A physically-motivated attention bias for molecular property prediction

Jay Shen [1] Oliver Tang [1] Andrew Ferguson [1]

[1]University of Chicago
Chicago, IL 60637

Introduction

Machine learning over molecule spaces is a critical task within accelerated workflows for drug discovery, material science, and more. Transformer architectures utilizing self-attention are commonly used for this task due to their flexibility, scalability, and easy of use.

Much effort thus far has been dedicated towards the attention biases and/or positional encodings transformers need for structural understanding. However, most existing so-called structural biases are inscrutable and frankly arbitrary.

Here, we propose a novel structural bias that is interpretable and physically motivated, then evaluate it against other popular structural biases.

Background

Attention mechanisms compute unnormalized pairwise attention scores $\tilde{a}_{ij}$ from token embeddings and then, after optional masking and/or biasing, normalize to obtain attention weights:

$$a_{ij} = \text{Softmax}\Big(\text{Mask}\big(\,\tilde{a}(\vec{e}_i, \vec{e}_j) + b_{ij}\,\big)\Big)$$

This framework encompasses a few popular variants:

- **Masked graph transformers** restrict information flow beyond atomic neighborhoods by using the molecule adjacency matrix as a mask, thus hard-coding bond structure.

$$a_{ij} = \text{Softmax}\Big(\text{Adj}\big(\,\tilde{a}(\vec{e}_i, \vec{e}_j)\,\big)\Big)$$

- **Graph transformers** derive positional encodings, such as centrality embeddings, and attention biases, such as shortest path distance embeddings, from molecule graph structure.

$$a_{ij} = \text{Softmax}\big(\,\tilde{a}(\vec{e}_i + \vec{a}_i, \vec{e}_j + \vec{a}_j) + b_{ij}\,\big)$$

- **Point-cloud transformers** treat molecules as clouds of particles in Euclidean space, and construct biases from atom coordinates:

$$a_{ij} = \text{Softmax}\big(\,\tilde{a}_{ij} + b(\vec{r}_i, \vec{r}_j)\,\big)$$

Normally, rotation- and translation-equivariance is required so that:

$$b(\vec{r}_i, \vec{r}_j) = b(\mathbf{R}_\theta \vec{r}_i + \vec{r}_0, \mathbf{R}_\theta \vec{r}_j + \vec{r}_0)$$

A wealth of literature exists which elaborates upon this design space. Some hybrid models even combine one or more of these approaches and achieve significant performance gains. Here, we assume model improvement is compositionally monotonic and examine various strategies in isolation.

A new attention bias

Notice that the attention bias becomes a scaling coefficient when passed through the softmax function:

$$a_{ij} = \text{Softmax}(\tilde{a}_{ij} + b_{ij}) = \frac{1}{Z} e^{b_{ij}} e^{\tilde{a}_{ij}}$$

Many existing biases thus scale attention weights unphysically with respect to interatomic distances. With this in mind, we propose the following bias:

$$b(\vec{r}_i, \vec{r}_j) = p \log \big|\big|\vec{r}_i - \vec{r}_j\big|\big|$$

where p is learned. The attention weights then become:

$$a_{ij} = \frac{1}{Z} \big|\big|\vec{r}_i - \vec{r}_j\big|\big|^p e^{\tilde{a}_{ij}}$$

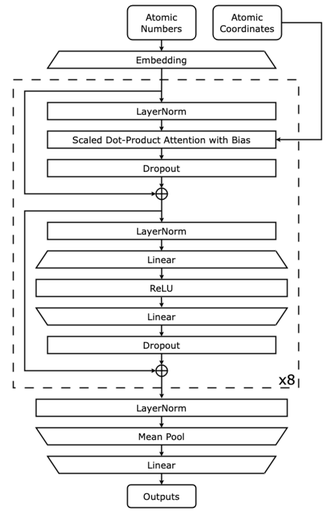
This follows strict power law scaling with respect to the interatomic distance. For example, in the case of $p = -2$, the bias effectively implements Coulomb's inverse-square law.

Experiment setup

We evaluate this bias on QM9 [3], a quantum chemistry dataset of 134k small molecules. Regression targets are normalized and the data is Murcko scaffold splitted.

All models implement a standard transformer architecture with 128-dimension embeddings, 8 attention heads, 8 transformer blocks, 512-dimension MLP hidden space, pre-layer norm, mean pooling, and linear readout. Atoms are minimally featurized using learned embeddings.

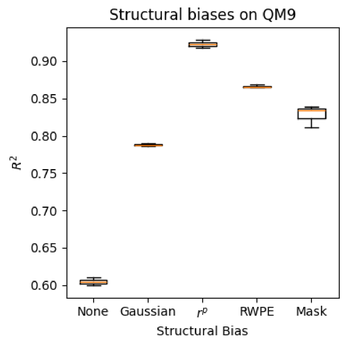
All training runs utilize a batch size of 64, fixed learning rate of 0.001, and weight decay of 0.001.

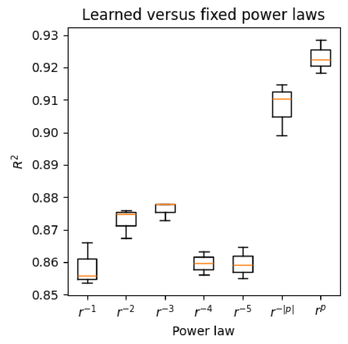


Reference models

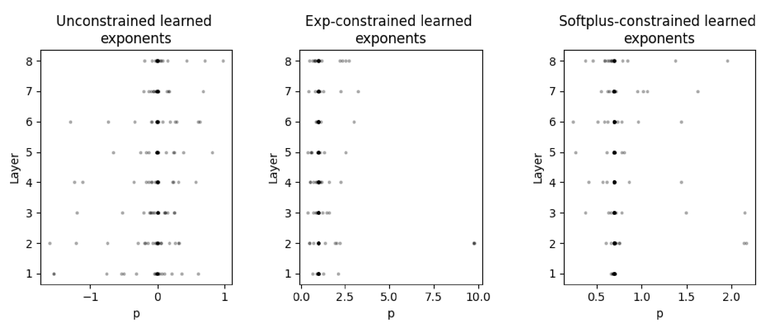
We compare the performance of our bias to that of the Gaussian kernel attention bias [2], random walk positional encoding [1], and masked graph attention. These baselines were chosen to represent standout examples of the three variants described earlier.

Results





As seen above, our bias outperforms other structural biases/encodings tested, including the Gaussian basis function bias. Furthermore, it is clear that learning exponents is better than keeping them fixed.



Analyzing the exponents learned, it appears there is little correspondence to any well-known power laws. Instead, the exponents seem normally distributed around zero. Even if we constrain the exponents to be negative, using an exponential or softplus function, the underlying parameters continue to distribute normally. Since this evidently improves model performance, we hypothesize that a variety of exponents enables multifaceted understanding of atomic configurations.

References

[1] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations, 2021.

[2] Dominic Masters, Josef Dean, Kerstin Klaser, Zhiyi Li, Sam Maddrell-Mander, Adam Sanders, Hatem Helal, Deniz Beker, Ladislav Rampášek, and Dominique Beaini. GPS++: An optimised hybrid mpnn/transformer for molecular property prediction, 2022.

[3] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. 2014.