Hybrid ML and Physics-Based Docking for Metalloproteins

Wojtek Treyde[a,b], Franz Görlich[a,c], and Fernanda Duarte[a]

[a]Chemistry Research Laboratory, and [b]Ineos Oxford Institute for Antimicrobial Research, 12 Mansfield Road, Oxford, OX1 3TA, UK. [c]Department of Chemistry and Applied Life Sciences, ETH Zurich, Vladimir-Prelog-Weg 1-5/10, 8093 Zurich, CH.

Motivation

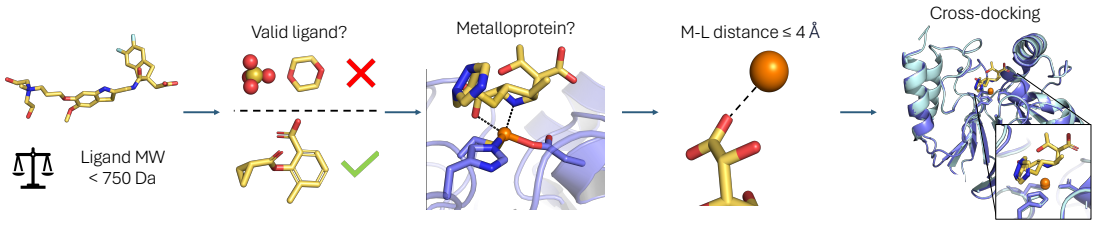
Docking has largely neglected M-L interactions.

Available tools[1] have limited scope, are closed-source, or require manual setup.

Goal: Improve pose prediction for metalloproteins.

Dataset Curation

Filtered based on quality and relevance.



The largest dataset of metalloprotein-ligand structures to date: **41k** training, **1.1k** validation, and **155** test complexes.

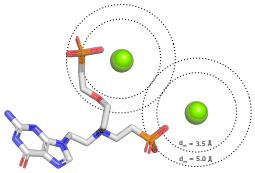
Data and splitting strategy based on PLINDER[2] .

Baseline Methods

Baseline docking tools
- AutoDock Vina[3]
- SMINA[4]
- GNINA[5]

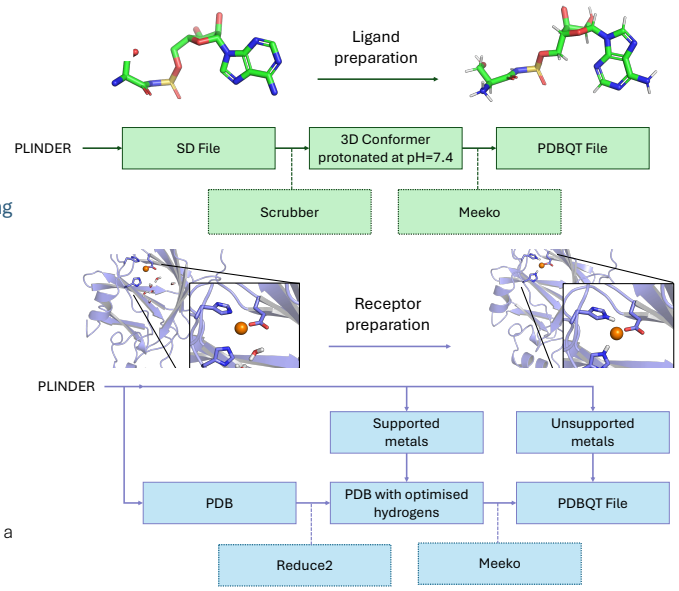
Evaluated using the metal-binding group (MBG) success rate

(Predictions with RMSD < 2 Å)



The MBG is defined as all atoms within a certain radius around metal ions.

Structures preparation



CHELATE

Coordination Heuristics for Estimating Ligand Association in Those Enzymes

CHELATE

CatBoost model trained on MBG RMSD.

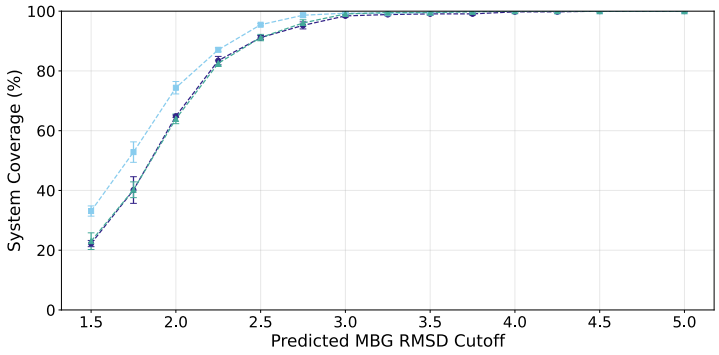
Features:
- Sum of inverse, mean, minimum, maximum and standard deviation of distances between ligand and protein atoms at different distance cut-offs, and their count.
- Docking score.
- One-hot encoded metals.

Code coming soon!

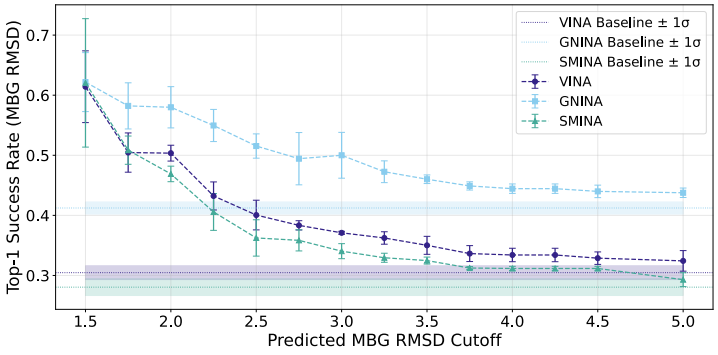
Benchmark Results

CHELATE filters poses based on predicted MBG RMSD, significantly improving the MBG success rate

Fraction of systems with at least one pose remaining



MBG success rate



Conclusions

Curated **the largest dataset of ligand-metalloprotein complex structures** to date.

Benchmarked **baseline models** (Vina, SMINA, GNINA), but found **poor performance**.

Developed CHELATE, a general metalloprotein docking tool.

Improves pose prediction based on MBG success rate.

Outlook

More accurate MBG placement
→ can be used for bonded models in MD simulations to further improve accuracy

References:
1. a) Wang *et al.*, *Brief. Bioinform.* **2023**, 24, 1–11. b) Jiang *et al.*, *Chem. Sci.* **2023**, 14, 2054–2069. c) Clemente *et al.*, *J. Chem. Inf. Model.* **2024**, 64, 1581−1592.
2. Durairaj, Adeshina, *et al.*, *bioRxiv* 2024.07.17.603955.
3. Trott, *et al.*, *J. Comput. Chem.* **2010**, 31, 455-461.
4. Koes, *et al.*, *J. Chem. Inf. Model.* **2013**, 53, 1893–1904.
5. McNutt, *et al.*, *J. Cheminform.* **2021**, 13, 43.

WT thanks the Ineos Oxford Institute for Antimicrobial Research, the Department of Chemistry and St John's College for their support.

Duarte group:

Email:
wojtek.treyde@sjc.ox.ac.uk