Euclidean Transformers for Macromolecular Structures: Lessons Learned

David D. Liu [1] Ligia Melo [1] Allan Costa [2] Martin Vögele [1] Raphael J. L. Townshend [1] Ron O. Dror [1]

Abstract

Recent years have seen significant efforts towards creating machine learning approaches for modeling molecular structure. In this work, we investigate a class of architectures of particular interest—translation- and rotation-equivariant transformers—across a number of important problems involving macromolecules with complex three-dimensional geometry. In particular, we build a representative model of this class that achieves state-of-the-art performance on a number of tasks in the ATOM3D collection. Surprisingly, we find that while equivariance is critical for achieving high performance, attention does not provide major improvements. We hope that these insights, combined with the overall robustness of the method, will help further machine learning architectural research on problems involving molecular structures. The model code is available out-of-the-box at https://github.com/drorlab/gert.

1. Introduction

Biomolecules such as proteins and RNAs adopt complex three-dimensional (3D) structures that enable fundamental biological processes, such as immune response and cell signaling. Accurately modeling how structures induce these phenomena is of interest for biologists and computer scientists alike and additionally provides practical guidance for drug discovery and design.

Recent years have seen a substantial increase in the number of catalogued macromolecular structures, in turn enabling the exploration of structural modeling through data-intensive methods, such as machine learning. The greater availability of structure predictions on individual desktops (Jumper et al., 2021; Baek et al., 2021) and data standardization through works such as ProteinNet (AlQuraishi, 2019), SidechainNet (King & Koes, 2020), MoleculeNet (Wu et al., 2018) and ATOM3D (Townshend et al., 2021) further enabled the investigation of biomolecules and the benchmarking of learning architectures, especially ones that operated on structural information.

In parallel to the growth of annotated data, advances in deep learning allowed for diverse architectural designs tackling tasks ranging from predicting protein-protein interfaces (Fout et al., 2017; Savojardo et al., 2020) or ligand binding affinities (Gomes et al., 2017; S Heck et al., 2017), to modeling mutation stability (Cao et al., 2019) or inverting the folding problem (Ingraham et al., 2019; Strokach et al., 2020; Hsu et al., 2022). This generation of architectures was largely inspired by the success of 3D convolutional neural networks, and later by the development of graph machine learning and the attention mechanism (Vaswani et al., 2017).

Learning architectures are able to better capture data when transformation structures of the problem domain need not to be learned. Such inductive bias can be implemented by constraining the representations and their interactions to satisfy symmetric properties of the domain. A model is equivariant to a specific transformation T if its produced representations behave cohesively with the action of T. Euclidean networks are models that implement roto-translation equivariance, and were demonstrated to achieve state-of-the-art results in molecular tasks (Eismann et al., 2020; Jing et al., 2021). Graph attention was incorporated in those networks (Fuchs et al., 2020; 2021) and similarly indicated competitive results.

In this work, we develop and extensively benchmark a rotation- and translation-equivariant transformer architecture for solving tasks on macromolecular structures. We call the model GERT, which stands for *G*eometric *E*ncoder *R*epresentations for *T*ransformers. This model is constructed from components introduced in (Fuchs et al., 2020) and (Thomas et al., 2018). In the same way that BERT (Devlin et al., 2018) provided embeddings for language, GERT represents an effort to provide embeddings for macromolecules. The model operates on individual atoms to construct these embeddings and can extract structural information at a finer resolution than residue-based models. We show that our geometric deep learning algorithm is *robust* over a broad range

[1] Department of Computer Science, Stanford University, Stanford, CA, USA [2] Center for Bits and Atoms, Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Raphael Townshend <raphael@cs.stanford.edu>, Ron Dror <ron.dror@stanford.edu>.

The 2022 ICML Workshop on Computational Biology. Copyright 2022 by the author(s).

of relevant tasks and achieves state-of-the-art or competitive results in benchmarks from ATOM3D; in other words, a single consistent set of GERT encoder hyperparameters yields performance at a level across tasks that is competitive with, if not better than the benchmarks. Moreover, we assess equivariant transformer components on molecular structure and show that equivariance is crucial towards model performance while attention is not as important, and provide initial evidence advocating for usage of structural models over sequence-based models. Overall, our work shows how to effectively learn on structural data and contributes tools for solving biomolecular problems.

2. Methods

2.1. Architecture

We design our architecture to operate over atomic labels and their 3D positions. The model (Figure 1) consists of two stages: an encoder component aimed at constructing a structural representation; and a task-specific head model that transforms these representations into the correct output for prediction. We use the same encoder architecture across all tasks, while the task-specific head varies from task to task based on the required output format. The encoder is specified by a *rotation order* parameter (Geiger et al., 2020); variants of the model with higher rotational orders are better for modeling higher angular resolution.

The encoder layer consists of alternating *equivariant attention* and *equivariant convolution* layers. The equivariant attention layers are equivalent to (Fuchs et al., 2020) and are computed over each atom in the region of interest in the molecule. The query matrix is generated by a linear transformation of the activations of the previous layer and the key matrix is generated by another equivariant convolution. The equivariant convolution layers are similar to the attention layers, but with the attention weights all held constant, which reduces them to the convolution layers described in (Thomas et al., 2018). These layers increase the capacity of the model to capture atom-atom interaction terms, a key driver of molecular structure.

Due to memory limitations, we are not always able to encode whole macromolecules in GERT at the atom-level, so we elect to focus on a region of interest for each macromolecule. For example, if we are interested in interactions happening at a certain amino acid in a protein, we focus on that residue's α-carbon and use only the atoms in the spherical region with radius R from the central α-carbon as our input signal. For some tasks, such as the residue deletion task, we represent the macromolecule at the residue level by preparing one atom (usually the α-carbon) per residue to pass as input into the model.

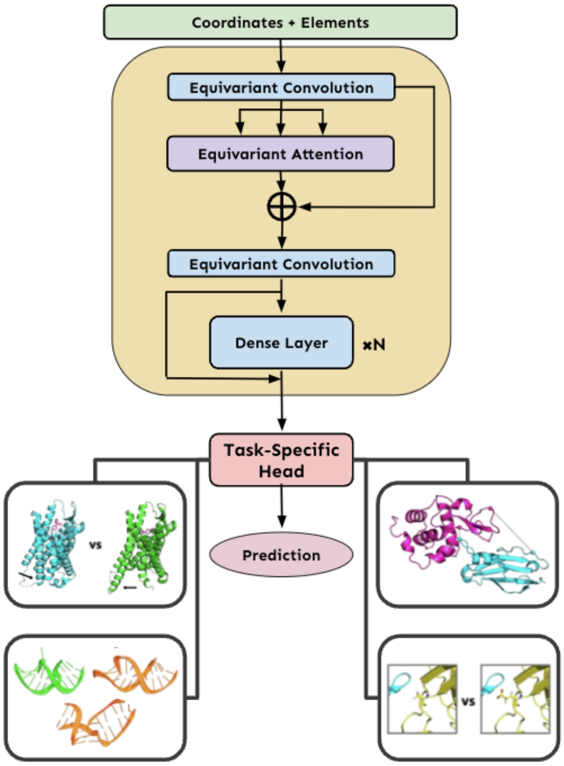

Figure 1. The GERT architecture. An encoder operates over three-dimensional coordinates and atomic labels to produce structural representations. These embeddings are fed into task-specific heads for direct supervised learning on different tasks. Encoder and decoder algorithms are described in more detail in Appendix A.

2.2. Datasets

We use the ATOM3D dataset (Townshend et al., 2021), a collection of 3D molecular tasks and benchmark GERT on all seven ATOM3D tasks that involve macromolecules. The datasets are split into training, validation and test sets by sequence identity. Data and training splits are described more thoroughly in Appendix B.

2.3. Training

One advantage of using GERT is that the architecture has a single set of encoder hyperparameters specifying the architecture, which is robust across all different tasks we experimented with (maximum rotation order of 4, 1 attention head, 1 encoder layer, 1 dense layer, and a consistent radial basis model used in the convolutions). For all tasks, we use the Adam optimizer (Kingma & Ba, 2014) and an initial learning rate of 1×10^{-4}. Task-specific heads necessarily vary based on the shape of the latent space representation (see Appendix A). The consistent configuration makes GERT easy to use out of the box for new tasks. The only differences between tasks are the number of training epochs. Training time is on the order of days for the largest ATOM3D datasets such as PIP with a single *NVIDIA TITAN Xp* GPU node due to computational limitations, because the high memory demand of equivariant representations restricts batch size to one structure at a time. With increased computational resources, we expect that training time can be reduced to just a few hours.

3. Results and Discussion

3.1. GERT is robust across many macromolecular tasks

To test our architecture, we evaluate GERT on all of the macromolecular datasets in ATOM3D and compare it to existing benchmarks. Table 1 summarizes the results of the experiments, benchmarking GERT against 3DCNN, GNN, and ENN (Anderson et al., 2019) reference architectures.

Many architectures are specifically engineered for a subset of macromolecular tasks, so they will perform well on those tasks and poorly on others, but GERT consistently has strong performance across all of the macromolecular tasks. The model outperforms the existing benchmarks on PIP (0.932 AUROC/0.891 on DB5) and MSP (0.642 AUROC) – both protein-protein interaction tasks – and delivers strong results consistently across all tasks. To the best of our knowledge, there is no other existing model that outperforms GERT on the protein interaction prediction (PIP) task. Even for tasks where GERT is not significantly out-performing all reference architectures, it is commonly still competitive with, if not exceeding, the best benchmark.

Additionally, for certain data-limited datasets such as MSP (only 4148 total mutant structures), some models, such as the 3DCNN, do not perform significantly better than random guessing (0.574 ± 0.005 AUROC), while other models such as the Cormorant ENN have unstable performance that is impacted by the initialization. GERT is more robust to random initialization compared to these baselines (0.642 ± 0.029 AUROC). This suggests that the model is helpful even on data-limited tasks. Since biological data is often scarce in nature, it is important to have data-efficient models, e.g., by explicitly encoding underlying physical symmetries. GERT is one such architecture and it works well on biomolecular systems in various contexts, even in regimes where data is not abundant.

3.2. Equivariance is crucial, but attention is interchangeable with convolutions

We investigate how the different components of the GERT architecture boost performance. In additional experiments, we test GERT across all macromolecular ATOM3D datasets while varying the maximum rotation order. Then, we perform the same set of experiments again, removing the attention layer in GERT. Throughout the experiments, we see that increasing the maximum rotation order significantly helps performance, but the attention layer does not provide a significant boost (Figure 2). We attribute this behavior to the convolutions used by the GERT architecture, which might be an apt substitute for attention since it also allows for interaction terms between molecules. This suggests that equivariant models that have weights dedicated to characterizing inter-atomic interactions perform similarly, and performance does not depend on how those weights are organized (e.g., in the convolution, or in attention).

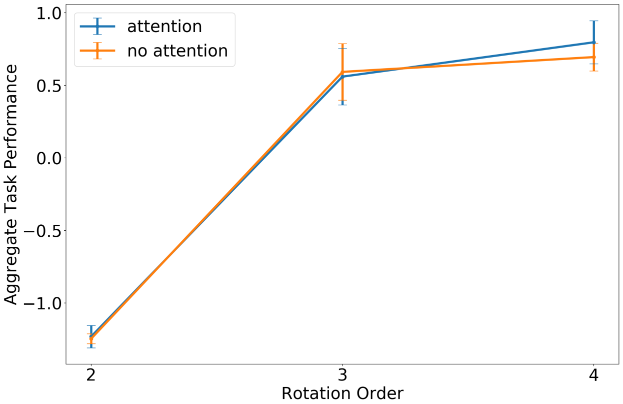

Figure 2. Attention and equivariance ablation study. The trend is that the *aggregate task performance* metric (an median of normalized per-task performances) increases with maximum rotation order and is not significantly affected by the presence of attention. Error bars are ± standard error. See Appendix C for details.

Our experiments on rotational orders indicate that performance increases significantly between maximum rotation orders 2 and 3 and then improves only marginally with higher orders than 3. This allows an assessment of whether the increase in performance is worth the additional effort in model complexity and training time. Since the size of the model increases quadratically with the maximum rotation order, it can be more practical to train a model of order 3 when resources are expensive or limited.

A non-equivariant version of our transformer performs worse than the equivariant transformer on all but one task (LEP). In this task, we had to cut down the maximum radius by 20% for the equivariant model due to memory limitations. The generally higher performance of the equivariant model suggests that having built-in understanding of basic physical symmetries improves prediction quality. The model can then learn additional information from the data beyond these symmetries.

To test the hypothesis that attention layers and equivariant convolutions largely served the same purpose of providing inter-atomic terms in the model, we built a similar non-equivariant version of GERT that does not use the equivariant convolutions, and similarly compared the performance of this model with and without the attention module in some

Table 1. GERT Performance on ATOM3D (Townshend et al., 2021) macromolecular benchmark tasks with standard deviations reported over three replicates, relative to baselines. Metrics are labeled with ↑ / ↓ if higher/lower is better, respectively. Results are reported as mean ± standard deviation over three training runs.

Task	Target	Metric	3DCNN	GNN	ENN	**GERT**
PSR	GDT-TS	mean R_S (↑)	**0.431 ± 0.013**	0.411 ± 0.006	—	0.279 ± 0.018
		global R_S (↑)	**0.789 ± 0.017**	0.750 ± 0.018	—	0.605 ± 0.017
RSR	RMSD	mean R_S (↑)	**0.264 ± 0.046**	**0.234 ± 0.006**	—	**0.236 ± 0.053**
		global R_S (↑)	0.372 ± 0.027	**0.512 ± 0.049**	—	0.318 ± 0.228
PIP	contacts	AUROC (↑)	0.844 ± 0.002	0.669 ± 0.001	—	**0.891 ± 0.005**
RES	res. type	accuracy (↑)	**0.451 ± 0.002**	0.082 ± 0.002	0.072 ± 0.005	0.168 ± 0.005
MSP	stab. incr.	AUROC (↑)	0.574 ± 0.005	0.609 ± 0.011	0.574 ± 0.040	**0.642 ± 0.029**
LBA	pK	RMSE (↓)	**1.416 ± 0.021**	1.601 ± 0.048	1.568 ± 0.012	1.453 ± 0.024
(30%)		global R_P (↑)	**0.550 ± 0.021**	**0.545 ± 0.027**	0.389 ± 0.024	**0.549 ± 0.010**
		global R_S (↑)	**0.553 ± 0.009**	0.533 ± 0.033	0.408 ± 0.021	**0.558 ± 0.008**
LEP	act./deact.	AUROC (↑)	0.589 ± 0.020	**0.681 ± 0.062**	**0.663 ± 0.100**	0.617 ± 0.054

toy experiments with smaller versions of the model (hidden dimension 32). The non-equivariant GERT model with 4 attention heads achieved a mean global R_S of 0.159 ± 0.010 on PSR, while the non-equivariant GERT without attention only achieved a mean global R_S of 0.146 ± 0.004. This difference in performance suggests that in regimes where the model is stripped of all other inter-atomic interactions, attention can be slightly helpful, especially on protein tasks like PSR where the relative atom locations are important indicators of structure. However, in general, most models will already have some built-in inter-atomic terms, and we were not able to find a regime where the presence of attention consistently and significantly boosted model performance.

3.3. Comparing structural models to sequence-based models

One open question in machine learning is whether geometric/structural models are viable alternatives to sequence models for macromolecular modeling. Primarily, some of the ATOM3D tasks (and many other important tasks aside from ATOM3D) can't be addressed in a meaningful way with sequence-based models. For example, PSR and RSR involve scoring different structural models with exactly the same sequence. Additionally, we benchmarked a standard language model, ESM-1b (Rives et al., 2019), with layer size 256 on PIP and compare the results to GERT, and the language model achieves across 10 trials an DB5 AUROC of 0.850 ± 0.082, while GERT achieves a DB5 AUROC of 0.891 ± 0.005. This result shows that even for tasks like PIP where language models are useful models, structural models like GERT still appear to edge them out in performance. Further experimentation is necessary to show that this holds in general; that is, in scenarios where the atomic structure is available, there is merit to favoring a model that operates on that structure over a sequence-based model.

4. Conclusion

In this work, we demonstrate the broad applicability of equivariant transformers to molecular tasks, covering tasks involving macromolecular structure, function, interaction, and design. These equivariant transformers have robust hyperparameters and architecture, as nearly identical configurations are applicable to a variety of tasks. This work represents a meaningful step towards a practical out-of-the-box machine learning model for molecular tasks. The simultaneous generalizability and predictive power contained in this class of models can have large implications on many high-impact problems in structural biology.

We benchmark our GERT equivariant transformer across various datasets and compare it to standard architectures such as the GNN, 3DCNN, and other equivariant neural networks. We showed that GERT is able to achieve performance comparable, if not better than state-of-the-art from other architectures on many tasks. Even on tasks where GERT is not the strongest model, it is still competitive; this consistency across the space of macromolecular tasks is unique. Additionally, we performed ablation studies to understand what the contributors are to the model performance. Our results show that having the built-in notion of physical symmetries is a major contributor to the state-of-the-art performance achieved in some tasks, while the attention module is often unnecessary.

In future work, we hope to explore the possibility of conducting transfer learning with our current architecture. Since GERT already generalizes very well to a variety of tasks, it would be interesting to study whether or not GERT can be pre-trained on a data-rich task first, to improve performance on a data-limited task if trained on a data-rich task first, as the latter tend to be quite common in biology.

References

AlQuraishi, M. Proteinnet: a standardized data set for machine learning of protein structure, 2019.

Anderson, B., Hy, T.-S., and Kondor, R. Cormorant: Covariant molecular neural networks, 2019.

Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754.

Cao, H., Wang, J., He, L., Qi, Y., and Zhang, J. Z. DeepDDG: Predicting the stability change of protein point mutations using neural networks. *Journal of Chemical Information and Modeling*, 59(4):1508–1514, February 2019. doi: 10.1021/acs.jcim.8b00697.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

Eismann, S. et al. Hierarchical, rotation-equivariant neural networks to select structural models of protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 89(5): 493–501, Dec 2020. ISSN 1097-0134. doi: 10.1002/prot.26033.

Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. Protein interface prediction using graph convolutional networks. pp. 6533–6542, 2017.

Fuchs, F. B., Worrall, D. E., Fischer, V., and Welling, M. Se(3)-transformers: 3d roto-translation equivariant attention networks, 2020.

Fuchs, F. B., Wagstaff, E., Dauparas, J., and Posner, I. Iterative se(3)-transformers, 2021.

Geiger, M. et al. Euclidean neural networks: e3nn, 2020.

Gomes, J., Ramsundar, B., Feinberg, E. N., and Pande, V. S. Atomic convolutional networks for predicting protein-ligand binding affinity, 2017.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022. doi: 10.1101/2022.04.10.487779.

Ingraham, J., Garg, V. K., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. 2019.

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

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, pp. 1–11, 2021.

King, J. E. and Koes, D. R. Sidechainnet: An all-atom protein structure dataset for machine learning, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014.

Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. 2019. doi: 10.1101/622803.

S Heck, G., O Pintro, V., R Pereira, R., MB Levin, N., F de Azevedo, W., et al. Supervised machine learning methods applied to predict ligand-binding affinity. *Current medicinal chemistry*, 24(23):2459–2470, 2017.

Savojardo, C., Martelli, P. L., and Casadio, R. Protein–protein interaction methods and protein phase separation. *Annual Review of Biomedical Data Science*, 3(1):89–112, July 2020. doi: 10.1146/annurev-biodatasci-011720-104428.

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P. M. Fast and flexible protein design using deep graph neural networks. *Cell Systems*, 11(4):402–411, 2020.

Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., and Riley, P. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds, 2018.

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, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: A benchmark for molecular machine learning, 2018.

A. GERT Algorithms

Algorithm 1 GERT Encoder

```
Input: Positions and atom types $(P, E) \in \mathbb{R}^{n\times 3} \times \mathbb{R}^{n\times F}$
Output: Latent space representation $X \in \mathbb{R}^{n\times d}$
$n$ is the number of atoms in $P$; $F$ is the number of possible discrete atom types; $d$ is the encoder representation dimension;
$n_H$ is the number of attention heads.
function Encoder($E, P$):
    $X \leftarrow$ Convolution$(E, P) \in \mathbb{R}^{n\times d}$
    for each encoder layer:
        $U \leftarrow$ Concat(Head$_1(X, P), \dots$ Head$_{n_H}(X, P)) \in \mathbb{R}^{n\times d\times n_H}$
        $V \leftarrow U$, summed along the last dimension $\in \mathbb{R}^{n\times d}$
        $X \leftarrow X +$ Convolution$(V, P) \in \mathbb{R}^{n\times d}$
        $T \leftarrow X \in \mathbb{R}^{n\times d}$
        for each dense layer:
            $T \leftarrow$ Convolution$(T, P) \in \mathbb{R}^{n\times d}$
        $X \leftarrow X + T \in \mathbb{R}^{n\times d}$
    return $X \in \mathbb{R}^{n\times d}$
```

Algorithm 2 Task-Specific Head for LBA, PSR, RES, RSR

```
Inputs: Positions and latent space representation $(P, X) \in \mathbb{R}^{n\times 3} \times \mathbb{R}^{n\times d}$
Output: Task-specific response $X \in \mathbb{R}$
$n$ is the number of atoms in $P$; $d$ is the encoder representation dimension; $W_1$ and $W_2$ are weight matrices.
The Norm$(\cdot)$ operation takes an $L_2$ norm across each rotation order and concatenates them for the output.
function Task-Specific Head($X, P$):
    $X \leftarrow$ Convolution$(X, P) \in \mathbb{R}^{n\times d}$
    $X \leftarrow$ Norm(X) $\in \mathbb{R}^{n\times d'}$
    $X \leftarrow X$, summed along the first (atom) dimension $\in \mathbb{R}^{d'}$
    $X \leftarrow W_2 \cdot$ LeakyReLU$(W_1 X) \in \mathbb{R}$
    return $X \in \mathbb{R}$
```

Algorithm 3 Task-Specific Head for LEP, MSP, PIP

```
Inputs: Positions and latent space representation $(P_1, P_2, X_1, X_2) \in \mathbb{R}^{n_1\times 3} \times \mathbb{R}^{n_2\times 3} \times \mathbb{R}^{n_1\times d} \times \mathbb{R}^{n_2\times d}$
Output: Task-specific response $X \in \mathbb{R}$
$n_1$ and $n_2$ are the number of atoms in $P_1$ and $P_2$, respectively; $d$ is the encoder representation dimension; $W_1$ and $W_2$
are weight matrices.
The Norm$(\cdot)$ operation takes an $L_2$ norm across each rotation order and concatenates them for the output.
function Task-Specific Head($X_1, X_2, P_1, P_2$):
    $X \leftarrow$ Concat($X_1$, $X_2$) $\in \mathbb{R}^{n_1+n_2\times d}$
    $P \leftarrow$ Concat($P_1$, $P_2$) $\in \mathbb{R}^{n_1+n_2\times 3}$
    $X \leftarrow$ Convolution$(X, P) \in \mathbb{R}^{n_1+n_2\times d}$
    $X \leftarrow$ Norm(X) $\in \mathbb{R}^{n_1+n_2\times d'}$
    $X \leftarrow X$, summed along the first (atom) dimension $\in \mathbb{R}^{d'}$
    $X \leftarrow W_2 \cdot$ LeakyReLU$(W_1 X) \in \mathbb{R}$
    return $X \in \mathbb{R}$
```

B. ATOM3D Tasks Descriptions

Protein Structure Ranking (PSR). This dataset is comprised of the data from the past 18 years of the CASP competition. The task is a regression task on predicting the Global Distance Test-Total Score (GDT-TS), and structures are split by

competition year.
RNA Structure Ranking (RSR). This dataset is comprised of the data from the first 21 released RNA puzzle challenges. The task is to predict the root mean squared deviation (RMSD) from the ground truth structure, and structures are split by competition year.
Protein Interface Prediction (PIP). The training set is the Database of Interacting Protein Structures (DIPS), and the test set is Docking Benchmark 5 (DB5). The task is to predict whether two amino acids will come into contact given that their respective proteins are bound to each other, and we try to maximize the AUROC over this task. Protein complexes are split by sequence identity at 30%.
Ligand Binding Affinity (LBA). The data is the PDBBind database (see (Townshend et al., 2021) for details), and we predict $-\log(K)$, where K is the binding affinity of the protein-ligand complex in molar units. Protein-ligand complexes are split by protein sequence identity at 30%.
Ligand Efficacy Prediction (LEP). The dataset consists of proteins with both known “active” and “inactive” states along with 527 small molecules with known activating or inactivating function with the proteins. The task is to predict whether or not a given molecule bound to the protein will be activating or not. Complex pairs are split by protein.
Residue Identity (RES). The data consists of atomic environments around selected residues from the PDB, and the task is to predict amino acid identity given the surrounding atomic environment. Residue environments are split by domain-level CATH protein topology class.
Mutation Stability Prediction (MSP). The data consists of pairs of structures, where one element is a protein and the other is the protein with a single point mutation. The task is to predict whether the stability of the complex increases as a result of the mutation. Protein complexes are split by sequence identity at 30%.

C. Aggregate Task Performance

In this section, we define the *aggregate task performance* metric $\tilde{z}_m$ for a model m. Consider a set of T tasks and M models. Let μ_{tm} and σ_{tm} be the mean and standard deviation on the associated metric across replicates for model m on task t. Let μ_t and σ_t be the mean and standard deviation of the set $\{\mu_{tm} : 1 \leq m \leq M\}$; that is, the mean and standard deviation of the mean model performances across a task t. We can compute a task-normalized score:

$$z_{tm} = \frac{\mu_{tm} - \mu_t}{\sigma_t}$$

Then, we take the median of task-normalized scores across tasks to obtain our aggregate task performance metric for a given model:

$$\tilde{z}_m = \text{Median}(\{z_{tm} : 1 \leq m \leq M\})$$

Assuming that the $\tilde{z}_m$ come from a normal distribution and that the μ_{tm} are independent random variables, the standard deviation s_m of $\tilde{z}_m$ can be approximated by:

$$s_m = \sqrt{\frac{\pi}{2}} \left(\frac{1}{T} \sqrt{\sum_t \frac{\sigma_{tm}^2}{\sigma_t^2}} \right)$$

And the standard error se_m:

$$se_m = \frac{s_m}{\sqrt{T}}$$

We use this metric in Figure 2, where $M = 6$ models (every combination of (*lmax, attention*) in $\{2, 3, 4\} \times \{\text{attention}, \text{no attention}\}$) and $T = 7$ tasks (all tasks in Appendix B). The metrics used are R_S for LBA; AUROC for LEP, MSP, and PIP; accuracy for RES; and per-target R_S for PSR and RSR.

D. Non-Equivariant GERT Performance

Table 2. Comparison of GERT and non-equivariant GERT over (Townshend et al., 2021) macromolecular benchmark tasks with standard deviations reported over three replicates, relative to baselines. Metrics are labeled with ↑ / ↓ if higher/lower is better, respectively. Results are reported as mean ± standard deviation over three training runs.

Task	Target	Metric	GERT	Non-Equivariant GERT
PSR	GDT-TS	mean R_S (↑)	0.273 ± 0.008	0.077 ± 0.007
		global R_S (↑)	0.600 ± 0.021	0.234 ± 0.011
RSR	RMSD	mean R_S (↑)	0.236 ± 0.053	0.207 ± 0.008
		global R_S (↑)	0.318 ± 0.228	0.439 ± 0.037
PIP	contacts	AUROC (↑)	0.891 ± 0.005	0.751 ± 0.002
RES	res. type	accuracy (↑)	0.168 ± 0.005	0.056 ± 0.010
MSP	stab. incr.	AUROC (↑)	0.642 ± 0.029	0.640 ± 0.005
LBA	pK	RMSE (↓)	1.453 ± 0.024	1.593 ± 0.043
(30%)		global R_P (↑)	0.549 ± 0.010	0.434 ± 0.025
		global R_S (↑)	0.558 ± 0.008	0.413 ± 0.024
LEP	act./deact.	AUROC (↑)	0.617 ± 0.054	0.718 ± 0.016