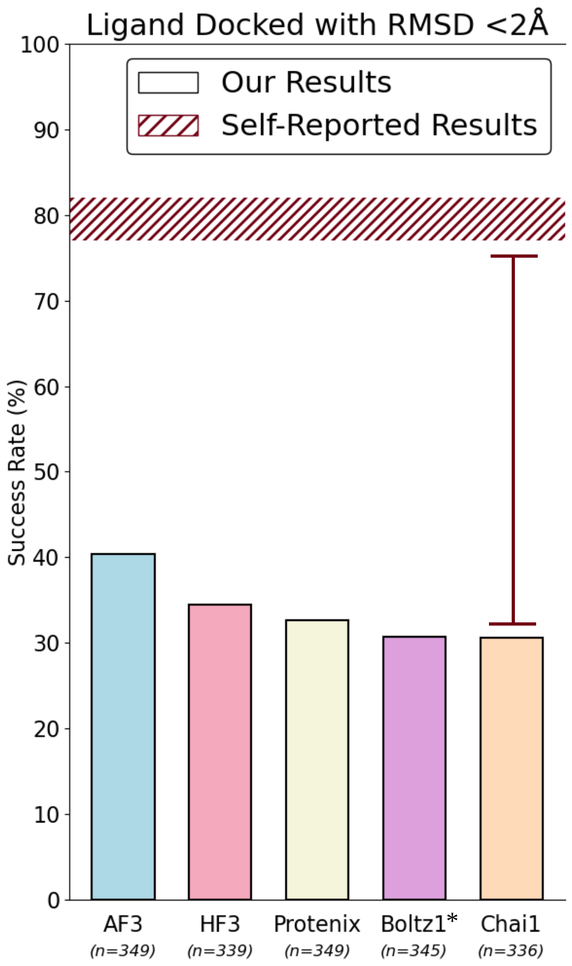
UNIVERSITY OF COPENHAGEN
CENTER FOR PHARMACEUTICAL DATA SCIENCE EDUCATION

CHALLENGING DEEP LEARNING DOCKING METHODS (DLDMS)

Johan Rasmussen and Victor Varming Rothe, Master Students (Pharmaceutical Sciences)
Supervised by: Albert J. Kooistra, Chris de Graaf and Jonas Verhellen.



WHY ARE SELF-REPORTED SUCCESS RATES SO INFLATED?

This can possibly be explained by differences in the testing set. The testing sets employed have several pitfalls. The widely used Posebusters v2 dataset[1] is used as example:

- **High Similarity** between training and testing set. 26% of complexes with ≤ 70% seq identity to testing set complexes[1].
- **Unvaried drug-like targets**: 80% enzymes, 0% GPCR, 2% channel/transporter[1].
- **Non-drug-like ligands**: High prevalence of nucleotides, flavinoids, sugars, amino acids, organic acids/ethers/alcohols[1].
- **Missing ligands**: 52% of dataset are missing a ligand[1].

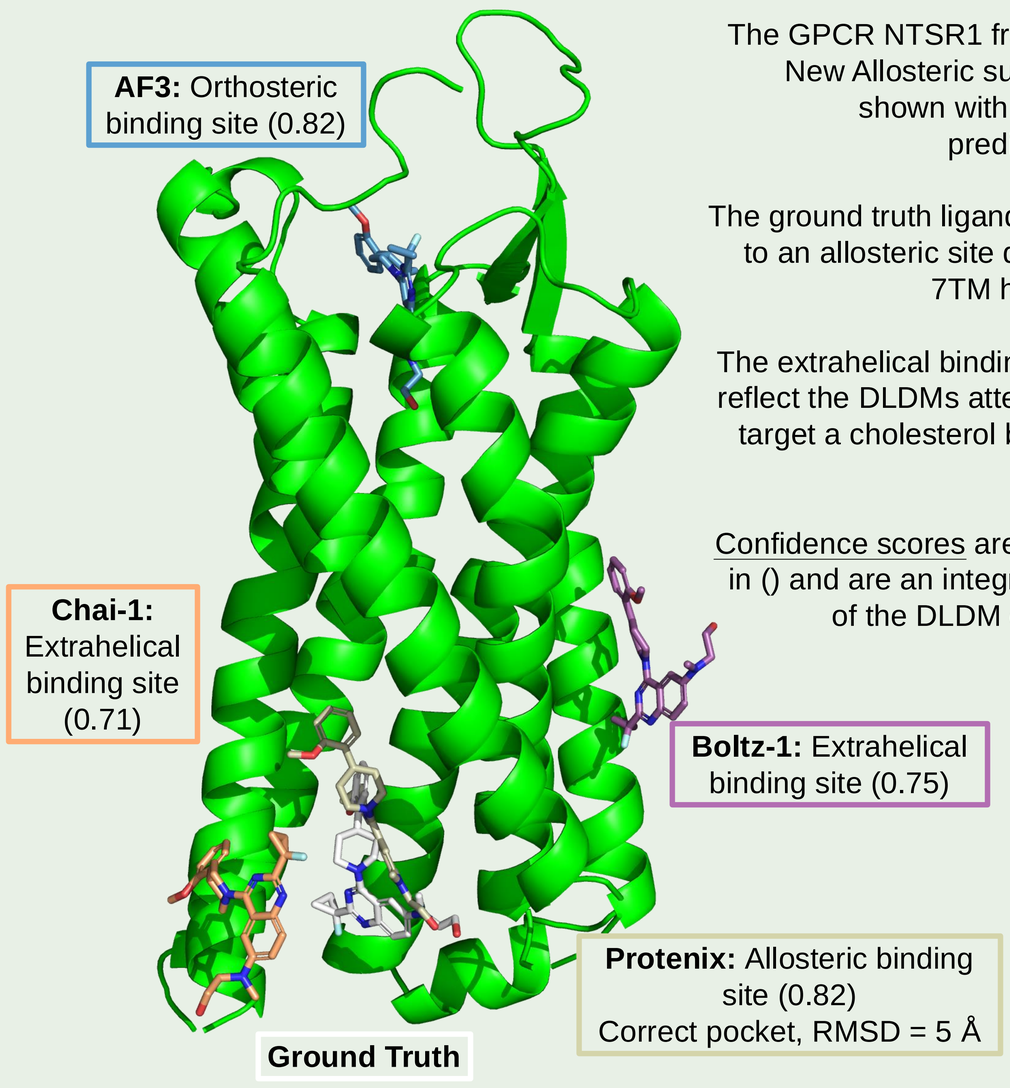
Do DLDM infer and learn, or do they just replicate scenarios from training?

WE PRESENT DEEPDOCKINGDARE

DeepDockingDare (DDD) is a curated dataset for testing of DLDMs such as AF3. DDD consists of 3 subsets, each providing different challenges for DLDMs:

- Sequence Dissimilarity: ≤ 30% sequence identity to any complex in the training set.
- New Modality: Different ligand type compared to training set e.g., agonist to inhibitor.
- New Allosteric: Novel allosteric sites on training set proteins

NEW ALLOSTERIC SITES ARE TRICKY



The GPCR NTSR1 from the New Allosteric subset is shown with DLDM predictions.

The ground truth ligand binds to an allosteric site deep in 7TM helices.

The extrahelical binding may reflect the DLDMs attempt to target a cholesterol binding site.

Confidence scores are noted in () and are an integral part of the DLDM output!

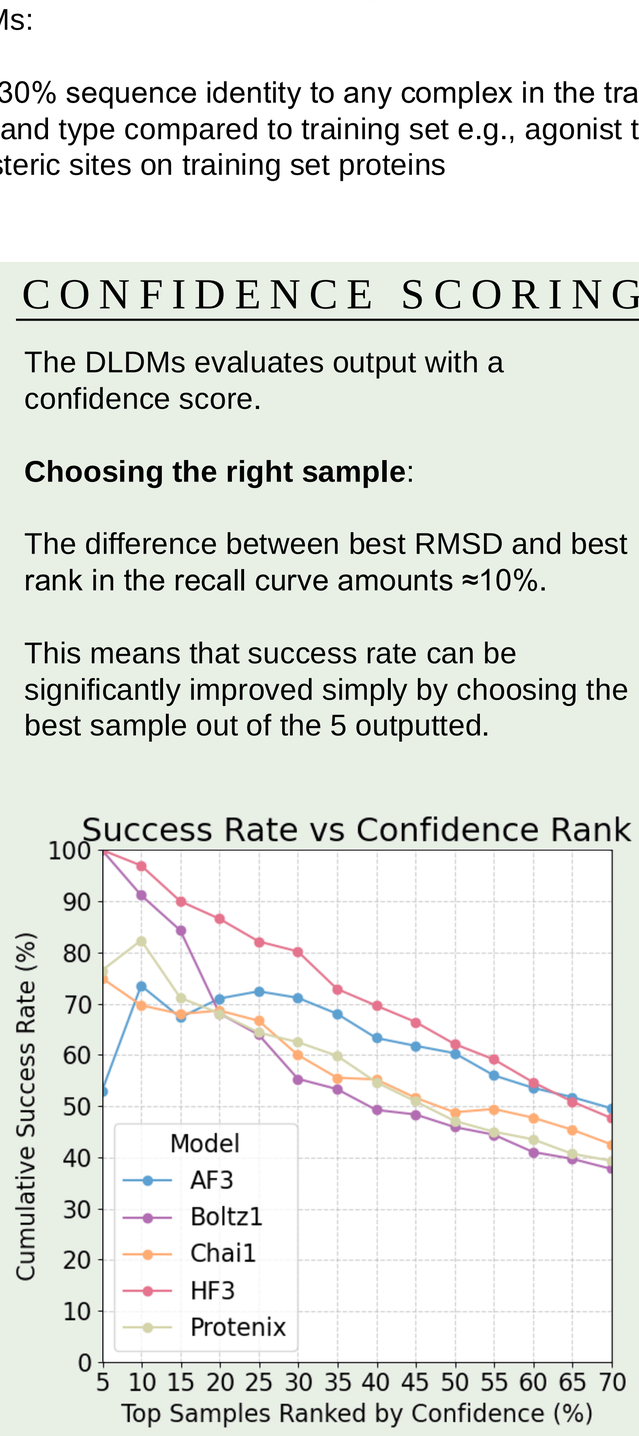
CONFIDENCE SCORING

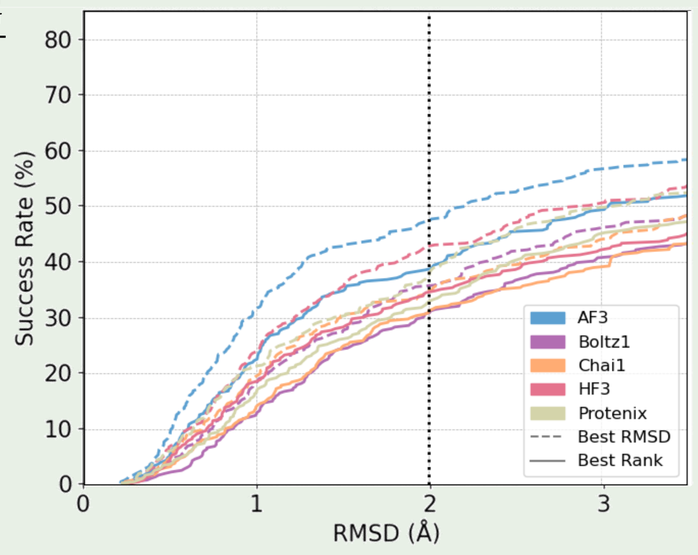
The DLDMs evaluates output with a confidence score.

Choosing the right sample:

The difference between best RMSD and best rank in the recall curve amounts ≈10%.

This means that success rate can be significantly improved simply by choosing the best sample out of the 5 outputted.





Using confidence scores for result validation:

DLDMs properly aren't ready for real world application yet, because of highly varying quality of results.
But perhaps reliability can be improved by only using results over a certain confidence threshold?

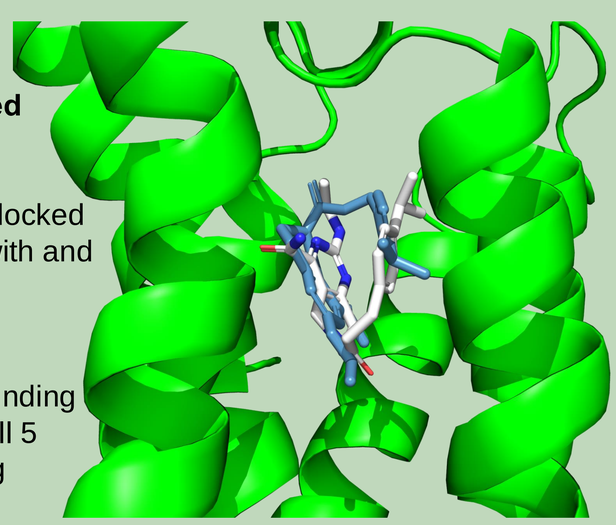
The plot shows the tested DLDMs ability correlate confidence to prediction quality. This can be used to filter out bad predictions.

GPCR MODELLING: CAN WE IMPROVE THE RESULT?

Should Gα be included in the DLDM input?

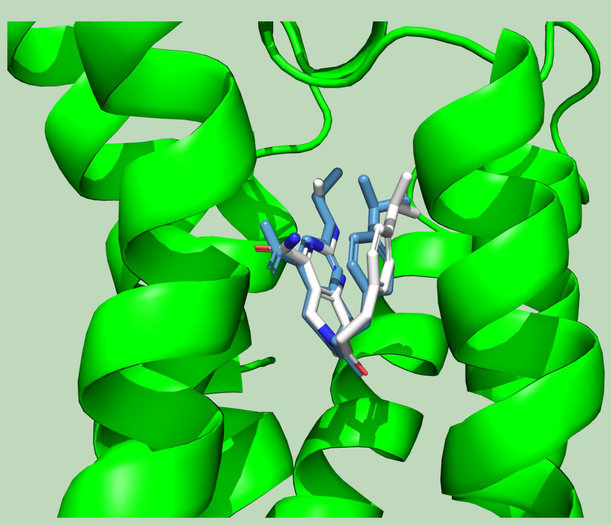
GPBAR is folded and docked in the orthosteric site with and without Gα by AF3.

Including Gα produces slight changes to the binding pocket which, across all 5 samples, turns a wrong prediction right!



Ground truth: White
AF3 predicted ligand: Blue
RMSD = 4.9 – 5.6 Å *(n=5)*

GPBAR without Gα

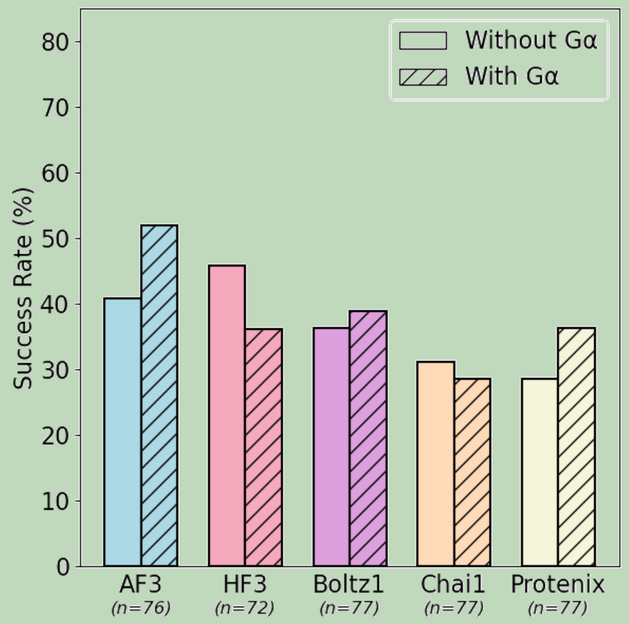


Ground truth: White
AF3 predicted ligand: Blue
RMSD = 0.8 – 0.9 Å *(n=5)*

GPBAR with Gα

Some methods show significantly **improved performance** (AF3 and Protenix) while others show a **decrease** (HF3) with Gα included.

This highlights the methods' sensitivity to which protein chains are included in the input even though they don't interact with the binding site.



WHAT'S NEXT?

Having assessed performance, we now aim to explore the factors driving these results.
We are also expanding our analysis to include more docking methods (classical and DLDM)

References:
[1] Buttenschoen M, Morris GM, Deane CM. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chem Sci. 2023;15(9):3130-3139. Published 2023 Dec 13
[2] Discovery C, Boitreaud J, Dent J, McPartlon M, Meier J, Reis V, et al. Chai-1: Decoding the molecular interactions of life. bioRxiv. 2024:2024.10.10.615955.
[3] Team BAAS, Chen X, Zhang Y, Lu C, Ma W, Guan J, et al. Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. bioRxiv. 2025:2025.01.08.631967.
[4] Wohlwend J, Corso G, Passaro S, Reveiz M, Leidal K, Swiderski W, et al. Boltz-1: Democratizing Biomolecular Interaction Modeling. bioRxiv. 2024:2024.11.19.624167.
[5] Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493-500.
[6] Liu L, Zhang S, Xue Y, Ye X, Zhu K, Li Y, Liu Y, Gao J, Zhao W, Yu H, Wu Z. Technical report of HelixFold3 for biomolecular structure prediction. arXiv preprint arXiv:2408.16975. 2024 Aug 30.
∗ Boltz1 uses their own dataset, which isn't directly comparable to the others applied. They compare results to Chai-1 with about the same performance in that dataset of around 60% success rate.