

# Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction

<https://doi.org/10.1101/2025.06.14.659707>

Saro Passaro\*, Gabriele Corso\*, Jeremy Wohlwend\*, Mateo Reveiz\*, Stephan Thaler\*, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, Regina Barzilay

Reviewed by: Karson M. Chrispens, James S. Fraser, Stephanie A. Wankowicz, 2025  
Conformational Ensembles Workshop

Stephanie A. Wankowicz - [0000-0002-4225-7459](https://orcid.org/0000-0002-4225-7459)

James S Fraser - [0000-0002-5080-2859](https://orcid.org/0000-0002-5080-2859)

Karson M. Chrispens - [0000-0002-2115-3132](https://orcid.org/0000-0002-2115-3132)

## Summary

The authors describe Boltz-2, a major update to the Boltz-1 macromolecular structure prediction model. While Boltz-1 focused on structure prediction and rivaled the accuracy of AlphaFold3, particularly on protein-ligand complexes, Boltz-2 introduces additional training data to the main structure prediction model, architecture modifications, and an additional module for protein-ligand affinity prediction. This new module, whose architecture parallels that of the main trunk of Boltz-1, utilizes curated ligand-binding data in the form of K<sub>d</sub> and IC<sub>50</sub> values from several publicly available assays for the supervised prediction of binding affinity. The module is trained to both classify binder/non-binder and also regress the affinity value. The new model is benchmarked against the competing models in the field for structure prediction and shows some marginal improvements. The performance of the affinity module is benchmarked against physics-based free energy perturbation (FEP) methods, as well as other machine learning-based approaches.

Boltz-2 represents a major engineering achievement. The incorporation of experimental method conditioning and new methods for guiding the model with experiments will advance the field of structure prediction and expand the applications of these models beyond affinity predictions.

The major weaknesses of the paper and the method are primarily related to the evaluation and analysis of the model's performance on realistic tasks. The protein and ligand training set lacks a principled split (the standard in the field is to use a scaffold split). This contributes to concerns about data leakage. We also believe that there should be more details on how and what data were used in the different stages of training, for example, when and how MD simulations were employed, where certain split datasets were used, and how diverse ligand affinity measurements were used, especially since the training data and processing pipeline has not

been released yet. Additionally, the paper lacks a clear evaluation of which pieces of this training were used for the classification vs. regression tests.

Overall, we are excited about the release of this model, including the open-source evaluation code, and view it as a significant step forward in the comprehensive modeling of macromolecular structure and interactions.

## Feedback

- While we understand the desire to ship the model weights quickly, we believe the community would benefit from directly seeing how different pieces of data are utilized in training and that this is nearly as important to the field as the model weights. To encourage transparency, please clarify the delay in releasing training data.
  - Note that this peer review was posted when the [Boltz-2 2.1.1 release](#) was the latest available, and did not have the training data processing pipeline.
- Overall, we are concerned about data leakage between training and testing. We are curious why the authors used the split they did, and would appreciate commentary here, given the challenge of evaluating these models thoroughly.
  - Ligands: scaffold splitting instead of Tanimoto split, as Tanimoto can be low, but still have substructures that are driving high affinity.
  - The sequence similarity cut-off is high. The field has also shown that proteins should be split on fold rather than sequence similarity.
  - Please clarify where and how data was split during training. Throughout the paper, it is unclear when the 30% v. 90% sequence similarity was used as cut off.
  - The majority of FEP benchmarks are likely built by simulating proteins that have structures deposited before 2023 (and thus included in training data). Please clarify how this possible data leakage was handled.
- The affinity module performs both classification and regression. What is the relationship or sequence of operations between these tasks?
- We would appreciate discussion on the extent to which the affinity module depends on the initialization of weights from the original trunk. Can this be tested with ablation studies? Or perhaps compare the performance of the original trunk with a simple regression head to the trained affinity module?
- We would like the authors to discuss or hypothesize why the physical validity of models is reduced with Boltz-2x compared to Boltz-1x.
- We encourage the authors to specify what data is being used in the main training run vs. additional tweaks.
  - For example, was there a specific rationale for when to add MD data during training?
  - Please provide additional clarification on what was tweaked or fine tuned after the main training run.
- It was exciting to see training data on conformational heterogeneity, but as the MD was only added partway through training, the MD is very short, and there is no water modeling, so it is challenging to determine where and how this is helping. We would like

to see additional discussion of how the short MD simulations have improved the model performance.

- Given the diverse set of training data used in Boltz-2 compared to Boltz-1 and previous models, can the authors comment on strategies or difficulties of combining these different data types? We expect that additional types of experimental data will be implemented in the future, such as Shape-Seq data for RNA. What sorts of experimental modalities are currently plausible to incorporate, and where do the authors see a particular need for more curation of experimental data so that the community can contribute to advancing the model's performance?
- The generative work on small molecules with the SynFlowNet is an interesting application, and we would love to see a comparison to how close these molecules are in the Enamine REAL dataset, and how useful this might be for generating/docking molecules to use as queries in a fuzzy search over the existing databases of synthesizable molecules.
- We would like to see more testing on edge cases and assessment of performance on hard problems. For example, where does Boltz-2 break down? For example, are there specific protein families, binding modes (e.g., allosteric vs orthosteric), or structural motifs where affinity predictions systematically underperform?
- Since Boltz-2 includes MD conditioning, it would be valuable to quantify how it performs on systems known to undergo large conformational changes upon binding.
  - The paper mentions that it does not generalize to these cases, but we think it would be helpful to know more about the specific details of the failure cases and how the authors think they might be overcome.