Unified Molecule Transformer

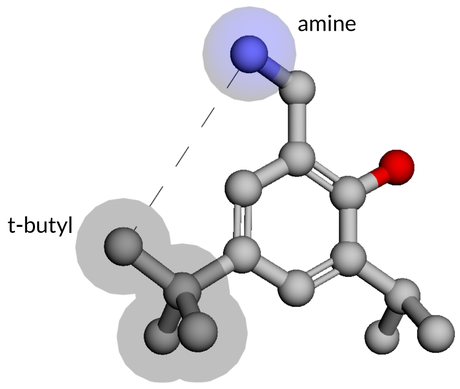
Łukasz Maziarka[1,2], Tomasz Danel[1,2], Sławomir Mucha[2],
Krzysztof Rataj[1], Stanisław Jastrzębski[3,4]
[1]Ardigen [2]Jagiellonian University [3]Molecule.one [4]New York University

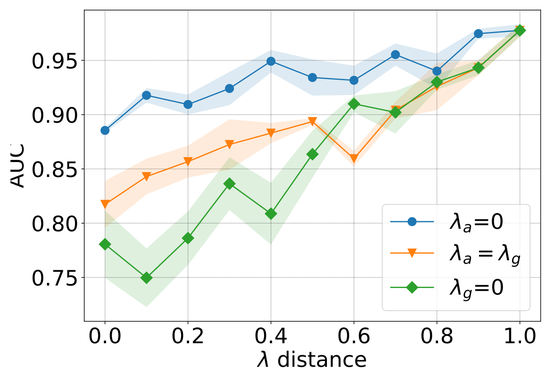
Abstract

Properties of a molecule depend on a variety of relationships between its atoms. On a high level, these relationships might include spatial proximity, the existence of a chemical bond, or simply a co-occurrence of two atoms. However, the commonly used graph-based models use only the chemical bonds to define the neighbourhood. Motivated by this **we propose Molecule Transformer (MT) model**. Our key innovation is augmenting the attention mechanism in Transformer using the inter-atomic distances, and the molecular graph structure. **Experiments on molecular property prediction tasks show that our method outperforms all the other tested models on multiple tasks**. We also show that individual attention heads implement different yet chemically interpretable functions.

Toy Task

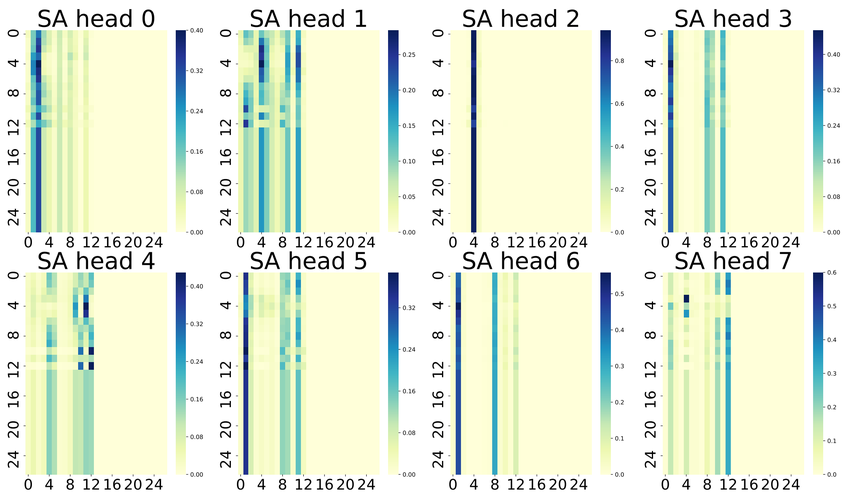
The toy task is to predict whether two substructures (-NH_2 fragment and *tert*-butyl group) co-occur within the given distance. MT can efficiently use the inter-atomic distances to solve the toy task (see left). Additionally, the performance is heavily dependent on λ_d (λ distance), which motivates tuning λ parameters in the main experiments (see right).

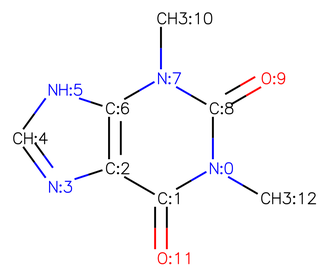




Interpretability

The self-attention outputs from each head are noticeably different and seem to be interpretable. Below we study such patterns on the example of the BBBP test dataset (see Experiments). For instance, we see that head 4 puts a large weight between the nitrogen atoms in the imidazole ring and the oxygen atoms of the oxo groups.





We noticed more patterns like this in the dataset. E.g.

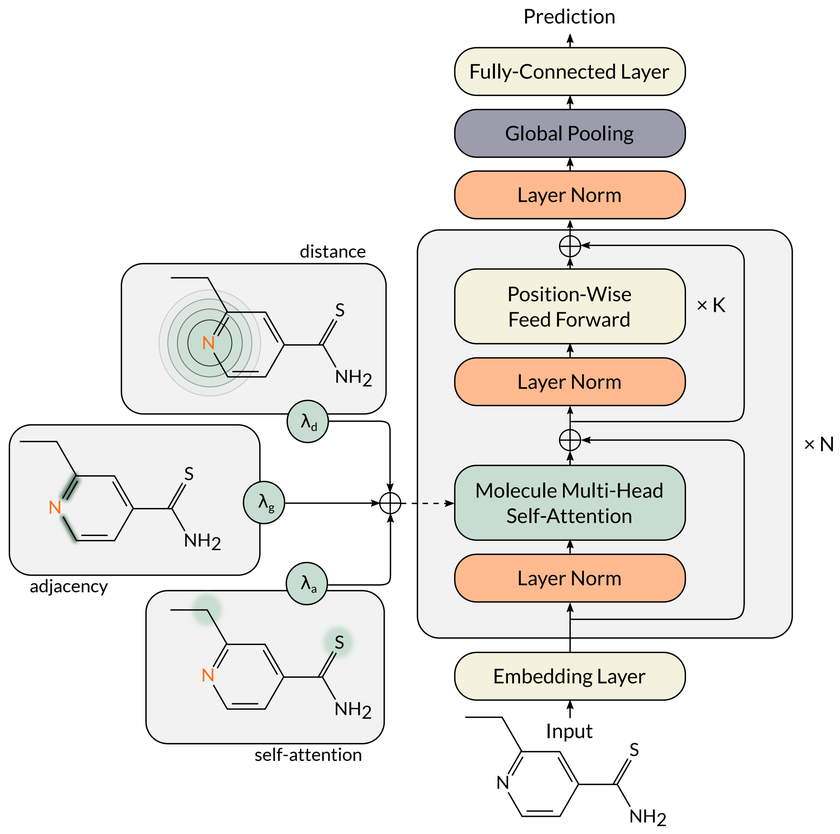
1. Head that seems to focus on the non-carbon atoms connected by only one bond;
2. Head that seems to focus on the carbon atoms within ring structures;
3. Head that seems to focus on the positions of electronegative atoms, but not nitrogen or oxygen.

For each head h and its associated atomic pattern a_h (e.g. "atom in a carbon ring") let $h(a_h)$ denote the overall attention strength assigned by the head h to the given atomic pattern a_h. We report in the left table $h(a_h)$ for the three selected heads and atomic pattern pairs. In the right table we report how often $h(a_h)$ is highest for all atoms a in the molecule.

	Head_1	Head_2	Head_3
Selected	5.23	1.24	1.27
Random	0.41	0.68	0.60

Head_1	Head_2	Head_3
81.3%	72.0%	68.7%

Model



We introduce Molecule Transformer (MT), a Transformer-based [1] model adapted to processing molecules. The architecture is shown in the figure above.
Molecule Transformer consists of N blocks followed by pooling and a classification layer. Each block is composed of a molecule multi-head self-attention layer, followed by a feed-forward block that includes a residual connection and layer normalization.
The multi-head self-attention is composed of H heads. Each head takes as input hidden state $\mathbf{H}$ and computes first $\mathbf{Q}_i = \mathbf{H}\mathbf{W}_i^Q$, $\mathbf{K}_i = \mathbf{H}\mathbf{W}_i^H$, and $\mathbf{V}_i = \mathbf{H}\mathbf{W}_i^V$. These are used in the attention operation as follows:

$$\mathcal{A}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \left(\lambda_a \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_k}}\right) + \lambda_d g(-\mathbf{D}) + \lambda_g \mathbf{A}\right)\mathbf{V}_i,$$

where the molecule structure is represented by the graph adjacency matrix $\mathbf{A} \in \{0,1\}^{N\times N}$, and the inter-atomic distance matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$. λ_a, λ_d, and λ_g denote scalars weighting the self-attention, distance, and adjacency matrices.

Experiments

Datasets. We run experiments on a wide range of datasets that represent typical tasks encountered in molecular modeling.

- **FreeSolv, ESOL.** Regression tasks. Popular tasks for predicting water solubility in terms of the hydration free energy (FreeSolv) and logS (ESOL).
- **Blood-brain barrier permeability (BBBP).** Binary classification task.
- **$\text{MetStab}_{\text{high}}$, $\text{MetStab}_{\text{low}}$.** Binary classification tasks. The metabolic stability of a compound is a measure of the half-life time of the compound within an organism.
- **hERG, Estrogen Alpha, Estrogen Beta.** Binary classification tasks for predicting binding affinity.

FreeSolv, ESOL and BBBP are popular benchmarks for predicting physical and toxicity-related properties, also included in the MoleculeNet benchmark [2]. The other datasets we consider are aimed to represent biophysical tasks, and were either taken from publications [3] (MetStab datasets) or extracted by us from ChEMBL [4] (hERG and Estrogen datasets).

Experimental setting. We compare Molecule Transformer to the following models: Graph Convolutional Networks (GCN), Random Forest (RF), Support Vector Machine with RBF kernel (SVM), Edge Attention-based Multi-relational Graph Convolutional Networks (EAGCN) [5], Message Passing Neural Networks (MPNN) [6] and Weave [7].
For all the models we tune the hyperparameters by a random search with a fixed budget of 100 trials. The MT model contains many hyperparameters and can benefit from extending or refining the search algorithm. We show this in the MT_{566} model, where the search budget is extended to 500 runs and additionally the lambda parameters have a second stage of search (with other parameters fixed).
We use random split for FreeSolv, ESOL and MetStab datasets. For all the other datasets we use scaffold split. Test performance is based on the best validation epoch. Each training was repeated three times.

Results

	BBBP (AUC)	ESOL (RMSE)	FreeSolv (RMSE)	Estrogen Alpha (AUC)	Estrogen Beta (AUC)	hERG (AUC)	$\text{MetStab}_{\text{low}}$ (AUC)	$\text{MetStab}_{\text{high}}$ (AUC)
SVM	0.603 ± 0.000	0.493 ± 0.000	0.391 ± 0.000	0.933 ± 0.000	0.765 ± 0.000	0.810 ± 0.000	0.828 ± 0.000	0.822 ± 0.0
RF	0.551 ± 0.005	0.533 ± 0.003	0.550 ± 0.004	0.928 ± 0.003	0.770 ± 0.004	0.769 ± 0.003	0.796 ± 0.004	0.706 ± 0.008
GC	0.690 ± 0.015	0.334 ± 0.017	0.336 ± 0.043	0.974 ± 0.005	0.726 ± 0.011	**0.917 ± 0.015**	**0.856 ± 0.013**	0.874 ± 0.014
Weave	0.703 ± 0.012	0.389 ± 0.045	0.403 ± 0.035	0.961 ± 0.005	0.766 ± 0.018	0.765 ± 0.034	0.612 ± 0.009	0.778 ± 0.039
MPNN	0.700 ± 0.019	**0.303 ± 0.012**	0.299 ± 0.038	-	-	-	-	-
EAGCN	0.664 ± 0.007	0.459 ± 0.019	0.410 ± 0.014	0.937 ± 0.031	0.724 ± 0.025	0.826 ± 0.011	0.779 ± 0.034	0.697 ± 0.019
MT	**0.711 ± 0.007**	0.330 ± 0.002	**0.269 ± 0.007**	**0.977 ± 0.003**	**0.790 ± 0.003**	0.906 ± 0.007	0.839 ± 0.009	**0.892 ± 0.005**
MT_{566}	**0.736 ± 0.009**	**0.298 ± 0.005**	**0.259 ± 0.014**	**0.981 ± 0.002**	**0.778 ± 0.006**	**0.92 ± 0.002**	**0.877 ± 0.013**	**0.894 ± 0.008**

References

[1] Ashish Vaswani et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
[2] Zhenqin Wu et al. Moleculenet: a benchmark for molecular machine learning† †electronic supplementary information (esi) available. see doi: 10.1039/c7sc02664a. In *Chemical science*, 2018.
[3] Sabina Podlewska and Rafał Kafel. Metstabon—online platform for metabolic stability predictions. *International journal of molecular sciences*, 19(4):1040, 2018.
[4] Anna Gaulton et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 09 2011.
[5] Chao Shang et al. Edge Attention-based Multi-Relational Graph Convolutional Networks. *arXiv e-prints*, page arXiv:1802.04944, Feb 2018.
[6] Justin Gilmer et al. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*.
[7] Steven Kearnes et al. Molecular graph convolutions: Moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30, 03 2016.