Power law attention biases for molecular transformers

Anonymous Author(s)
Affiliation
Address
email

Abstract

Transformers [9] are the go-to architecture for many data modalities. While they have been applied extensively to molecular property prediction, they do not dominate like they do in language and vision [8, 4]. One cause may be the lack of effective structural biases that capture relevant interatomic relationships. Here, we investigate attention biases as a simple and natural way to encode that structure. Motivated by physical power laws, we propose a family of simple attention biases $b_{ij} = p \log ||\mathbf{r}_i - \mathbf{r}_j||$ which weights attention probabilities according to interatomic distances. On the QM9 dataset [7], this approach outperforms positional encodings and graph attention while remaining competitive with more complex Gaussian kernel biases [6]. We also show that good attention biases can compensate for a complete ablation of scaled dot-product attention, suggesting a low-cost path toward interpretable molecular transformers.

1 Background

1.1 Vanilla scaled dot-product attention

Most contemporary transformer flavors rely upon scaled dot-product attention, which computes the attention probabilities as:

$$A_{ij} = \text{softmax}_j \left[\frac{\mathbf{q}_i \mathbf{k}_j^T}{\sqrt{d_k}} \right] \tag{1}$$

Here, $\mathbf{q}_i = \mathbf{x}_i \mathbf{W}_Q$ is the ith token's query vector, $\mathbf{k}_j = \mathbf{x}_j \mathbf{W}_K$ is the jth token's key vector, and d_k is the query/key dimension. Implicitly, the attention mechanism assumes all information about both the tokens themselves and their structural relationships to each other are contained in their embeddings $\mathbf{x}_i$. Most language and vision transformers provide that structural information by adding positional encodings, conditioned on token position, to the embeddings [9, 2].

Positional encodings have also proved effective for molecular transformers [11]. For example, the random walk positional encoding proposed by Dwivedi et al. [3], which outperforms other encodings on many molecular property prediction tasks, samples random walks to capture information about an atom's bond neighborhood.

As structural biases, positional encodings are well-entrenched in the transformer literature, and rightfully so.

1.2 Attention biasing

Additive positional encodings have a few disadvantages. For one, they are associated with individual tokens rather than global structures. For language or vision, this is not an issue, as each token can be

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

assigned an absolute, discrete position within an ordered sequence or grid. However, molecules are most naturally modeled as graphs or Euclidean point clouds, and neither of these modalities admit absolute and discrete positions that can be associated with an encoding. There is also the more general concern of polysemanticity; in principle, superposing structural information upon token embeddings may muddle representations and hinder interpretability.

An attractive alternative to positional encodings is the attention bias. Attention biases alter pre-softmax attention logits by adding some values b_{ij}, which are computed separately from the scaled dot-product:

$$A_{ij} = \text{softmax}_j \left[\frac{\mathbf{q}_i \mathbf{k}_j^T}{\sqrt{d_k}} + b_{ij} \right] \tag{2}$$

An advantage of attention biases is that they are implicitly pairwise and can be naturally constructed given an inter-token distance metric. In the vision domain, Liu et al. [5] construct learned attention biases from the grid displacement between two image patches. In the molecular domain, Luo et al. [6] learn pair-type-aware Gaussian kernel functions that compute attention biases from interatomic distances, achieving strong results on several property prediction tasks.

1.3 Other structural biases

Attention masking is another common structural bias used for molecular property prediction. Attention masks prevent information exchange between token pairs by setting the attention logit to $-\infty$. For most molecular applications, masking is used to block interactions between non-bonded pairs:

$$A_{ij} = \text{softmax}_j \left[\frac{\mathbf{q}_i \mathbf{k}_j^T}{\sqrt{d_k}} + M_{ij} \right] \quad \text{where} \quad M_{ij} = \begin{cases} 0 & (i,j) \in \mathcal{E} \\ -\infty & \text{otherwise} \end{cases} \tag{3}$$

Attention masking is closely connected to graph machine learning, specifically graph attention kernels like graph attention networks [10, 1].

2 Power law attention biases

Motivated by various power laws from physics such as Coulomb's force, we propose the following attention bias:

$$b_{ij} = p \log ||\mathbf{r}_i - \mathbf{r}_j|| \tag{4}$$

This bias term, after the softmax, weights attention probabilities according to a power law of the interatomic distance:

$$\begin{aligned} A_{ij} &= \text{softmax}_j \left[\frac{\mathbf{q}_i \mathbf{k}_j^T}{\sqrt{d_k}} + p \log ||\mathbf{r}_i - \mathbf{r}_j|| \right] \\ &\propto \exp \left[\frac{\mathbf{q}_i \mathbf{k}_j^T}{\sqrt{d_k}} + p \log ||\mathbf{r}_i - \mathbf{r}_j|| \right] \\ &= ||\mathbf{r}_i - \mathbf{r}_j||^p \exp \left[\frac{\mathbf{q}_i \mathbf{k}_j^T}{\sqrt{d_k}} \right] \end{aligned} \tag{5}$$

Here, p, which represents the power law exponent, is a parameter that can be held fixed, learned per layer, or learned per attention head for greater expressivity. The attention probability is now modulated by two properties of the token pair, the query-key compatibility and the power of the interatomic distance. This is similar to Coulomb's force, whose magnitude is modulated by both the electric charges of the particles as well as the distance between them.

Note that issues with infinities appear when $i = j$ and $||\mathbf{r}_i - \mathbf{r}_j|| = 0$. We remedy this by simply masking out the diagonal elements of the bias before the softmax. We also tried adding a constant ϵ to the bias as well as learning separate self-interaction terms b_{ii}. However, we found no approach had any significant advantage over the others.

Table 1: QM9 ensemble MSE losses

Inductive bias	# Params	HOMO (eV)	LUMO (eV)	U (eV)
Vanilla transformer [9]	0	0.320 ± 0.002	0.694 ± 0.002	18.616 ± 1.371
Masked attention	0	0.099 ± 0.001	0.118 ± 0.002	25.591 ± 2.398
Random walk PE [3]	2176	0.115 ± 0.002	0.141 ± 0.002	32.347 ± 1.254
Gaussian kernel bias [6]	1456×8	$\mathbf{0.059 \pm 0.001}$	$\mathbf{0.068 \pm 0.002}$	21.825 ± 3.809
$b_{ij} = -2 \log \|\|\mathbf{r}_i - \mathbf{r}_j\|\|$	0	0.079 ± 0.001	0.094 ± 0.002	26.845 ± 1.276
$b_{ij} = p \log \|\|\mathbf{r}_i - \mathbf{r}_j\|\|$	8×8	0.081 ± 0.002	0.095 ± 0.002	22.712 ± 2.531
$b_{ij} = p_{<0} \log \|\|\mathbf{r}_i - \mathbf{r}_j\|\|$	8×8	0.076 ± 0.002	0.088 ± 0.001	$\mathbf{18.520 \pm 0.694}$

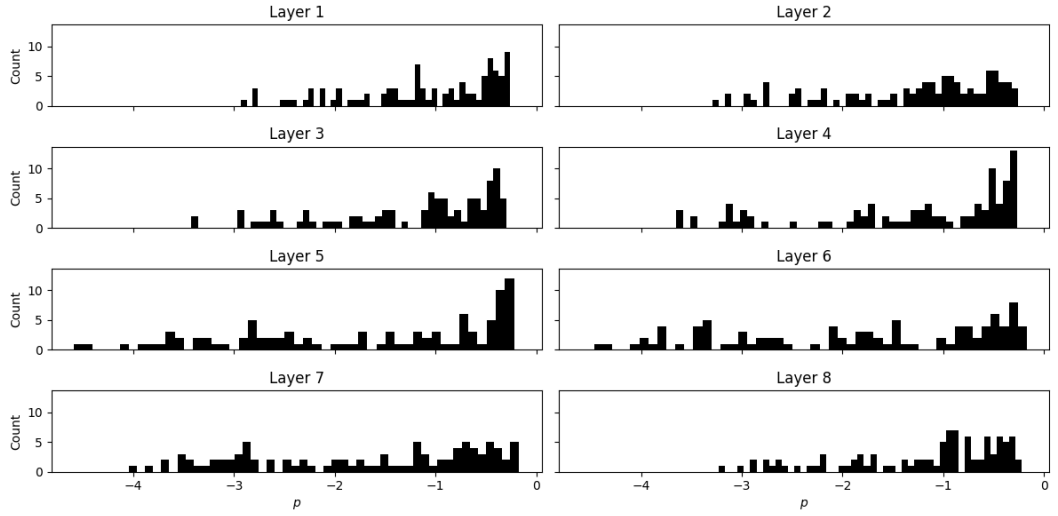

Figure 1: 48-bin histograms of learn constrained exponents $p_{<0}$ by layer, accumulated from all attention heads and 12 independent training runs.

3 Results

3.1 Predicting quantum chemical properties

To compare the structural biases described above, we benchmark on the HOMO, LUMO, and U labels from the QM9 quantum chemistry dataset [7]. The data was prepared by normalizing labels and scaffold splitting molecules using Murcko scaffolds into 8/1/1 train, validation, and test sets. All models are transformers with 128-dimension embeddings, 8 layers, and 8 attention heads per layer, differing only in the structural bias used. Excluding structural bias modules, the models all contain 1,583,491 trainable parameters. Training is also the same—128 epochs with learning rate warmup and cosine decay phases that ensure convergence. Table 3.1 reports test mean squared error losses, as well as the number of parameters added by each structural bias.

In our experiments, we find that attention biases outperform other types of structural biases. We find the best power law bias is $b_{ij} = p_{<0} \log ||\mathbf{r}_i - \mathbf{r}_j||$, in which the learned, per-head exponents are constrained < 0. It is more accurate than all tested models at predicting the internal energy U, and is second only to the more expensive Gaussian kernel biases when predicting the HOMO and LUMO energies. This result is intuitive, as the underlying physics should be governed by inverse power laws.

Figure 1 shows the distribution of exponents $p_{<0}$ learned by models trained with the inverse power law attention bias $b_{ij} = p_{<0} \log ||\mathbf{r}_i - \mathbf{r}_j||$. The exponents do not seem to adhere to any specific power laws, as we would have liked to see, and instead distribute themselves randomly. This is expected, as neural networks often extract their conclusions in highly abstract and difficult-to-interpret ways.

Table 2: Ensemble mean percent change in MSE loss after ablating scaled dot-product attention

$\mathbf{b_{ij}}$	**HOMO**	**LUMO**	**U**
$-2\log\|\|\mathbf{r}_i - \mathbf{r}_j\|\|$	$+11.80\%$	$+15.63\%$	$+38.06\%$
$p_h \log\|\|\mathbf{r}_i - \mathbf{r}_j\|\|$	$+7.93\%$	$+4.23\%$	$+0.93\%$
$-\|p_h\| \log\|\|\mathbf{r}_i - \mathbf{r}_j\|\|$	$+9.10\%$	$+12.09\%$	-11.92%
Gaussian kernel	$+3.25\%$	$+3.93\%$	-9.71%

3.2 Ablating scaled dot-product attention

One interesting use of attention biases is as substitutes for scaled dot-product logits. Namely, instead of computing Equation (2), we could compute attention patterns as:

$$A_{ij} = \text{softmax}_j\big[b_{ij}\big] \tag{6}$$

This "fixed" attention completely decouples structure from token representations, unlike the "dynamic" attention in Equation (2), which allows embeddings to influence structural information flow via the scaled dot-product. It has the immediate benefit of being cheaper to evaluate, scaling as $\mathcal{O}(N^2_{\text{tokens}})$ rather than $\mathcal{O}(N^2_{\text{Tokens}} d_k)$. There are also modest memory benefits from the removal of the query and key weight matrices: the "fixed" attention models we tested were smaller than their "dynamic" attention counterparts by around 12.5%.

Table 2 shows the percent change in loss when using "fixed" attention as opposed to "dynamic" attention. In general, the loss increases slightly, but in a few cases actually decreases, specifically for the internal energy U. This is probably because the internal energy is less dependent on molecule structure, and thus does not rely as heavily on information carried by "dynamic" attention.

We hypothesize that "fixed" attention layers may have use cases situated within large models, reducing compute and memory footprints while retaining or even enhancing accuracy.

4 Conclusions

Here, we proposed a simple attention bias motivated by physics, showed its effectiveness at quantum chemical property prediction compared to baselines, and examined trained model weights. We also tested the hypothesis that attention biases can act as substitutes for scaled dot-product attention logits, and demonstrated the feasibility of that idea. In addition to empirical evidence, we conceptually argued that attention biases naturally model pairwise structure and may leave more opportunities for interpretability.

Our study is limited in that we only experimented with one model size, one dataset, and compared against a small selection of baseline models. Future experimentation should explore scaling and task transfer to clarify where the true advantages lie. We also hypothesized that attention biases promote disentangled representations, and further interpretability analysis could test this.

All the code written for this work is available on GitHub at `https://anonymous.4open.science/r/molecular_attention_bias/`.

References

[1] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks?, 2022.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[3] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations, 2022.

[4] Xiang Fu, Brandon M. Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction, 2025.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[6] Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. One transformer can understand both 2d and 3d molecular data, 2023.

[7] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

[8] Afnan Sultan, Jochen Sieg, Miriam Mathea, and Andrea Volkamer. Transformers for molecular property prediction: Lessons learned from the past five years. *Journal of Chemical Information and Modeling*, 64(16):6259–6280, 2024.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[10] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

[11] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation?, 2021.