Atomwise

Synthetic benchmarks to evaluate attribution for structure-based graph convolutional networks in drug discovery

Andreana Rosnik, PhD; Kate A. Stafford, PhD; Brandon Anderson, PhD; Henry van den Bedem, PhD

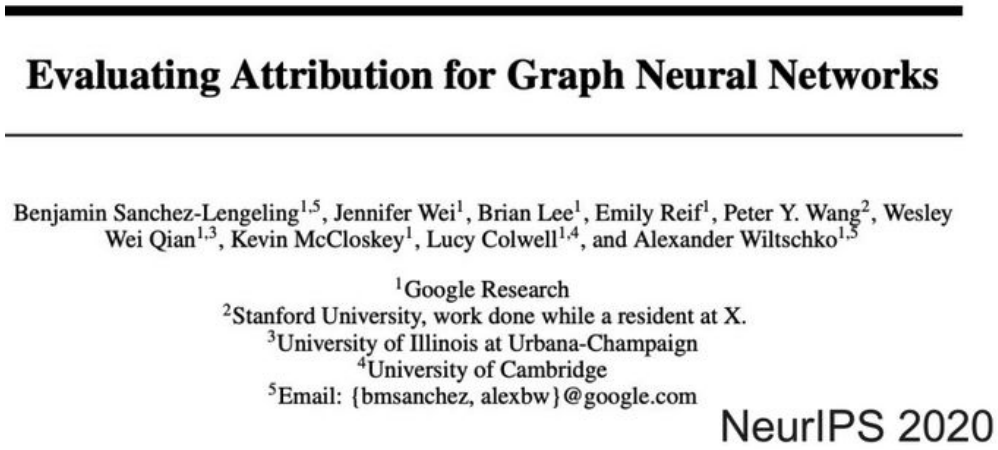
March 23, 2022

Creating interpretable deep learning models

- Interpreting a deep learning (DL) model is challenging
- For drug discovery it's important to understand what our models are learning:
  - Does it make chemical sense?
  - Are the models biased or cheating?
- We use *attention mechanisms* to *attribute* information to nodes (atoms) and edges
- Which atoms contribute to …
  - Pose sensitivity?
  - Activity prediction?
  - pKi regression?

Attention as an attribution technique

- Attention is a process of mapping intermediate DL model outputs back to initial inputs to determine which features were best captured by the model
- "using attention as an attribution technique performed poorly"
  - the degree to which (raw) edge weights map onto *absolute* contribution varies by context



Evaluating Attribution for Graph Neural Networks

Benjamin Sanchez-Lengeling[1,5], Jennifer Wei[1], Brian Lee[1], Emily Reif[1], Peter Y. Wang[2], Wesley Wei Qian[1,3], Kevin McCloskey[1], Lucy Colwell[1,4], and Alexander Wiltschko[1,5]

[1]Google Research
[2]Stanford University, work done while a resident at X.
[3]University of Illinois at Urbana-Champaign
[4]University of Cambridge
[5]Email: {bmsanchez, alexbw}@google.com

NeurIPS 2020

- See also:
  - Interpretation of QSAR Models by Coloring Atoms According to Changes in Predicted Activity: How Robust Is It, *J. Chem. Inf. Model.* (doi: 10.1021/acs.jcim.8b00825)

Attention as an attribution technique

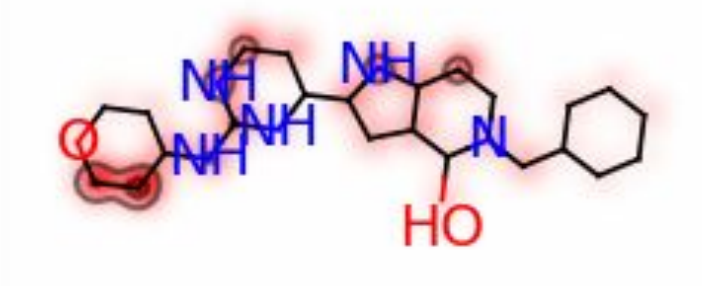
- Hypothesis: Atom-based attention is most informative for information gained from atoms that most closely interact with the receptor
- Ligand atoms share edges with many receptor atoms. Which edges are important?
- Edge entropy = entropy of the edges connecting each node to its neighbors
- For any node/atom, we can calculate the Shannon entropy over the edge attention weights $\{\alpha_{ij}\}_{j \in \mathcal{N}_i}$

$$H(\alpha_i) = -\sum_{j \in \mathcal{N}_i} \alpha_{ij} \log(\alpha_{ij})$$

- Relative entropy: way to measure information gain

$$-\sum_{j \in \mathcal{N}_i} \alpha_{ij} \log(\alpha_{ij} |\mathcal{N}_i|)$$

Good pose, information gained



Bad pose, information unavailable

blue: atom has no contact within 5Å (provide no information gain)
red: one or more contacts have greater weight

Benchmarks move interpretability towards quantification

- Most work in the field is anecdotal
- Our project addresses:
  - Is attention a valid attribution technique for drug discovery problems?
  - Can we develop a synthetic benchmark to evaluate attention's applicability?
  - Can attention models (and benchmarks) help us understand ligand poses?

Matveieva and Polishchuk *J Cheminform* (2021) 13:41
https://doi.org/10.1186/s13321-021-00519-x

Benchmarks for interpretation of QSAR models

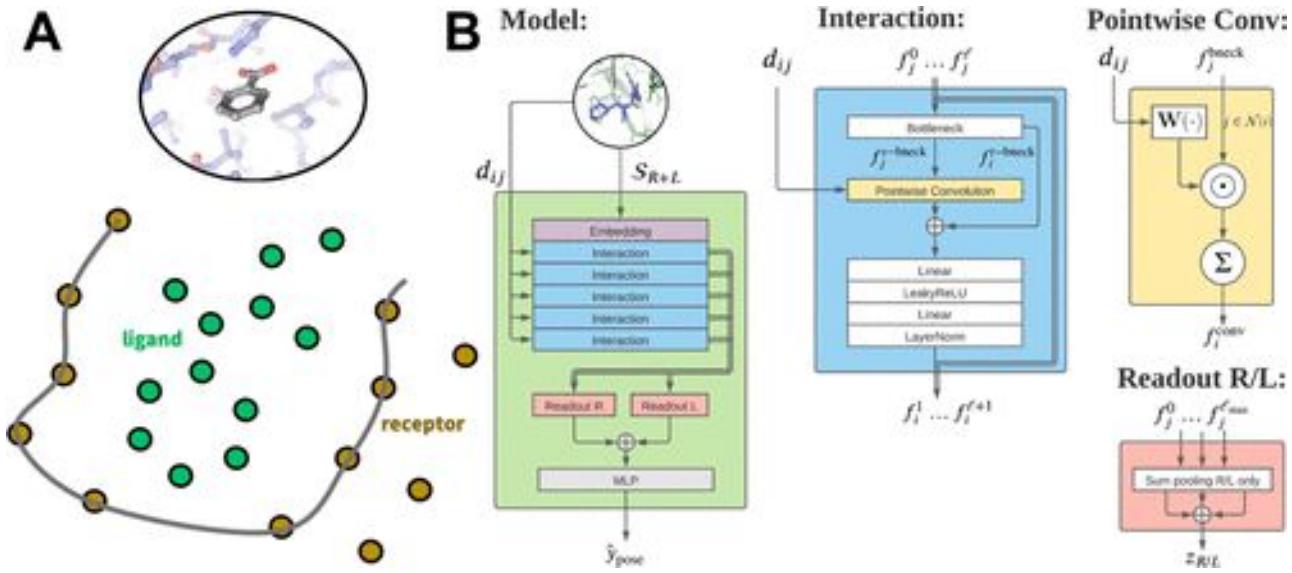
Dataset	Property type	End-point	Train/test set size	Expected atom contribution
N	Regression	Sum(N)	6995/2999	Nitrogen atoms: 1; others: 0
N − O	Regression	Sum(N) − sum(O)	6893/2969	Nitrogen atoms: 1; Oxygen atoms: − 1; others: 0
N + O	Regression	(Sum(N) + sum(O))/2,where sum(N) = sum(O)	7000/3000	Nitrogen and Oxygen atoms: 0.5; others: *0*
Amide_reg	Regression	Sum(NC=O)	7000/3001	Any atom of amide groups: 1; others: 0
Amide_class	Classification	Active: if sum(NC=O) > 0; inactive: if sum(NC=O) = 0	6998/3000	Any atom of amide groups: 1; others: 0
Pharmacophore	Classification	Active: at least one conformer with exactly one pharmacophore match (same two atoms in all conformers); inactive: no pharmacophore matches for all conformers; pharmacophore match: HBD and HBA 9–10 Å apart	7000/3000	Atoms which are HBA or HBD of the pharmacophore: 1; others: 0

Hydrogen bond synthetic benchmark: data preparation

- Model trained to predict number of h-bonds
  - This is a quantity for which we know the ground truth. The question is, will the attention mechanism figure that out?
- Hydrogen bond calculation: count number of {N,O} h-bonds
  - distance cut-off 3.5 Å
  - angle cut-off 45°
- Dataset: PDB data containing binding affinity data, cross-docked, split by 50% sequence similarity (7049 / 3126 targets train/test, ~7.5M poses)
  - cross-docking = docking compounds to different crystal structures of same protein (identified by uniprot)
  - PDB data contains data from: PDBBind, BindingDB, BindingMOAD

Model architecture

- Model:
  - AtomNet® model variant GRAPHite
  - convolutional layers: [rl,rl,rl,l,l,l]
  - attention: [true,true,true,false,false]
- Trained on three different labels, for the sake of comparison:
  - hydrogen bond count prediction
  - pose classification
  - pKi regression

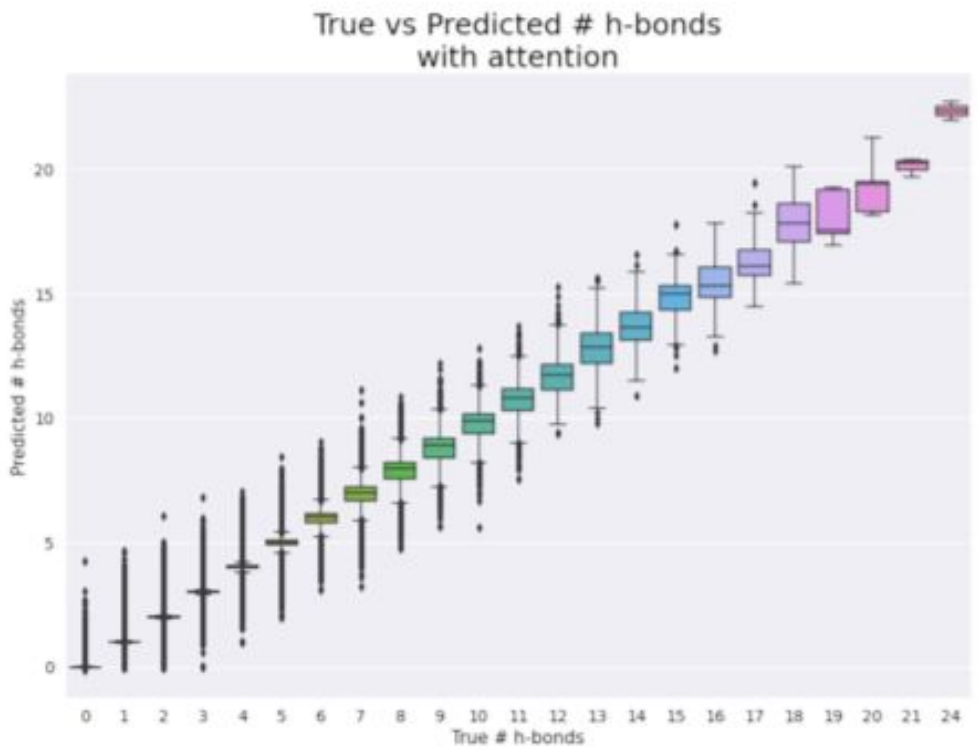


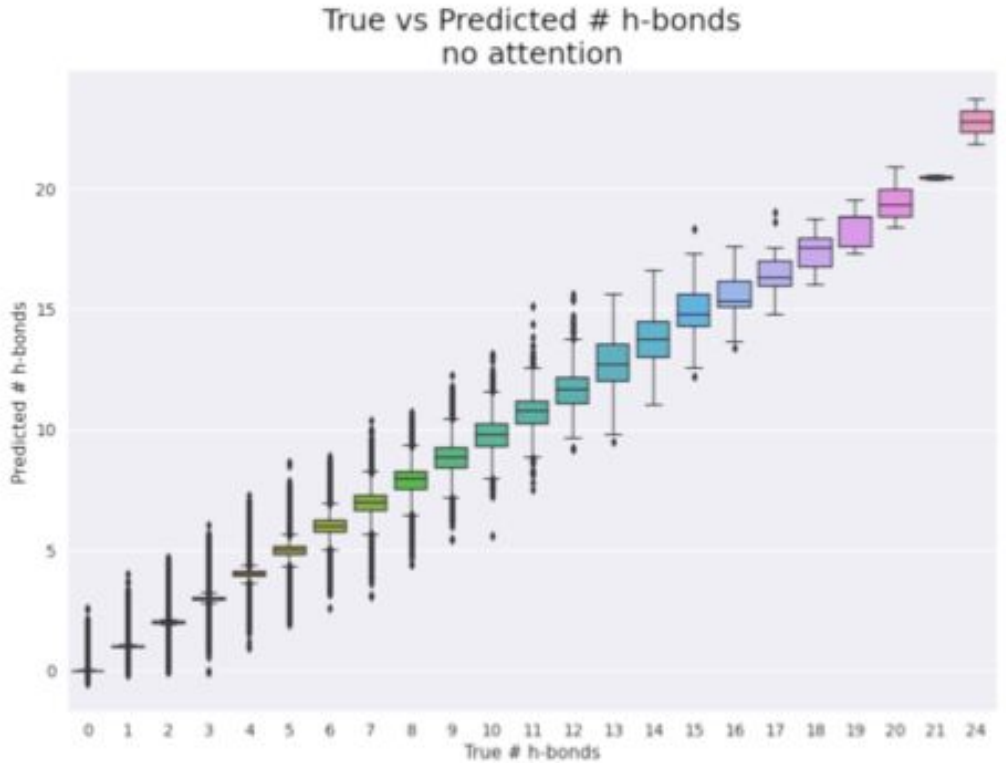
Architecture of AtomNet® model variant GRAPHite, a directional Message Passing Neural Network, Fig 3 from:

AtomNet® PoseRanker: Enriching Ligand Pose Quality for Dynamic Proteins in Virtual High-Throughput Screens, *J. Chem. Inf. Model.* (doi: 10.1021/acs.jcim.1c01250)

Model training

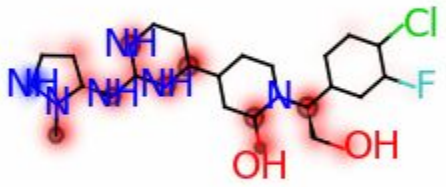
- Hydrogen bond counting is a very easy task
- Attention doesn't alter convergence much, though it slows training 2-3x



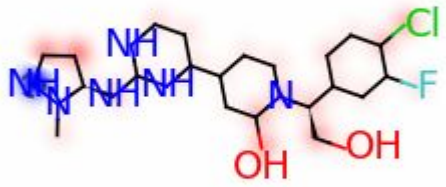


Anecdotes: same architecture, different labels

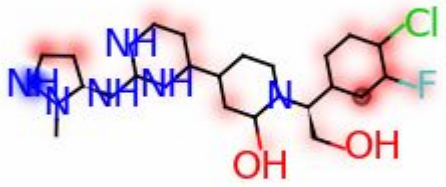
Good poses – **with** attention



hydrogen bond labels

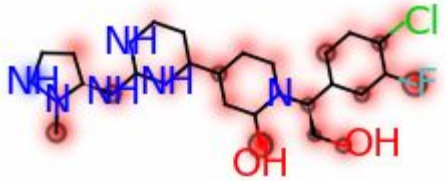


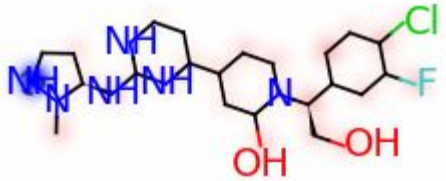
pose labels

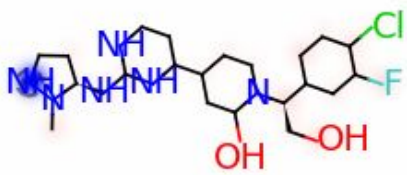


pKi labels

Good poses – **no** attention



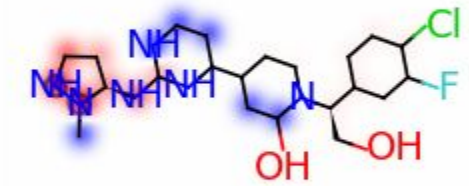




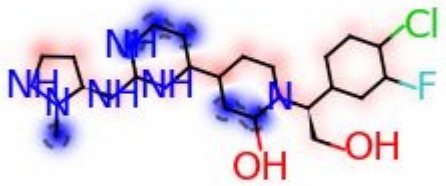
- Models trained on different labels tend to highlight features chemically intuitive to their label
  - hydrogen bond models light up for hydroxyl or amine groups
  - pose models have diffuse areas of entropy gain, since the entirety of the scaffold is relevant
  - pKi models highlight R-groups

Anecdotes: same architecture, different labels

Bad poses – **with** attention



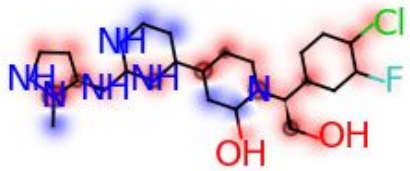
hydrogen bond labels

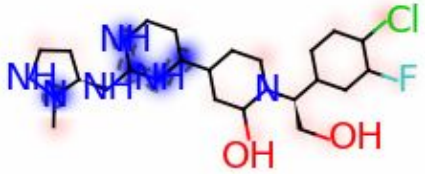


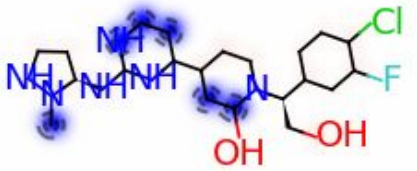
pose labels

pKi labels

Bad poses – **no** attention

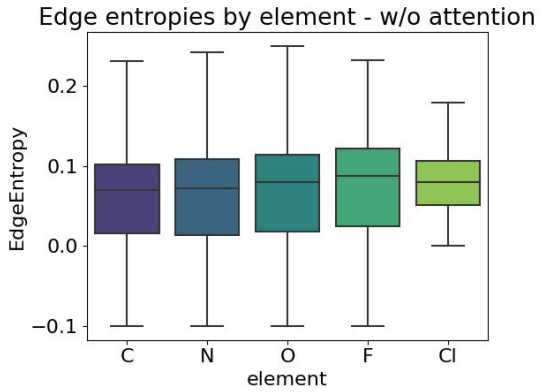
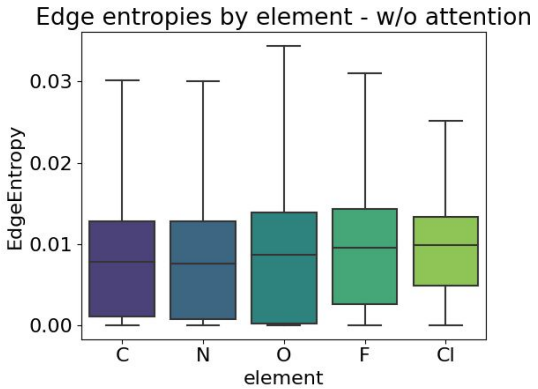
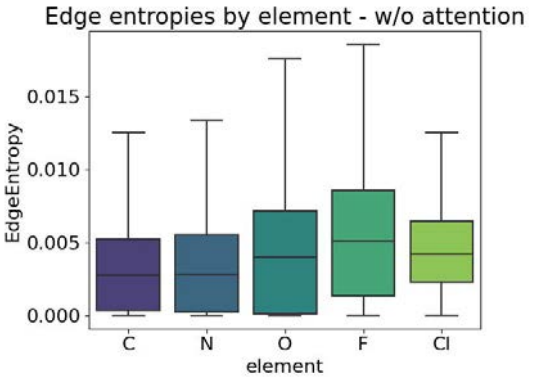
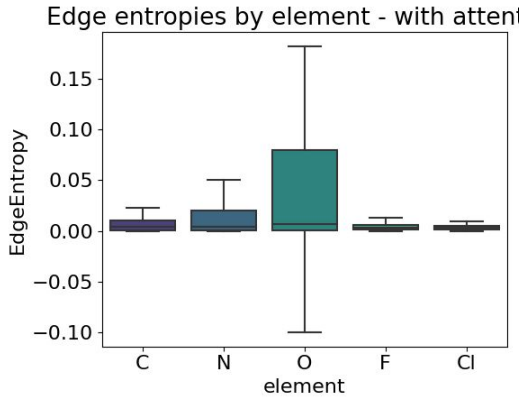
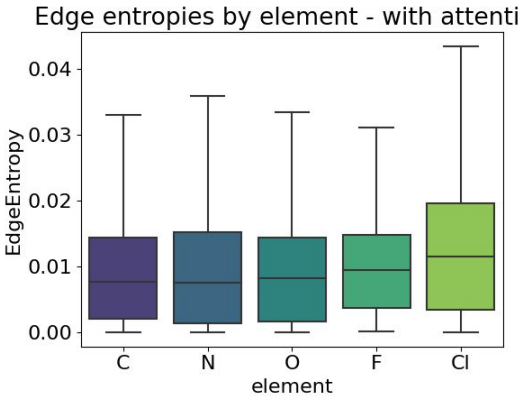
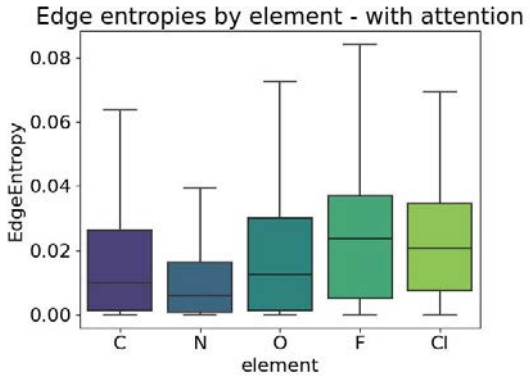






- Models trained on different labels tend to highlight features chemically intuitive to their label
  - hydrogen bond models light up for hydroxyl or amine groups
  - pose models have diffuse areas of entropy gain, since the entirety of the scaffold is relevant
  - pKi models highlight R-groups

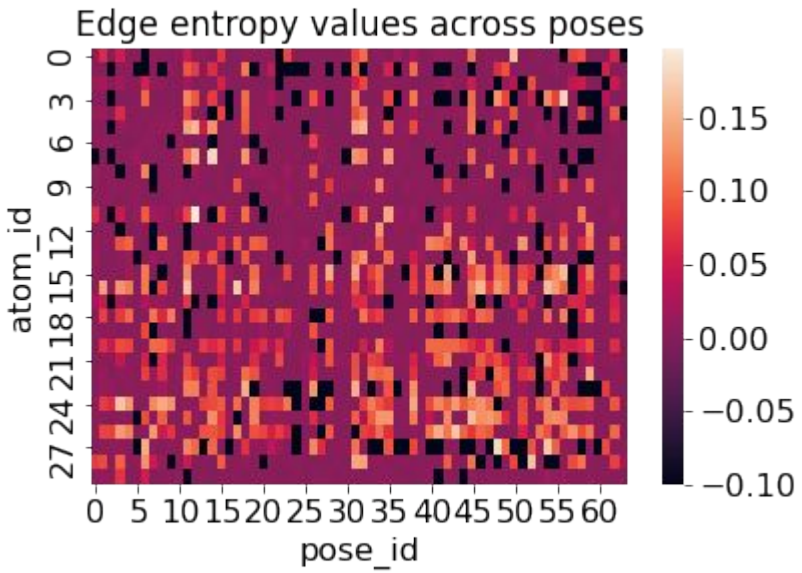
Edge entropy values by element

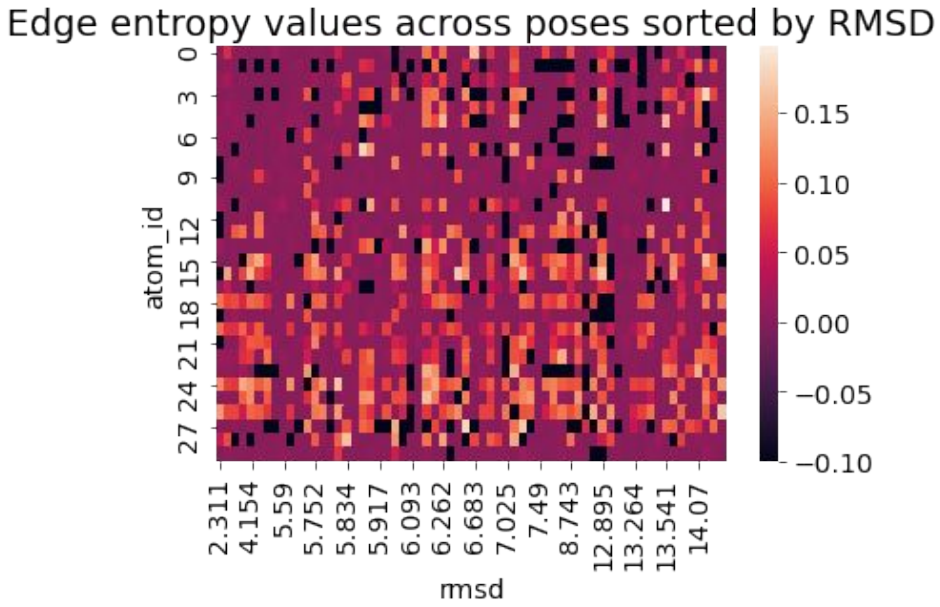
	hbond labels (n=5844)	pose labels (n=5844)	pKi labels (n=5844)
No attention			
With attention			

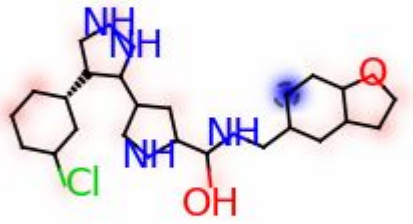
- Adding attention to edges makes a big difference in hbond models: {N,O} indeed get attention
- Attention on the pose labels highlights atoms most prevalent in scaffolds

Anecdotes: what can we learn about pose quality?

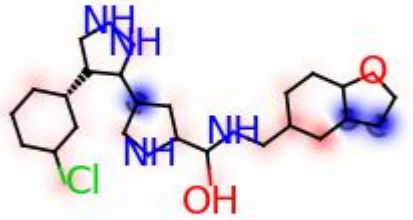
- Hypothesis: regardless of labels, more low entropy regions for worse poses, since there may be more atoms further from receptor atoms
- In some cases, there are more low entropy spots when sorting by *either* pose rank or RMSD
  - tends to occur for ligands with halogens







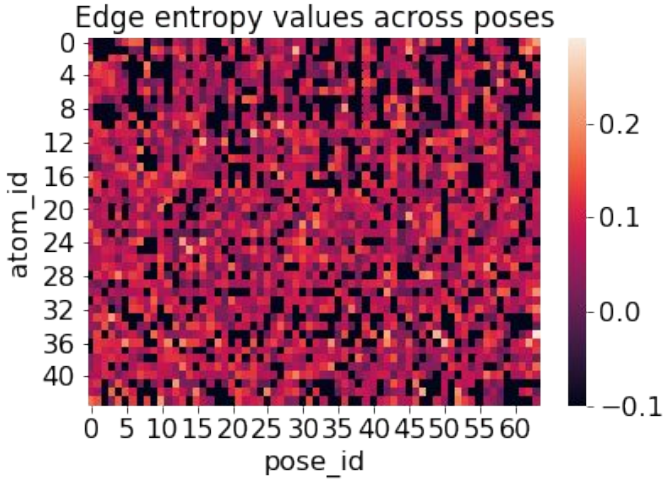
Good pose

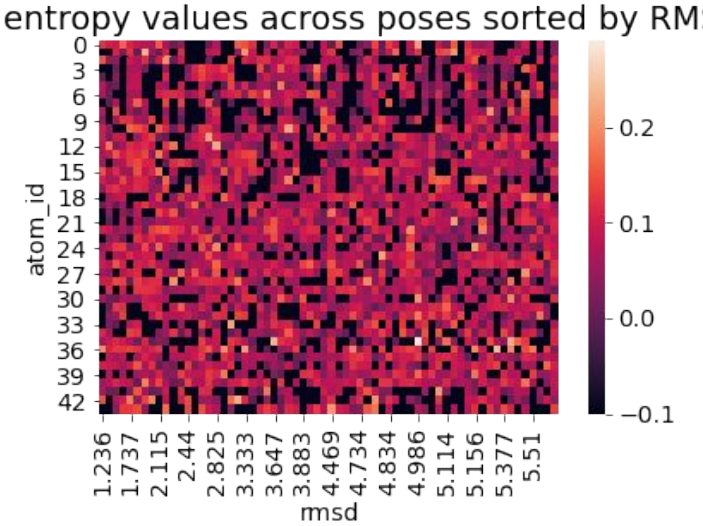


Bad pose

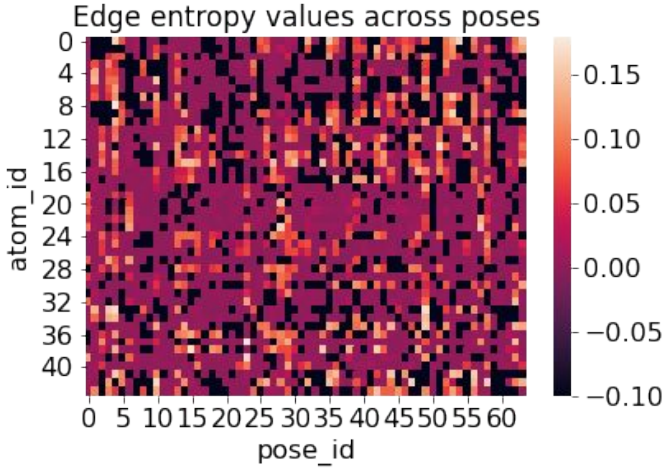
Anecdotes: what can we learn about pose quality?

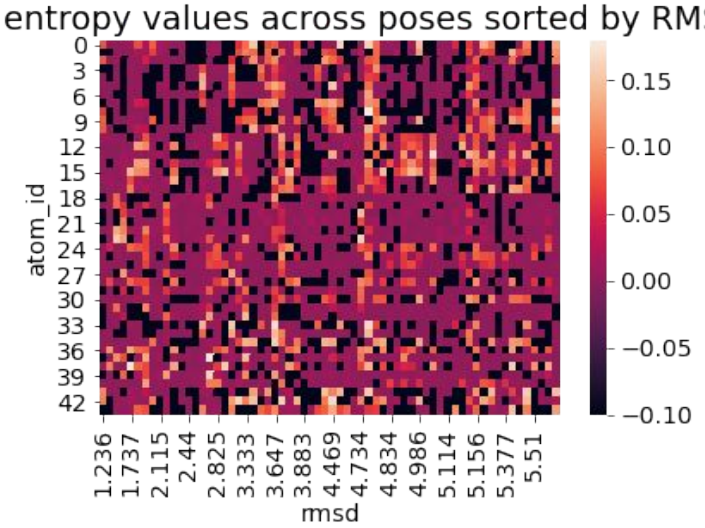
No attention



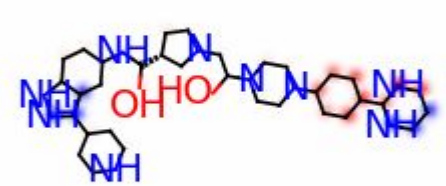


With attention

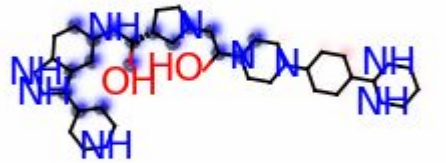




- large ligands have the noisiest plots (higher RMSDs)
- Attention helps reduce noise in edge entropy heat maps



Good pose

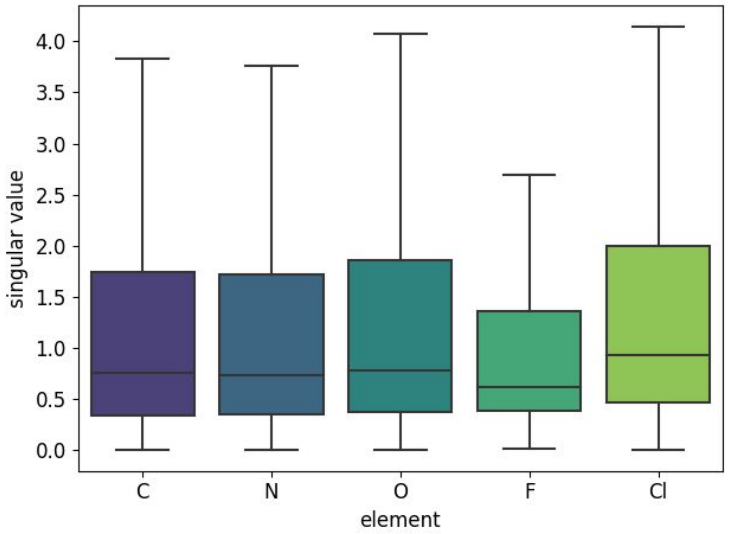


Bad pose

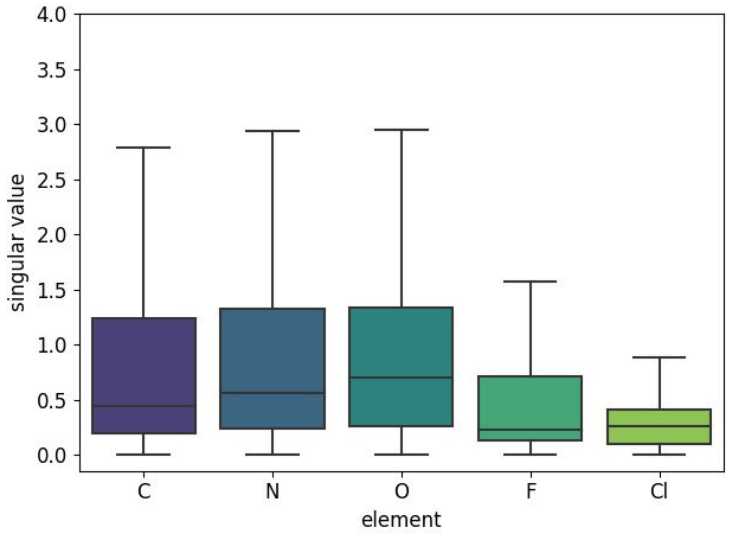
Heatmap signatures: pose quality signals arise

- Singular values of heatmap matrices for hydrogen bond models differ between models with and without attention
- Aggregating by element, we get results similar to simple aggregation of edge entropy values by element
- Perhaps singular value decomposition can uncover more features…

No attention

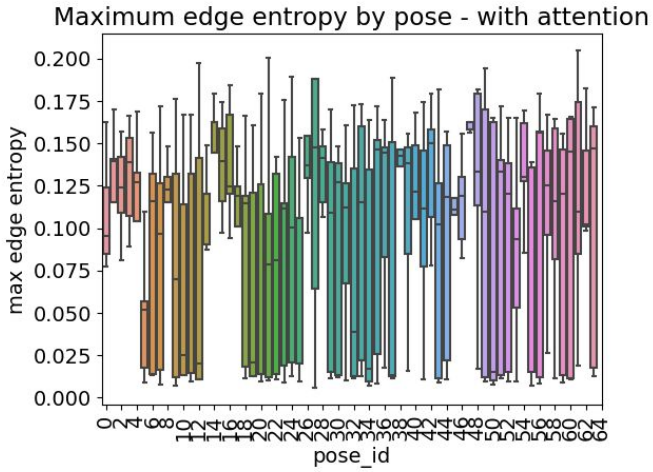


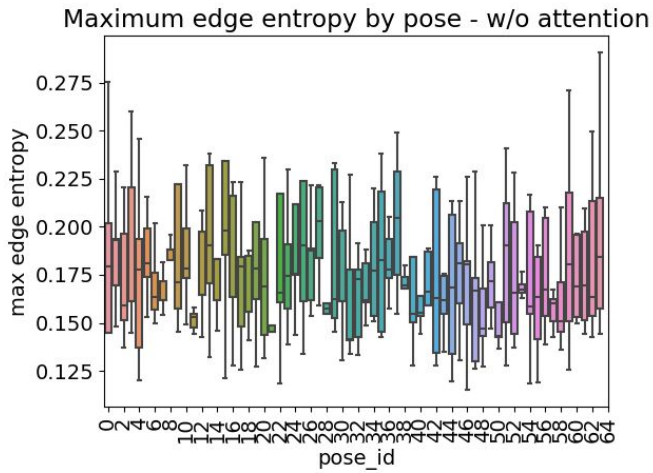
With attention

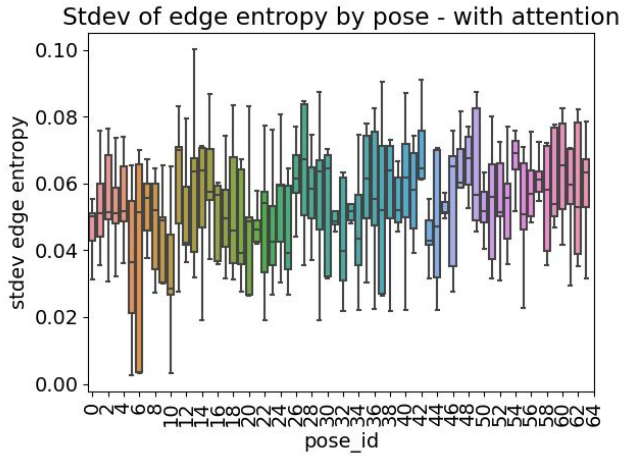


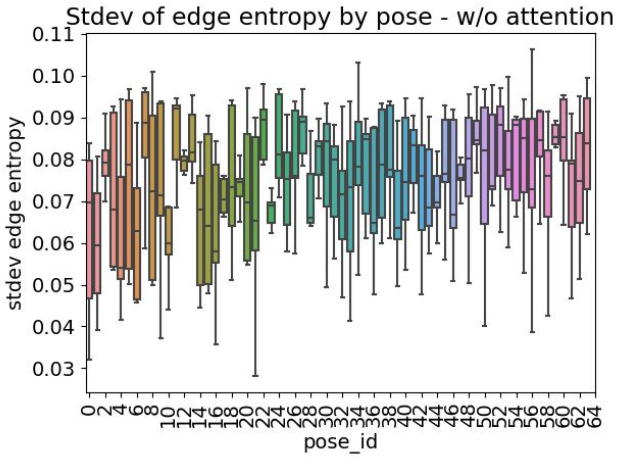
Attention highlights top poses from the rest

- For 5844 randomly selected target-compound pairs,
  - best-ranked poses have the narrowest distribution in maximum entropy values
  - standard deviation in relative edge entropy trends upward as pose rank increases
- The trends for poses 0-5 disappear when attention is absent → attention is (slightly) pose sensitive!









Summary

- Attention works as an attribution mechanism for simple, atom-based properties (like hydrogen bonds) for which we have a ground truth answer
- Hydrogen bond benchmarks are easy to train and appear to learn hydrogen bonds: attention-based models have higher entropy for N and O atoms
- Atom-based attention models perform less well for noisier problems (like pose classification); further investigation is necessary to understand what can be gleaned from attention models in these contexts

Acknowledgements

- My co-authors:
  - Kate
  - Brandon
  - Henry
- The entire cheminformatics team at Atomwise
- You, for listening!

Atomwise

Andreana Rosnik, PhD

Cheminformatics Scientist II
Atomwise Inc.

andreana@atomwise.com

We're hiring!

Learn more about Atomwise:
https://www.atomwise.com/careers/
Read more about our work:
https://blog.atomwise.com/
Atomwise at ACS:
https://info.atomwise.com/acs_spring2022