Power law attention biases for molecular transformers

Jay Shen
Department of Physics
University of Chicago
Chicago, IL 60637
jshe@uchicago.edu

Oliver Tang
Pritzker School of Molecular Engineering
University of Chicago
Chicago, IL 60637
yifengt@uchicago.edu

Andrew Ferguson
Pritzker School of Molecular Engineering
University of Chicago
Chicago, IL 60637
andrewferguson@uchicago.edu

Abstract

Transformers [1] are the go-to architecture for most data modalities due to their scalability. While they have been applied extensively to molecular property prediction, they do not dominate the field as they do elsewhere [2, 3]. One cause may be the lack of structural biases that effectively capture the relationships between atoms. Here, we investigate attention biases as a simple and natural way to encode structure. Motivated by physical power laws, we propose a family of low-complexity attention biases $b_{ij} = p \log ||\mathbf{r}_i - \mathbf{r}_j||$ which weigh attention probabilities according to interatomic distances. On the QM9 [4] and SPICE [5] datasets, 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, this formulation assumes that all information about both the tokens themselves and their structural relationships are contained in their embeddings $\mathbf{x}_i$. Most language and vision transformers incorporate that structural information by adding positional encodings conditioned on token position to the embeddings [1, 7].

Positional encodings have been used in molecular transformers [8] with modest success. For example, the random walk positional encoding of Dwivedi et al. [9], which tends to outperform others, samples random walks to capture information about an atom's bond neighborhood.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

1.2 Attention biases

Conceptually, positional encodings have a few disadvantages. For one, by their additive nature they are forced to associate with individual tokens rather than inter-token structures. In language or vision, this is not an issue, as tokens can be labeled with discrete, absolute positions. Molecules, however, are most naturally modeled as graphs or Euclidean point clouds, and neither modality admits such convenient indexes.

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

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

Attention biases naturally model pairwise relationships, and can easily be constructed given a measure of distance. For example, Liu et al. [10] construct attention biases for vision transformers based on the grid displacement between image patches. In the domain of molecular modeling, Luo et al. [6] learn Gaussian kernel functions that compute attention biases from interatomic distances. This approach achieves good performance on several molecular property prediction tasks.

1.3 Other structural biases

Attention masking is another common structural bias which modulates information exchange between token pairs by setting the attention logit to $-\infty$. When applied to molecular models, it is often 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 related to graph machine learning, especially graph attention kernels like GAT [11].

2 Power law attention biases

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

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

Because of the softmax operation, this bias term 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. It modulates attention probability separately from the query-key compatibility: this is conceptually similar to many interactions in physics. For example, Coulomb's law modulates the electric force using both particle charge and separation.

One issue with this bias is that singularities appear when $i = j$ and $||\mathbf{r}_i - \mathbf{r}_j|| = 0$. We remedy this by simply masking out diagonal elements. We also experimented with adding a constant ϵ to the bias as well as learning separate self-interaction terms b_{ii}, but found no approach had any significant advantage over the others.

3 Experiments

3.1 Setup

To compare the structural biases described above, we benchmark on the HOMO, LUMO, U, G, and H energy targets from QM9 [4], as well as the U_F energy target from SPICE [5]. Both QM9 and SPICE are quantum chemistry datasets generated using density functional theory. All regression targets were normalized, and samples were split into 8/1/1 train, validation, and test sets. QM9 was split by Murko scaffold while SPICE was split randomly.

We standardized the parameter and compute budgets used across all tested models (Table 1). Models with attention biases only use them in the first four blocks, with the remaining blocks being vanilla transformers.

Table 1: Standardized hyperparameters across all evaluated transformer models. Excluding structural bias modules, all models contained around 1.5M parameters.

Embedding Dimension	128
Number of Layers	8
Attention Heads per Layer	8
Dropout	0.1
Training Epochs	128
Batch Size	64
Learning rate	0.0001
Weight decay	0.00001
Learning Rate Warmup	1 epoch
Learning Rate Schedule	On-plateau decay

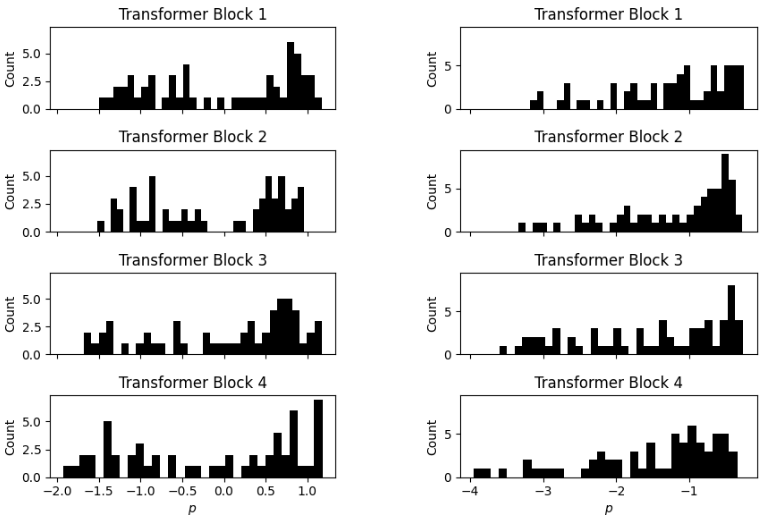
3.2 Results

Table 2: Test mean absolute errors of transformer models with various structural biases when trained to predict energies from QM9 and SPICE. Reported values are the mean over an ensemble, with the standard deviations included, when significant, in parentheses. The number of additional parameters associated with each structural bias is reported, up to multiplicative and additive constants, in terms of the embedding dimension E, number of attention heads H, token dictionary size T, and number of transformer blocks with biased attention N.

		QM9					SPICE
Structural bias	# Params	HOMO	LUMO	U	H	G	U_F
Transformer [1]	0	0.34	0.74	24 (1)	24 (1)	24 (1)	99 (8)
Masked Attention	0	0.12	0.14	29 (4)	29 (4)	30 (4)	-
RWPE [9]	kE	0.14	0.20	42 (7)	42 (7)	42 (8)	-
Attention bias							
Gaussian kernel [6]	kHT^2N	0.08	0.10	21 (1)	21 (1)	21 (1)	-
Power law $p = -1$	0	0.11	0.13	26 (1)	26 (1)	26 (1)	7 (1)
Power law $p \in \mathbb{R}$	HN	0.12	0.14	25 (1)	25 (1)	25 (1)	5 (1)
Power law $p \in \mathbb{R}_-$	HN	0.11	0.12	21 (1)	21 (1)	21 (1)	5 (1)

The results of our experiments are described in 2. We find that attention biases consistently outperform the other classes of structural biases tested. Of the attention biases, Gaussian kernel biases are the most performant in general, but power law biases where $p \in \mathbb{R}^-$ match their accuracy on the U, H, and G targets at much lower computational cost.

It makes sense why the learned constrained power law $p \in \mathbb{R}^-$ is superior to power laws $p = -2$ and $p \in \mathbb{R}$, as it is more expressive than the former and, since physics is governed by inverse power laws, more principled than the latter.



(a) $p \in \mathbb{R}$

(b) $p \in \mathbb{R}^-$

Figure 1: Histogram of learned power law exponents. Exponents are collected from an ensemble of 7 independently trained models.

In Figure 1, we report the distribution of exponents learned by ensembles of transformers with $p \in \mathbb{R}$ and $p \in \mathbb{R}$ power law attention biases. The exponents do not seem to tend toward any specific power laws and instead distribute randomly. This is to be expected, as neural networks are known to be highly uninterpretable.

3.3 Ablating scaled dot-product attention

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

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

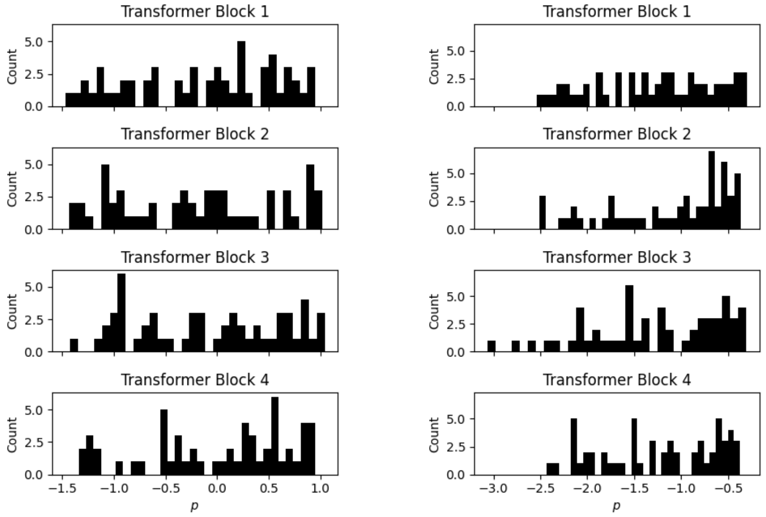
This "fixed" attention completely decouples structure from token representations, unlike "dynamic" attention (Equation (2)) which allows embeddings to influence structural information flow via the dot-product logit. It this way, "fixed" attention is essentially message-passing on a fully connected graph.

"Fixed" attention has the immediate benefit of being cheaper to evaluate. It still scales quadratically as $\mathcal{O}(T^2)$, but does not scale with the embedding dimension $\mathcal{O}(ET^2)$ like "dynamic" attention. Here, T refers to the token dictionary size and E to the embedding dimension. "Fixed" attention also provides modest memory benefits, as the query and key weight matrices are no longer needed.

Table 3: Ensemble mean percent change in MSE loss after ablating scaled dot-product attention

Attention bias	HOMO	LUMO	U	H	G
Gaussian kernel	+7.7%	+7.4%	-5.2%	-5.6%	-6.6%
Power law $p = -1$	+13.8%	+18.9%	+3.3%	+2.2%	+4.3%
Power law $p \in \mathbb{R}$	+7.2%	+5.6%	+2.7%	+0.6%	+1.9%
Power law $p \in \mathbb{R}_-$	+3.8%	+3.0%	+16.0%	+16.2%	+16.9%

Table 3 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 models with Gaussian kernel attention biases predicting the U, H, and G targets. This may be because these energies are less dependent on complex molecule structure, and thus do not rely as heavily on information carried by "dynamic" attention.



(a) $p \in \mathbb{R}$

(b) $p \in \mathbb{R}^-$

Figure 2: Histogram of learned power law exponents when dot-product attention is ablated. Exponents are collected from an ensemble of 7 independently trained models.

Figure 2 shows the distribution of exponents learned by ensembles of $p \in \mathbb{R}$ and $p \in \mathbb{R}$ power law bias models when scaled dot-product attention is ablated. The distribution of exponents is similarly uninterpretable like that of Figure 1.

We hypothesize that "fixed" attention layers may have use cases situated within larger models with "dynamic" attention layers. They can reduce 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 learned model representations. We also tested the hypothesis that attention biases can act as substitutes for scaled dot-product attention logits, and demonstrated that idea's feasibility. In addition to demonstrating empirical evidence in support of attention biases, we argued that they provide a more natural way to encode inter-atom structure in transformer models.

The fundamental limitation of our study is scale: we experimented with one model size and two relatively small datasets. Future experimentation should scale up data, model size, and compute to clarify how our findings hold in different scenarios. For one, we proposed that replacing scaled dot-product attention may be useful when developing larger models—this hypothesis could be tested given greater scale.

The code written for this work is available at `https://github.com/jshe2304/molecular_attention_bias`.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL `https://arxiv.org/abs/1706.03762`.

[2] 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. doi: 10.1021/acs.jcim.4c00747. URL `https://doi.org/10.1021/acs.jcim.4c00747`.

[3] 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. URL `https://arxiv.org/abs/2502.12147`.

[4] 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.

[5] Peter Eastman, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni De Fabritiis, and Thomas E. Markland. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials, 2022. URL `https://arxiv.org/abs/2209.10702`.

[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. URL `https://arxiv.org/abs/2210.01765`.

[7] 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. URL `https://arxiv.org/abs/2010.11929`.

[8] 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. URL `https://arxiv.org/abs/2106.05234`.

[9] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations, 2022. URL `https://arxiv.org/abs/2110.07875`.

[10] 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.

[11] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. URL `https://arxiv.org/abs/1710.10903`.