Prediction of Hydrogen Positions with Machine Learning

University of Stuttgart

Artificial Intelligence Software Academy

Project Coordinator: Christian Pfaendner

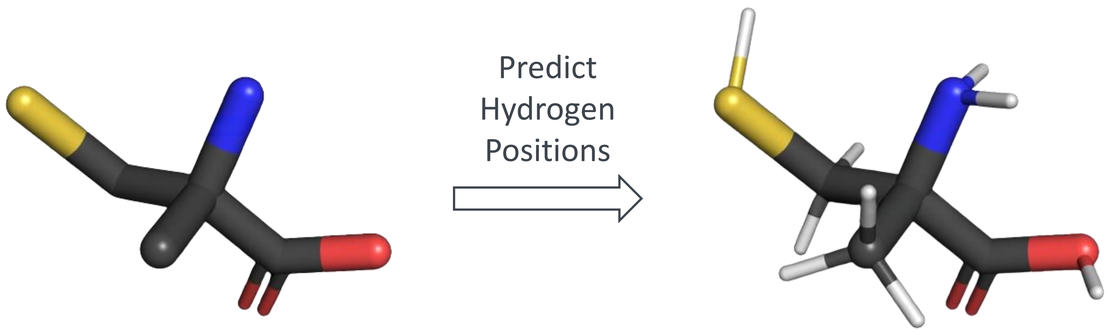
AISA

https://www.aisa.uni-stuttgart.de/

Markus Escher
Oussama Barhoumi
Hasan Evci

Goal

Problem: Experimentally resolved protein structures often lack hydrogen atoms.



Dataset

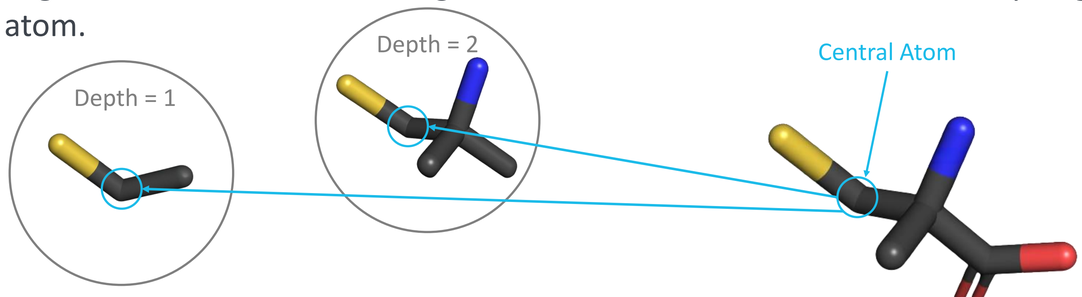
The PDB **Chemical Component Dictionary** contains information about bonds of the atoms and their 3D coordinates.

- 41440 protein molecules
- 286717 Carbon atoms
- 53803 Oxygen atoms

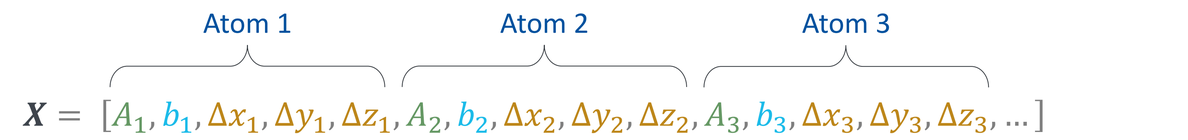
WORLDWIDE PDB PROTEIN DATA BANK

Approach

1. Segment the molecule into neighborhoods centered around each non-hydrogen atom.



2. Write the **neighborhood** as an input feature.



3. Train a model to add the hydrogen atom connected to the central atom.

Training

Data Splitting

Specifications

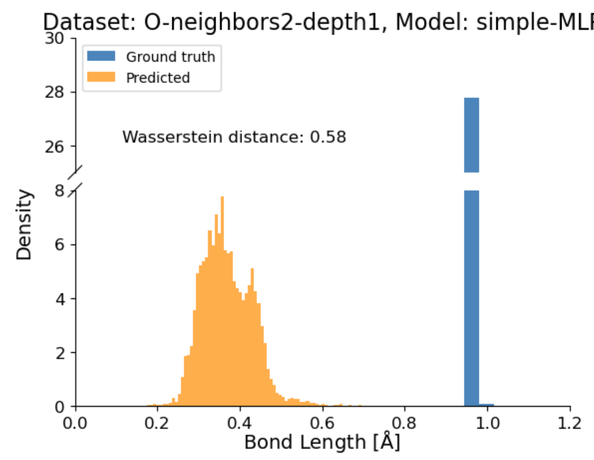
- A different model is trained for each:
  - type of central atom
  - depth of the neighborhood
  - number of bonded non-hydrogen atoms to the central atom
- This project's scope includes predicting the position of **one** hydrogen with:
  - **Carbon** central atom with **3 non-hydrogen** bonded atoms
  - **Oxygen** central atom with **1 non-hydrogen** bonded atom

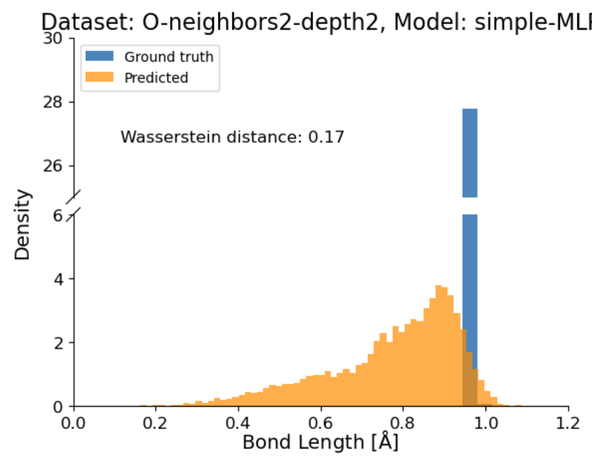
Both with a **neighborhood depth of 1 and 2**

Models used

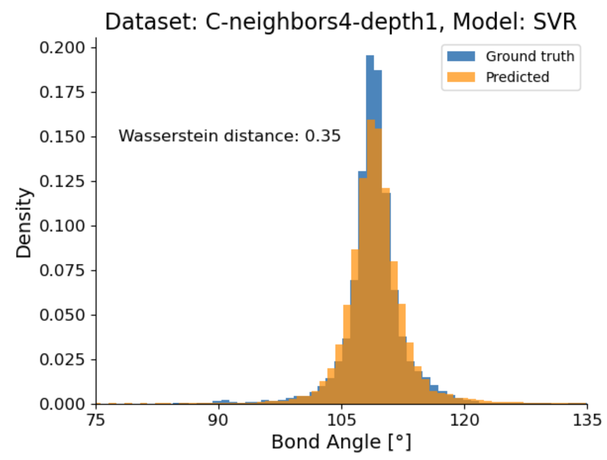
In total, we trained 11 different models from the following model types:

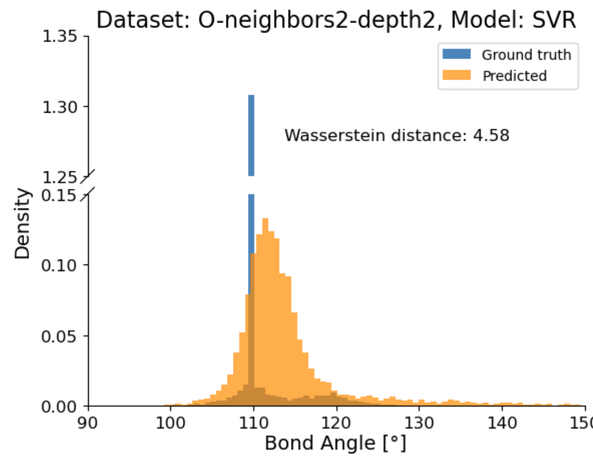
Bond Length





Bond Angle

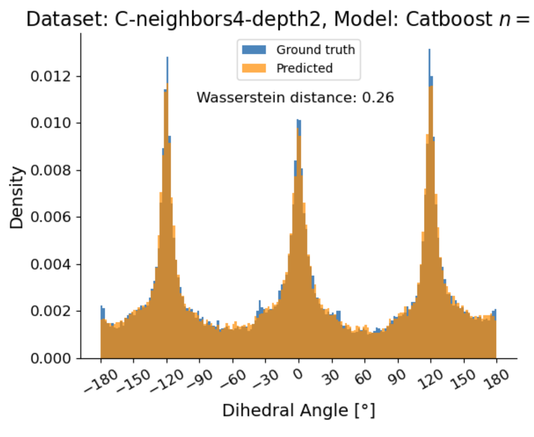


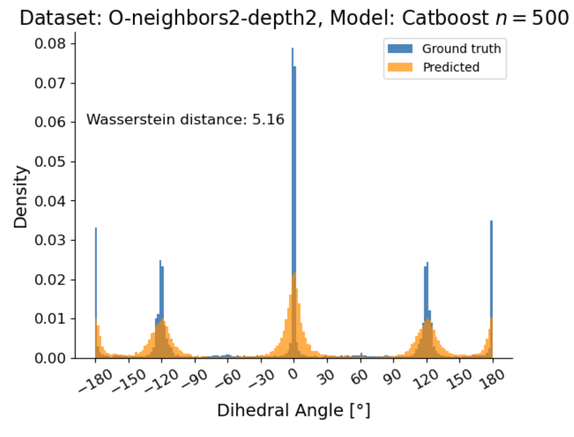


Evaluation

Dataset	Model	Metrics					
		MSE	R^2	Angle error (Degree)	Bond length error (Ångström)	Bond angles (Wasserstein distance)	Dihedral angles (Wasserstein distance)
C-neighbors4-depth1	support vector regressor	0.01	0.98	**3.5**	0.02	**0.35**	n/a
	simple MLP	**0.01**	**0.98**	4.86	**0.01**	1.09	n/a
	catboost $n = 500$	0.01	0.98	4.62	0.02	0.95	n/a
C-neighbors4-depth2	catboost $n = 500$	0.01	0.97	4.94	**0.02**	1.17	0.26
	randomforest $n = 500$	0.01	0.98	**4.28**	0.06	**0.36**	0.2
	simple MLP	**0.01**	**0.98**	5.25	0.03	1.29	0.32
O-neighbors2-depth1	catboost $n = 200$	**0.26**	**0.14**	**65.04**	**0.55**	**39.36**	n/a
	gradient boosting $n = 500$	0.26	0.14	65.48	0.56	41.76	n/a
	simple-MLP	0.26	0.14	65.92	0.58	46.31	n/a
O-neighbors2-depth2	simple MLP	0.1	0.66	24.19	0.17	7.74	4.86
	catboost $n = 500$	**0.1**	**0.66**	24.96	0.2	8.82	5.16
	support vector regressor	0.11	0.65	**22.24**	**0.12**	**4.58**	3.89

Dihedral Angle





Conclusion

- Hydrogen position was predicted accurately
- The model with depth 1 is performing well with carbon but poorly with the oxygen as central atoms
- The oxygen model improves when including more neighbors

Outlook

- Predicting hydrogen positions for different atom types
- Predicting positions of more than one missing Hydrogen bonded to the central atom
- Implement a Graph Neural Network for improved input invariance