$ s.t. $a \neq x_i$ and $b \neq x_j$, we say that:

$$\mathbf{m}_i^a \text{ is better than } \mathbf{m}_j^b \iff S(i, a) > S(j, b) \quad (2)$$

Positional ranking. The second approach follows when we prioritise the positions we want to mutate instead of the amino-acids we mutate to. Formally, this can be quantified as:

$$\begin{aligned} &\mathbf{m}_i^a \text{ is better than } \mathbf{m}_j^b \\ \iff &\Big(S(i, x_i) < S(j, x_j) \Big) \vee \\ &\Big(S(i, x_i) = S(j, x_j) \wedge S(i, a) > S(j, b) \Big) \end{aligned} \quad (3)$$

Note that when we perform positional ranking, we only keep the 3 top mutations for each position.

2.5. Protein fitness prediction

The GVP and EQGAT trained on the ATOM3D RES task can be thought of as unsupervised models that can suggest amino-acid mutations. We extend our original approach to perform fitness prediction using a ridge regression model augmented with the positional scores generated by equivariant GNNs, in a similar manner to that introduced by Hsu et al. (2021). For a given sequence of amino-acids $x_1 \ldots a_i \ldots x_n$ with a single-point mutation at position i, we embed each amino-acid using either the one-hot encoding or *AAIndex* embeddings (Kawashima et al., 1999) on which we perform PCA to render 19-dimensional features per amino-acid. We flatten and concatenate these

encodings to render feature vectors $\mathbf{h}_{\text{one-hot}} \in \mathbb{R}^{20 \times n}$ and $\mathbf{h}_{\text{aa-index}} \in \mathbb{R}^{19 \times n}$. To this feature vector we concatenate the score predicted by the GNN model for amino-acid a_i at position i:

$$\mathbf{x}_{\text{one-hot}} = [\mathbf{h}_{\text{one-hot}} \,||\, S(i, a_i)] \quad (4)$$
$$\mathbf{x}_{\text{aa-index}} = [\mathbf{h}_{\text{aa-index}} \,||\, S(i, a_i)] \quad (5)$$

Here, $S(i, a_i)$ is the same scoring function defined in Equation 1. These features are then used to train a ridge regression model to predict protein fitness using subsets of single-point mutated sequences for each of the ProteinGym DMS assays we have model scores for.

3. Results

3.1. Mutation generation

We generate single-point mutations for 49 out of the 87 DMS assays in the ProteinGym substitutions dataset (Notin et al., 2022). When we generate mutations, we discard any that we cannot find in the experimental dataset of the target sequence. We are interested in understanding how good our models are at suggesting mutations that are *better than the wildtype* sequence, hence we propose three metrics through which to perform comparisons: (1) Spearman's rank correlation restricted to better than wildtype sequences, (2) the precision of the top 10 mutations, and (3) the recall of the top 10 mutations. To compute the last two metrics we only considered whether a mutation proposed by the model is better than the wildtype, disregarding its actual score. Table 2 shows the performance of our models, depending on the type of ranking used. We note that the equivariant models have a competitive performance to Tranception (Notin et al., 2022) when ranking mutations that are better than the wildtype, indicating that they represent a viable strategy for aiding the discovery process in protein engineering. Per-dataset performance metrics can be found in A.2.

EGNN models require a significantly smaller number of protein structures during training in order to reach a similar ranking correlation coefficient to Tranception for mutations that are better than the wildtype. While Tranception is trained on the UniRef100 database (Suzek et al., 2015), which contains over 4 million source sequences, our models are trained on the ATOM3D RES dataset (Townshend et al., 2022), which contains fewer than 22k molecules from which local environments are sampled.

Structure vs Sequence. We believe EGNNs may require less training because structure is more informative than sequence for fitness prediction. While sequence-based models attend the full protein and subsequently learn to focus on the important bits, EGNNs attend only local environments, thus learning to identify important atoms faster. Further experiments could be run to compare the power of sequence-

Table 2. Ranking performance of the models across 49 DMS assays. Numbers in **bold** represent the highest score per column, while numbers with an <u>underline</u> represent the second highest score per column. We note that two equivariant GNNs have the highest rank correlation for better than wildtype mutations.

MODEL	RANKING STRATEGY	TOP 10 PRECISION	TOP 10 RECALL	SPEARMAN'S RANK CORRELATION		
				AVERAGE	WORSE THAN WT	BETTER THAN WT
EQGAT	POSITIONAL	0.486	<u>0.187</u>	0.223	0.128	0.118
EQGAT	GLOBAL	0.491	0.072	0.262	0.154	<u>0.157</u>
GVP	POSITIONAL	0.462	**0.419**	0.106	−0.009	**0.276**
GVP	GLOBAL	0.426	0.100	0.202	0.128	−0.011
TRANCEPTION		<u>0.619</u>	0.012	<u>0.429</u>	0.299	0.143
ESM-1V		0.618	0.018	0.407	0.288	0.135
MSA TRANSFORMER		**0.638**	0.018	**0.434**	**0.327**	0.135

and structure-based models when their level of training is comparable. However, we point out that training EGNNs to the same level as present state-of-the-art sequence models may be infeasible due to both data and computational constraints.

Correlation to sequence-based models. As part of our analysis, we also compute the correlation between the better than wildtype predictions made by our EGNN models and Tranception. Per-model and per-dataset statistics can be found in A.3; we note that the highest rank correlation we find is **0.212**, in the case of the EQGAT model. Since these approaches seem to be weakly correlated, we believe there are improvements to be gained from ensembling both structure- and sequence-based approaches.

The impact of design choices. As mentioned in Section 2.4, we discard mutations at positions where the EGNNs make the wrong prediction, as we find that incorporating these is detrimental to the overall performance (see A.4). This indicates that these structure-based models are still undertrained, with potential for improvement coming both from larger datasets and more data engineering.

3.2. Protein fitness prediction

We train 4 types of ridge regression models on each of the 49 DMS datasets separately. The baseline non-augmented model uses only features $\mathbf{h}_{\text{one-hot}}$ or $\mathbf{h}_{\text{aa-index}}$ defined in Section 2.5; the remaining 3 models are augmented with single-point mutation scores from GVP, EQGAT, and Tranception, respectively. For each model type and each DMS array we first set aside 20% of the single-point mutated sequences for testing; We train the regression on increasingly larger training subsets. We repeat the process 20 times with different random subsets and report the average Spearman rank correlation on better than wildtype sequences, as seen in Figure 2. The performance on other metrics can be found in A.8.
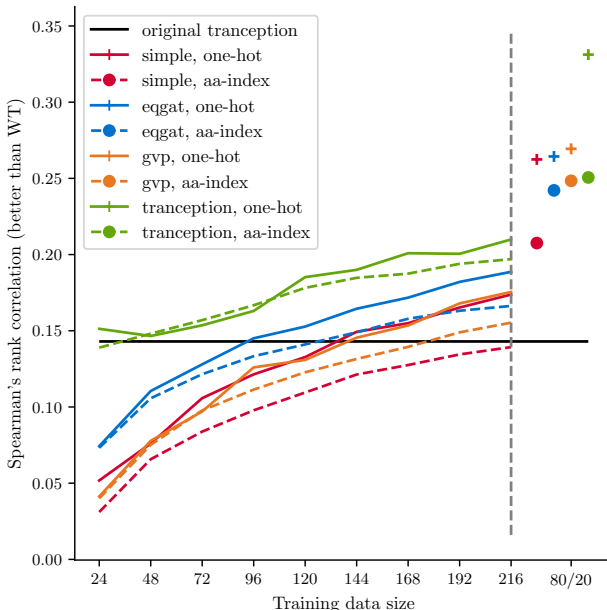

Figure 2. Performance on mutations that are better than the wild-type for four regression models using two types of embeddings. Statistics are aggregated across 49 DMS assays. We note that we can improve the fitness prediction performance above the Trancep-tion baseline (in black) across all regression models by training on as few as 144 data points.

Similar to the results reported by Hsu et al. (2021), the augmented linear models allow us to surpass the baseline zero-shot fitness prediction models with as few as 100 datapoints in the case of the model augmented with EQGAT scores. While the linear model augmneted with Tranception scores performs best overall, we point out that Tranception is fine-tuned to predict protein fitness, while the scores retrieved from our models merely represent the confidence in a certain amino-acid for a target position.

4. Limitations and future work

We apply pre-trained EGNNs to both mutation generation and protein fitness prediction, and find that structural approaches are a competitive approach to sequence-based language models for the prediction of mutations that are better than the wildtype, while also requiring **181x** fewer molecules to train.

While the results look promising, this comparison is limited in scope, as our approach does not deal with more complex (hetero-oligomeric) structures from the ProteinGym dataset.

Types of fitness. Additionally, the benchmarking dataset contains a wide range of sequences for which "fitness" can be interpreted in many different ways. DMS assays in the ProteinGym dataset come from humans, viruses, prokaryotes, and eukaryotes. In particular, in the subset of 49 DMS assays used in this paper, **5** come from eukaryotes, **21** from humans, **18** from prokaryotes, and **5** from viruses. Fitness, in the case of viruses, is interpreted as infectivity or the likelihood of mutation. In the rest of the cases, fitness can range from stress resistance to efficiency. For example, in their experimental paper, Lu et al. (2022) focused on improving thermal stability. Hence, the fitness score used by Notin et al. (2022) in the ProteinGym dataset represents a "fuzzy" concept that is context-dependent. Future work could focus more closely on identifying the types of fitness structure-based approaches excel at.

References

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. The protein data bank (2000). *Nucleic Acids Research*, 28 (1):235–242, 2000. URL https://doi.org/10.1093/nar/28.1.235.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, October 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3095381. URL https://doi.org/10.1109/TPAMI.2021.3095381.

Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919, Nov 1992. ISSN 0027-8424. doi: 10.1073/pnas.89.22.10915. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC50453/. Comparative Study, Journal Article, Research Support, U.S. Gov't, P.H.S.

Hsu, C., Nisonoff, H., Fannjiang, C., and Listgarten, J. Combining evolutionary and assay-labelled data for protein fitness prediction. *bioRxiv*, 2021. doi: 10.1101/2021.03.28.437402. URL https://www.biorxiv.org/content/early/2021/03/29/2021.03.28.437402.

Jing, B., Eismann, S., Soni, P. N., and Dror, R. O. Equivariant graph neural networks for 3d macromolecular structure. *arXiv preprint arXiv:2106.03843*, 2021.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Kawashima, S., Ogata, H., and Kanehisa, M. Aaindex: Amino acid index database. *Nucleic Acids Res*, 27:368–9, Jan 1999. doi: 10.1093/nar/27.1.368. URL https://doi.org/10.1093/nar/27.1.368.

Le, T., Noe, F., and Clevert, D.-A. Representation learning on biomolecular structures using equivariant graph attention. In *The First Learning on Graphs Conference*, 2022. URL https://openreview.net/forum?id=kv4xUo5Pu6.

Lu, H., Diaz, D. J., Czarnecki, N. J., Zhu, C., Kim, W., Shroff, R., Acosta, D. J., Alexander, B. R., Cole, H. O., Zhang, Y., Lynd, N. A., Ellington, A. D., and Alper, H. S. Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature*, 604(7907):662–667, Apr 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04599-z. URL https://doi.org/10.1038/s41586-022-04599-z.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

Notin, P., Dias, M., Frazer, J., Hurtado, J. M., Gomez, A. N., Marks, D., and Gal, Y. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.),

Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pp. 16990–17017. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/notin22a.html.

Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T., and Rives, A. Msa transformer. *bioRxiv*, 2021. doi: 10.1101/2021.02.12.430858. URL https://www.biorxiv.org/content/early/2021/02/13/2021.02.12.430858.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Shroff, R., Cole, A. W., Diaz, D. J., Morrow, B. R., Donnell, I., Annapareddy, A., Gollihar, J., Ellington, A. D., and Thyer, R. Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS Synthetic Biology*, 9(11):2927–2935, 2020. doi: 10.1021/acssynbio.0c00345. URL https://doi.org/10.1021/acssynbio.0c00345.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6): 926–932, Mar 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739. URL https://doi.org/10.1093/bioinformatics/btu739.

Torng, W. and Altman, R. B. 3d deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinformatics*, 18(1):302, 2017. doi: 10.1186/s12859-017-1702-0. URL https://doi.org/10.1186/s12859-017-1702-0.

Townshend, R. J. L., Vögele, M., Suriana, P., Derry, A., Powers, A., Laloudakis, Y., Balachandar, S., Jing, B., Anderson, B., Eismann, S., Kondor, R., Altman, R. B., and Dror, R. O. Atom3d: Tasks on molecules in three dimensions, 2022.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining, 2023.

A. Appendix

A.1. Extended RES task formalism

For a given point-cloud atomic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes $i, j \in \mathcal{V}$ and edges $(i \to j) \in \mathcal{E}$ for which we have initial scalar and vector node features $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}| \times n} \times \mathbb{R}^{|\mathcal{V}| \times 3 \times \nu}$, as well as scalar and vector edge features $\mathbf{E} \in \mathbb{R}^{|\mathcal{E}| \times m} \times \mathbb{R}^{|\mathcal{E}| \times 3 \times \eta}$, we first consider a *masked* version of this graph, $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$, for which we have masked all the atoms of the side-chain attached to the node t (representing an C_α atom).

These features become the input to an EGNN model trained on the RES task, that returns the probability that the masked residue t is any of the 20 naturally occurring amino-acids. Formally, if we define the *node classification function* f_γ^t : $\mathbb{R}^{|\mathcal{V}| \times n} \times \mathbb{R}^{|\mathcal{V}| \times 3 \times \nu} \times \mathbb{R}^{|\mathcal{E}| \times m} \times \mathbb{R}^{|\mathcal{E}| \times 3 \times \eta} \to \mathbb{R}^{20}$ with learnable parameters γ, then the score of amino-acid a at position i in a wildtype sequence can be defined as:

$$S(i, a) = [f_\gamma^{g(i)}(\mathbf{H}, \mathbf{E})]_a \tag{6}$$

We then formally define the GNN model $\mathbf{G}_{\theta_1}$: $\mathbb{R}^{|\mathcal{V}_t| \times n} \times \mathbb{R}^{|\mathcal{V}_t| \times 3 \times \nu} \times \mathbb{R}^{|\mathcal{E}_t| \times m} \times \mathbb{R}^{|\mathcal{E}_t| \times 3 \times \eta} \to \mathbb{R}^{|\mathcal{V}_t| \times o}$ that takes as input the node and edge features and returns final node features $\mathbf{H}_{\text{out}}$:

$$\mathbf{H}_{\text{out}} = \mathbf{G}_{\theta_1}(\mathbf{H}_t, \mathbf{E}_t) \tag{7}$$

where $\mathbf{H}_t$ and $\mathbf{E}_t$ represent the node and edge features for all nodes and edges that exist in the masked graph $\mathcal{G}_t$.

Since we are interested in predicting the type of amino-acid corresponding to the masked side-chain of node t, we pass its final features $[\mathbf{H}_{\text{out}}]_t$ through a multi-layer perceptron MLP_{θ_2} : $\mathbb{R}^o \to \mathbb{R}^{20}$ to obtain the final scores associated to each of the 20 naturally occuring amino-acids.

$$f_\gamma^t(\mathbf{H}, \mathbf{E}) = \text{MLP}_{\theta_2}([\mathbf{G}_{\theta_1}(\mathbf{H}_t, \mathbf{E}_t)]_t) \tag{8}$$

Given a wildtype sequence $x_1 x_2 \ldots x_n$ of length n with $x_i \in \mathcal{A} = \{1, 2, \ldots, 20\}$ representing the index of amino-acid i, we construct the atomic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and build a scoring function of the positions $S : \{1, 2, \ldots, n\} \times \mathcal{A} \to \mathbb{R}$ that we define by extending the formalism in Equation 8:

$$S(i, a) = [f_\gamma^{g(i)}(\mathbf{H}, \mathbf{E})]_a \tag{9}$$

Where $g : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, |\mathcal{V}|\}$ is a mapping function from positions to the index of the node representing the central C_α of the residue at every position in the sequence.

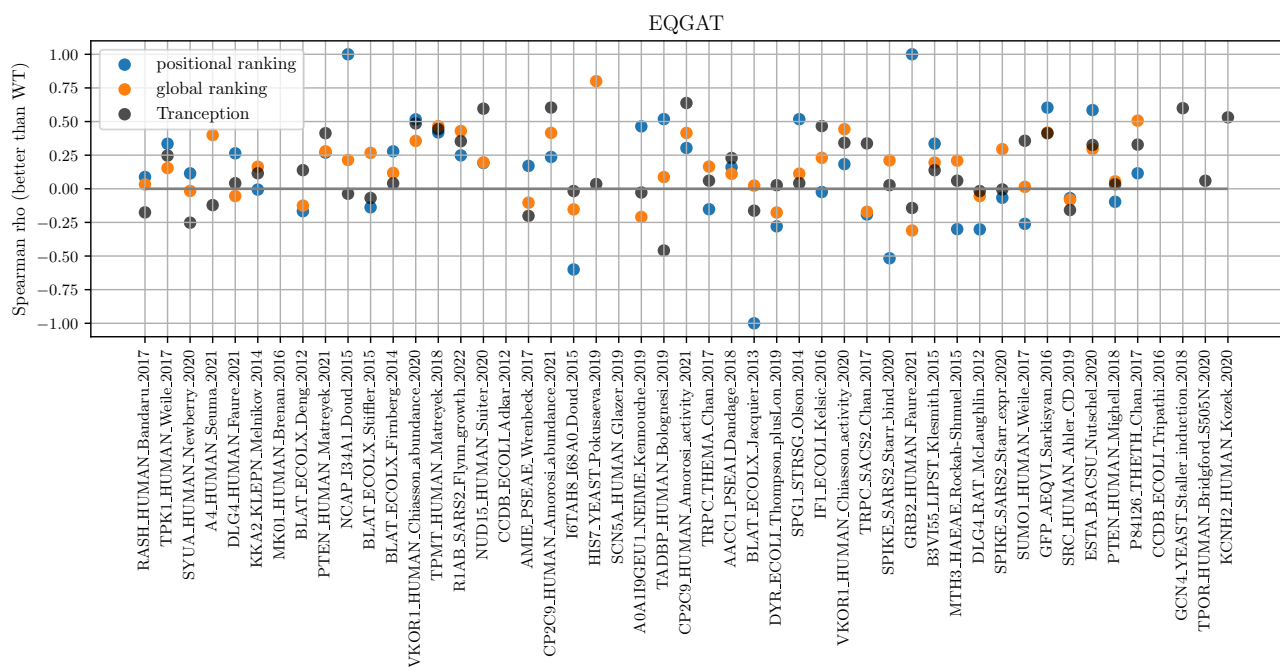## A.2. Per-dataset performance of our ranking approach

Figure 3. Spearman's rank correlation on better than wildtype mutations per dataset for the EQGAT model.
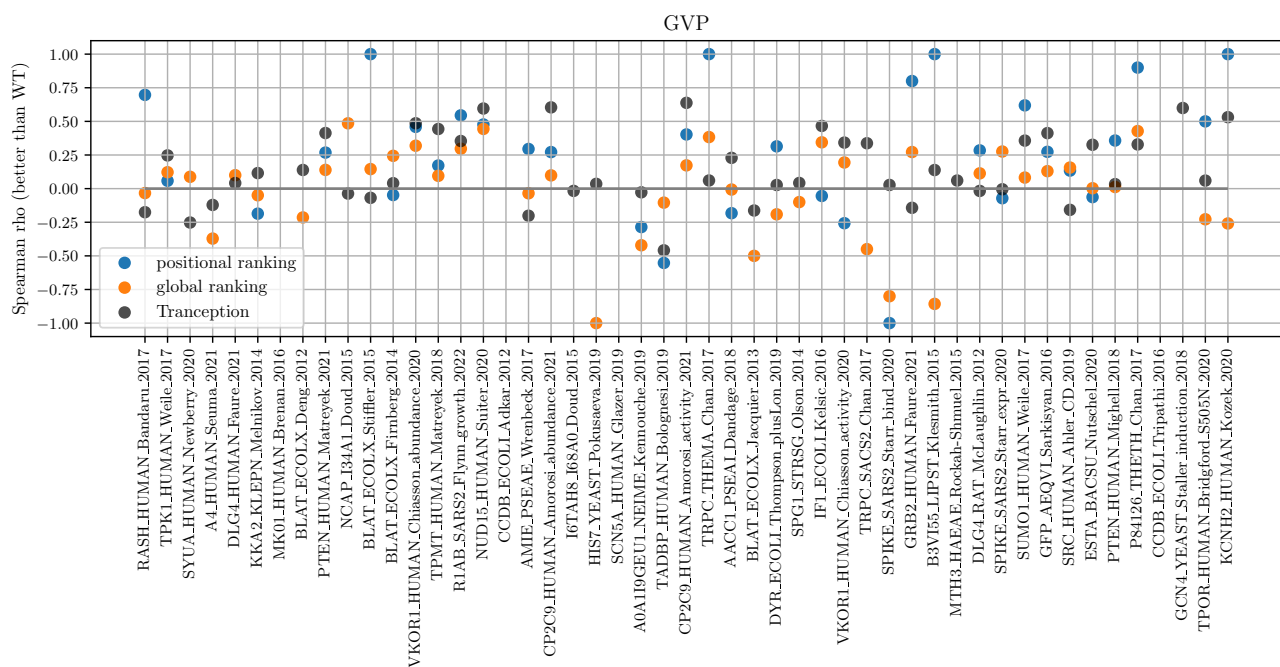

Figure 4. Spearman's rank correlation on better than wildtype mutations per dataset for the GVP model.

A.3. Correlation to Tranception ranking

Table 3. Average Spearman rank correlation (for better than wildtype predictions) between Tranception and structure-based models.

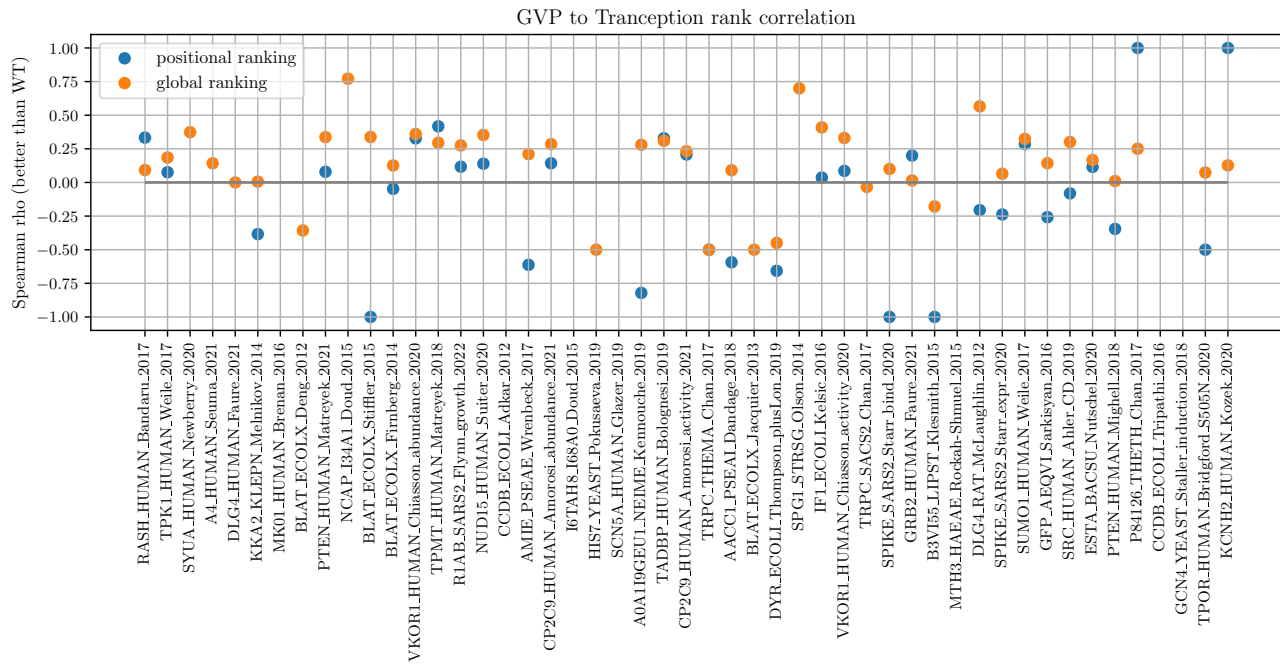| Model	Ranking	Correlation to Tranception
EQGAT	Global	0.212
	Positional	0.146
GVP	Global	0.147
	Positional	0.073

Figure 5. Spearman's rank correlation between the ranking made by the GVP model and Tranception.
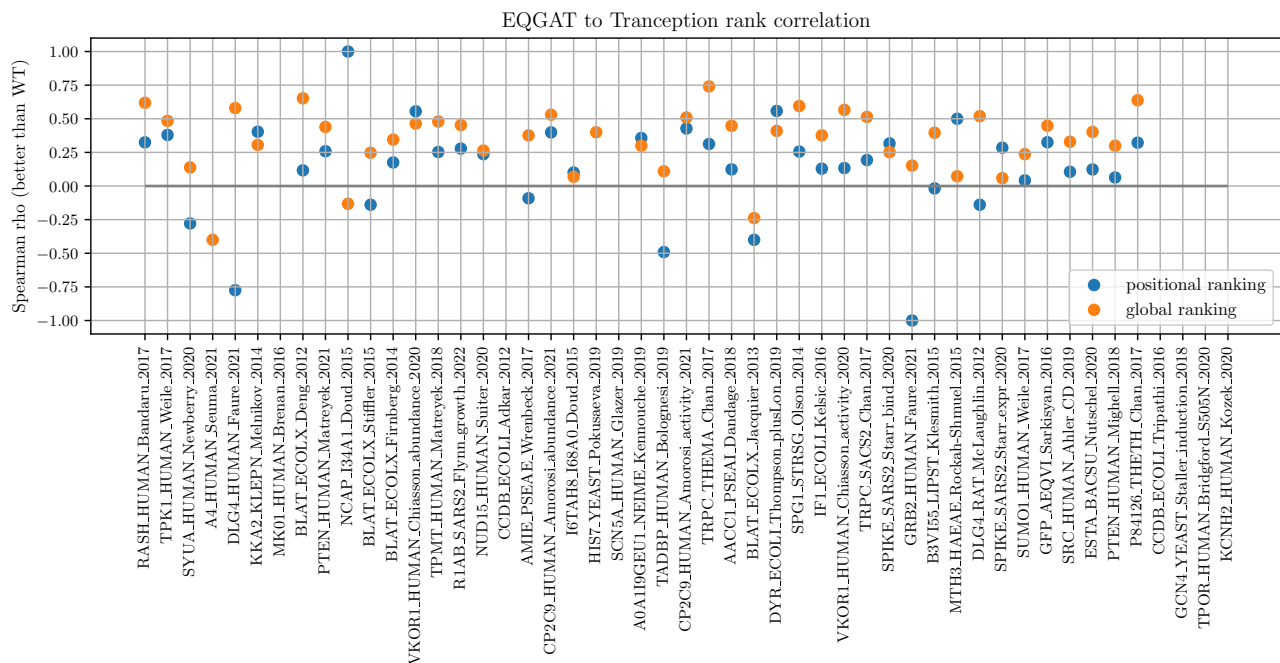
Figure 6. Spearman's rank correlation between the ranking made by the EQGAT model and Tranception.

A.4. Design choice: discarding wrongly predicted positions

As mentioned at the beginning of Section 2.4, we add an additional filter to our ranking techniques in order to increase the quality of the mutations proposed by our models. More specifically, we discard the scores of all mutations at a position that was classified incorrectly by the EGNN.

The reasoning behind this is that an incorrect classification of a residue may indicate that the models do not have a good understanding of the biophysical properties of the residue's local environment. This is particularly damaging to the generation of meaningful mutations if the models assign similarly low confidences to all 20 amino-acids that are candidate for a position, because they resort to random guessing.

We provide an ablation study to back up this design choice. Tables 4 and 5 show how all three types of Spearman rank correlations are higher when discarding wrongly predicted positions.

Table 4. Model performance when performing global ranking using both wrongly and correctly predicted positions. Statistics averaged across for 49 sequences.

MODEL	POSITIONS USED	SPEARMAN'S RANK CORRELATION		
		AVERAGE	WORSE THAN WT	BETTER THAN WT
EQGAT	ALL	0.20	0.153	0.069
EQGAT	CORRECT ONLY	**0.262**	**0.154**	**0.157**
GVP	ALL	0.093	0.076	**0.014**
GVP	CORRECT ONLY	**0.202**	**0.128**	-0.011
TRANCEPTION		0.429	0.299	0.143

We can get a better idea of the types of information that is learned by the EGNN models by looking at the confusion matrices between the true labels and the predicted labels in the RES ATOM3D test dataset, illustrated in Figure 7.

The BLOSUM62 matrix. We compare these two matrices to the BLOSUM62 matrix, a scoring matrix commonly used in bioinformatics for sequence alignment. It stands for "BLOcks SUbstitution Matrix" and was developed by Henikoff &

Table 5. Model performance when performing positional ranking using both wrongly and correctly predicted positions. Statistics averaged across for 49 sequences.

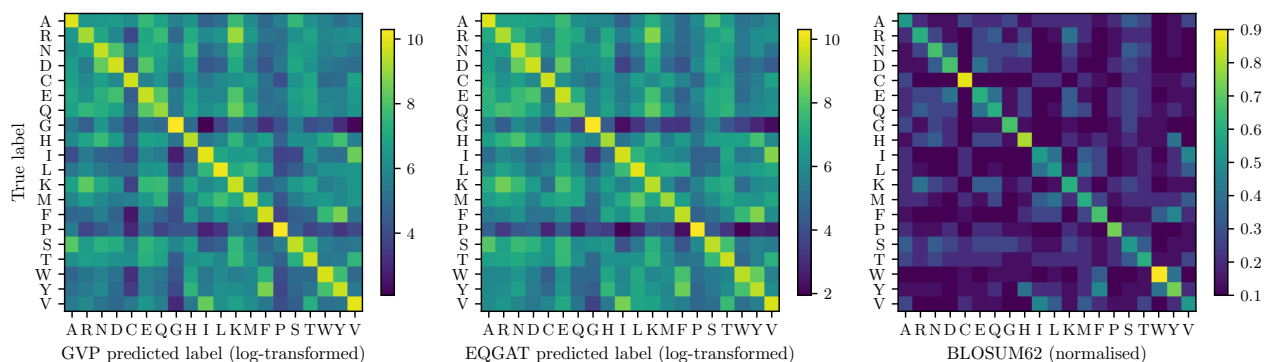| MODEL	POSITIONS USED	SPEARMAN'S RANK CORRELATION		
		AVERAGE	WORSE THAN WT	BETTER THAN WT
EQGAT	ALL	0.124	0.100	0.061
EQGAT	CORRECT ONLY	**0.223**	**0.128**	**0.118**
GVP	ALL	0.083	**0.076**	-0.018
GVP	CORRECT ONLY	**0.106**	-0.009	**0.276**
TRANCEPTION		0.429	0.299	0.143

Figure 7. Comparison between the confusion matrices of EQGAT and GVP to the BLOSUM62 matrix.

Henikoff (1992). In the BLOSUM62 matrix, each cell represents the score for substituting one amino acid with another. The matrix is symmetric, and the scores are typically positive integers. Higher scores indicate a higher degree of conservation or similarity between the substituted amino acids.

We expect that model with meaningful inferred knowledge about the biophysical properties of molecules and amino-acids to have a similar confusion matrix to the BLOSUM62 matrix, yet we do not necessarily see this trend arising. We do, however, notice that the confusion matrices of the EQGAT and the GVP look similar to each other, indicating that they must be learning the same features from the training dataset.

A.5. Design choice: experimental vs. AlphaFold structures

We expect the usage of AlphaFold structures to be detrimental to the overall performance of our models. To check whether this is indeed the case, we perform on ablation study on 14 of the 49 sequences we use in this project. These sequences have both a complete experimental structure and an AlphaFold structure, so we can compare the performance of our models using either one or the other. Since the ablation study presented in Section A.4 makes it clear that mutations at wrongly predicted positions are detrimental to our models, we perform this second ablation study by also discarding mutations for positions that models get wrong.

As presented in Tables 6 and 7, we find that there isn't a clear relation between using the AlphaFold structure and a decrease in the "better than WT" Spearman correlation, as it seeems to depend on both the model and the ranking strategy used. However, we *do* notice a clear trend for the models rank *worse than wildtype* mutations better when using the AlphaFold structure.

A.6. Design choice: full structure vs. local environment

For the main results, we input the entire molecule in the GNN (with the exception of the masked amino-acid position we wish to predict scores for). This may not necessarily be the best approach, because the models are trained on *samples of local environments*, which contain on average 600 nodes, whereas a full molecular structure can have even 4000 nodes.

Table 6. Model performance when performing global ranking using either AlphaFold or experimental features. Statistics are averages across 14 DMS assays.

MODEL	STRUCTURE	SPEARMAN'S RANK CORRELATION		
		AVERAGE	WORSE THAN WT	BETTER THAN WT
EQGAT	ALPHAFOLD	**0.311**	**0.177**	0.136
EQGAT	EXPERIMENTAL	0.262	0.154	**0.157**
GVP	ALPHAFOLD	**0.237**	**0.211**	**0.049**
GVP	EXPERIMENTAL	0.202	0.128	−0.011

Table 7. Model performance when performing positional ranking using either AlphaFold or experimental features. Statistics are averages across 14 DMS assays.

MODEL	STRUCTURE	SPEARMAN'S RANK CORRELATION		
		AVERAGE	WORSE THAN WT	BETTER THAN WT
EQGAT	ALPHAFOLD	**0.235**	0.097	**0.149**
EQGAT	EXPERIMENTAL	0.223	**0.128**	0.118
GVP	ALPHAFOLD	**0.253**	**0.332**	0.172
GVP	EXPERIMENTAL	0.106	-0.009	**0.276**

Tables 8 and 9 show that the EQGAT model benefits from using the entire structure, while the GVP model benefits from using the local environment.

Table 8. Model performance when performing positional ranking using either local environments or the full molecule. Statistics are averaged across 49 DMS assays.

MODEL	STRUCTURE	SPEARMAN'S RANK CORRELATION			TOP 10 PRECISION	TOP 10 RECALL
		AVERAGE	WORSE THAN WT	BETTER THAN WT		
EQGAT	FULL	**0.223**	**0.128**	**0.118**	0.486	**0.187**
EQGAT	LOCAL	0.203	0.039	0.041	**0.516**	0.176
GVP	FULL	0.106	-0.009	0.276	**0.462**	**0.419**
GVP	LOCAL	**0.203**	**0.104**	**0.311**	0.451	0.382

Table 9. Model performance when performing global ranking using either local environments or the full molecule. Statistics are averaged across 49 DMS assays.

MODEL	STRUCTURE	SPEARMAN'S RANK CORRELATION			TOP 10 PRECISION	TOP 10 RECALL
		AVERAGE	WORSE THAN WT	BETTER THAN WT		
EQGAT	FULL	**0.262**	**0.154**	**0.157**	0.491	**0.072**
EQGAT	LOCAL	0.254	0.149	0.134	0.491	0.047
GVP	FULL	0.202	0.128	**-0.011**	**0.426**	0.100
GVP	LOCAL	**0.216**	0.233	**-0.031**	0.392	**0.126**

A.7. Summary of design decisions for mutation generation

Overall, our ablation studies suggest the following:

- Discarding positions where the models cannot correctly identify the true wildtype amino-acid increases the quality the mutation ranking;

- Using the AlphaFold structure is *not* detrimental to the generation of better than wildtype mutations;

- The EQGAT model works best with the full molecular structure, while the GVP model works best with the a local environment.

A.8. Performance metrics for the ridge regression models

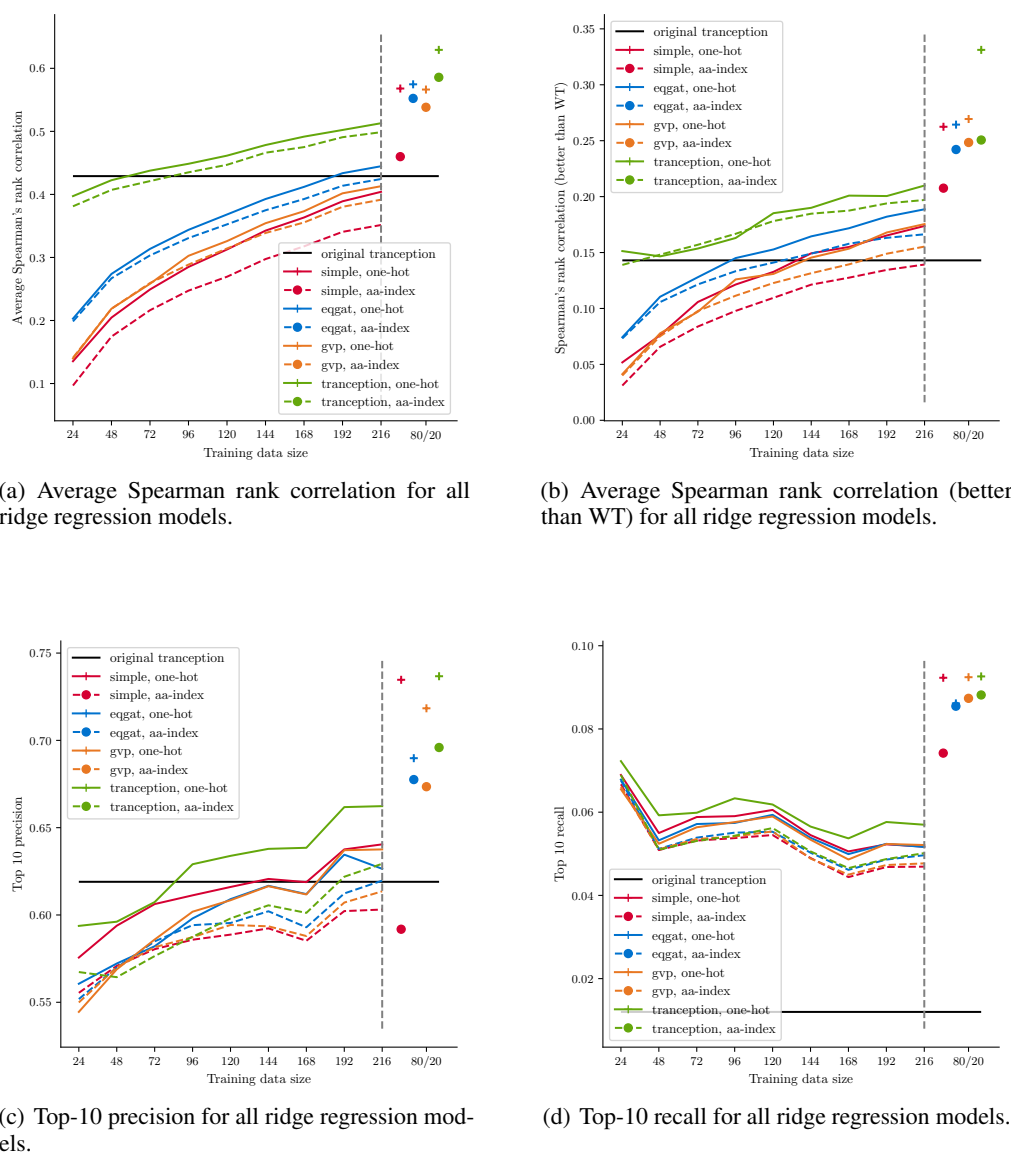Figure 8 summarises the performance of the ridge regression models.

(a) Average Spearman rank correlation for all ridge regression models.

(b) Average Spearman rank correlation (better than WT) for all ridge regression models.

(c) Top-10 precision for all ridge regression models.

(d) Top-10 recall for all ridge regression models.

Figure 8.

A.9. RES Task learning hyperparameters

We train two equivariant GNN models on the RES task using the ATOM3D RES dataset (Townshend et al., 2022). Both models are trained using one NVIDIA A100 GPU.

Table 10 summarises the training configuration used for both models. Table 11 summarises the architectures of the EQGAT and GVP.

Table 10. Training configuration for both models.

Hyperparameter	Value
Learning rate	$1e^{-4}$
Patience scheduler	10
Decay rate	0.75
Dropout	0.1
Batch size	64
Epochs	40

Table 11. Model architectures.

Model	Hyperparameter	Value
GVP	Message-passing layers	5
	Node features (scalar, vector)	(100, 16)
	Edge features (scalar, vector)	(32, 1)
EQGAT	Message-passing layers	5
	Node features (scalar, vector)	(100, 16)
	RBF function	Bessel
	RBFs	32
	RBF cutoff	4.5 Å