Boltzina: Efficient and Accurate Virtual Screening via Docking-Guided Binding Prediction with Boltz-2

Kairi Furui
Department of Computer Science
School of Computing
Institute of Science Tokyo
Kanagawa, 226-8501, Japan
furui@li.comp.isct.ac.jp

Masahito Ohue
Department of Computer Science
School of Computing
Institute of Science Tokyo
Kanagawa, 226-8501, Japan
ohue@comp.isct.ac.jp

Abstract

In structure-based drug discovery, virtual screening using conventional molecular docking methods can be performed rapidly but suffers from limitations in prediction accuracy. Recently, Boltz-2 was proposed, achieving extremely high accuracy in binding affinity prediction, but requiring approximately 20 seconds per compound per GPU, making it difficult to apply to large-scale screening of hundreds of thousands to millions of compounds. This study proposes Boltzina, a novel framework that leverages Boltz-2's high accuracy while significantly improving computational efficiency. Boltzina achieves both accuracy and speed by omitting the rate-limiting structure prediction from Boltz-2's architecture and directly predicting affinity from AutoDock Vina docking poses. We evaluate on eight assays from the MF-PCBA dataset and show that while Boltzina performs below Boltz-2, it provides significantly higher screening performance compared to AutoDock Vina and GNINA. Additionally, Boltzina achieved up to 11.8× faster through reduced recycling iterations and batch processing. Furthermore, we investigated multi-pose selection strategies and two-stage screening combining Boltzina and Boltz-2, presenting optimization methods for accuracy and efficiency according to application requirements. This study represents the first attempt to apply Boltz-2's high-accuracy predictions to practical-scale screening, offering a pipeline that combines both accuracy and efficiency in computational biology. The Boltzina is available on github; https://github.com/ohuelab/boltzina.

1 Introduction

In drug discovery research, structure-based drug design (SBDD) is a method that utilizes three-dimensional structural information of target proteins to design and evaluate novel compounds [1]. Among these approaches, virtual screening (VS) [2] has been widely used to select promising candidate molecules from vast compound libraries. In conventional VS, pose generation by molecular docking and binding affinity prediction using scoring functions have been standard methods [3, 4], but scoring functions based on physical models or empirical rules have limitations in accuracy [5, 6].

To address this challenge, machine learning-based scoring functions (MLSFs) have been proposed [7, 8, 9]. Diverse MLSF approaches range from classical Random Forests using interaction features [10] to neural network methods such as convolutional networks based on contact information [11] and graph neural networks operating on atomic and interaction graphs [12]. While MLSFs sometimes show higher accuracy than conventional methods, they still face issues such as data dependency and insufficient generalization to unknown targets, leaving reliability challenges in actual drug discovery applications [9, 13].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat: AI for Accelerated Materials Design.

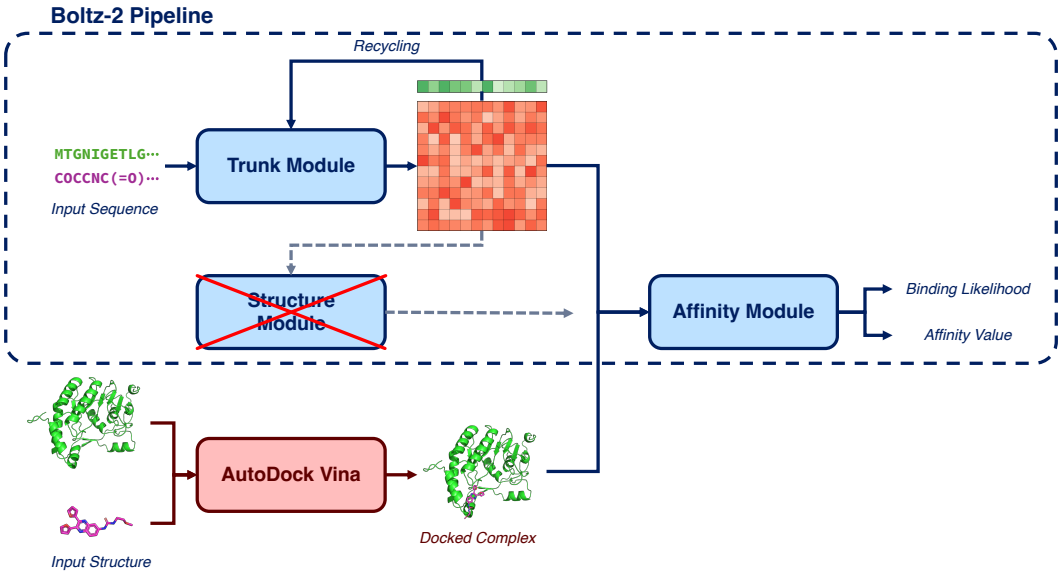

Figure 1: Overview of the Boltzina pipeline. The dashed blue box represents the original Boltz-2 affinity prediction pipeline, and the red × mark on the structure module indicates that it is omitted in Boltzina.

The recently proposed Boltz-2 [14] integrates structure prediction and binding affinity prediction using an AlphaFold3-like [15] diffusion model, achieving performance that significantly exceeds conventional docking-based methods and MLSFs. Boltz-2 demonstrates performance approaching molecular simulation-based free energy calculations [16, 17] by incorporating an affinity module for protein–ligand complexes in addition to Boltz-1's [18] structure prediction capabilities. Furthermore, evaluation using MF-PCBA [19] reported screening performance that significantly surpasses existing compound-protein interaction (CPI) prediction models [20] and empirical scoring methods [21].

However, Boltz-2 requires approximately 20 seconds per ligand for prediction, making direct application to large-scale libraries exceeding one million compounds impractical [22, 23]. This is because Boltz-2 requires a diffusion process for structure prediction, which causes the computational time bottleneck in affinity prediction. To solve this challenge, methods that maintain Boltz-2's high accuracy while improving computational efficiency are required.

This study proposes Boltzina, a framework that achieves rapid compound screening by directly predicting Boltz-2's affinity and binding from poses generated by the existing docking method AutoDock Vina [4]. We compare these methods on the MF-PCBA dataset and examine their applicability to large-scale screening. Furthermore, we decompose Boltzina's components to clarify the accuracy–speed trade-off, and evaluate multi-pose selection strategies and two-stage screening with Boltz-2 to discuss applicability in actual large-scale screening.

2 Materials and Methods

2.1 Boltzina Pipeline

Boltzina is based on Boltz-2's architecture as shown in Figure 1. In the original Boltz-2, binding affinity is predicted through staged information processing using the trunk module, structure module, and affinity module. First, the trunk module extracts latent structural features from input protein sequences and ligands. The trunk module primarily consists of PairFormer and MSA modules, generating pairwise representations that capture intermolecular interactions. This latent representation implicitly contains structural information. Next, the structure module predicts 3D structures based on the latent representation from the trunk module, determining atomic coordinates and the geometric arrangement of protein–ligand complexes. Finally, the affinity module predicts binding affinity using intermolecular interaction information obtained from the trunk module and explicit 3D coordinate

Table 1: Information on the 8 MF-PCBA test set assays used for evaluation and their corresponding ligands. PDB IDs were used as references for grid positions when available. Active and Inactive indicate the number of active and inactive compounds in each assay, with failed ligands excluded.

PubChem AID	PDB ID/CID	Active	Inactive	Failed Ligand
743445	6UE6	144	49851	5
485317	3MBG	976	49004	20
2097	1J1B	522	49393	8
493091	3PGL	782	49203	15
2650	7SXF	612	49270	11
504329	CID1474141	466	49516	18
588689	3EVG	486	49499	15
588549	CID2277079	159	49828	11

information predicted by the structure module. The affinity module performs processing specialized for protein–ligand interface interactions through a dedicated PairFormer architecture, outputting two predicted values: binding likelihood and affinity value. In Boltzina, instead of using the 3D coordinate information from the structure module, poses generated by rigid docking with the external docking software AutoDock Vina are directly input to the affinity module, thereby omitting the structure module. This allows Boltzina to retain Boltz-2's high-accuracy intermolecular interaction analysis capabilities while avoiding the computational cost required for structure prediction.

In our implementation, PDB files of docking structures generated by AutoDock Vina were converted to appropriate MMCIF-format complex structures using PDB-tools [24] and the MAXIT suite [25]. These converted structures were then processed into the model's input format by reusing Boltz-2's template structure processing implementation. Furthermore, we constructed a batch processing pipeline for efficiently handling multiple complexes and improved processing efficiency by making the batch size variable, whereas it was fixed at 1 in Boltz-2.

2.2 Dataset

For evaluation, we used a test set independently constructed from the MF-PCBA dataset [19], following [14]. This dataset is a virtual screening benchmark for developing and evaluating machine learning methods in drug discovery, containing multiple targets collected from PubChem, and was also used in Boltz-2 to evaluate model screening performance. In this study, we conducted experiments on 8 out of 10 MF-PCBA assays evaluated in Boltz-2, as listed in Table 1. Protein sequences similar to these test data were appropriately filtered in Boltz-2's training for affinity prediction tasks [14]. Here, AID489030 was excluded because the clear binding pocket was unknown and the grid could not be determined, and AID485273 was excluded because the active ligands included a high proportion of large molecules with more than 60 heavy atoms, which would compromise pose estimation by AutoDock Vina. For large molecules with more than 60 heavy atoms, AutoDock Vina's pose estimation often requires more than 5 minutes, which is outside the scope of fast screening intended by Boltzina, and was therefore excluded from evaluation in this study.

In preprocessing, after applying PAINS (Pan-Assay Interference Compounds) filtering [26], all binders and non-binders were randomly sampled to total 50,000 cases, then duplicates were removed. Furthermore, molecules for which AutoDock Vina failed (e.g., those containing arsenic or boron atoms) were excluded. Additionally, in docking calculations, cases that required more than 5 minutes of computation time were also treated as timeouts and considered failures. As mentioned above, most molecules that failed due to timeouts were large molecules with 60 or more heavy atoms. Screening performance was finally evaluated with the numbers shown in Table 1. To evaluate execution time and the effects of using multiple docking poses, 1,000 ligands were randomly sampled from the MF-PCBA test set.

2.3 Docking Settings

AutoDock Vina [27] v1.2.7 was used for docking pose generation. The grid box size for AutoDock Vina docking was set to 20 Å, a size widely adopted in general molecular docking studies [28], and

exhaustiveness was set to the default value of 8. For the main screening performance evaluation, only the best pose predicted by AutoDock Vina was evaluated with Boltz-2. The effects of varying AutoDock Vina parameters are shown in Supplementary Figure 7. Docking was performed using one protein–ligand complex structure predicted by Boltz-2 as a reference structure for each target. For the six assays where clear known complex structures existed in the PDB database, those ligands were used; for the remaining two assays, the most potent ligand among MF-PCBA binders was used as input to predict holo structures (Table 1). Because Vina requires explicit specification of binding pockets, the centroid of the holo ligand was used as the grid center. We confirmed that most binders in Boltz-2's complex prediction structures bound to the same region as the aforementioned predicted holo structures, validating the appropriateness of the grid position.

For the experiments, we used computational nodes from the TSUBAME4 supercomputer at the Institute of Science Tokyo. The operating system was Red Hat Enterprise Linux Server 9.3. For each computation, one H100 GPU and 48 of the 96 cores from AMD EPYC 9654 2.4GHz processors were used, with 192GiB of memory available. Docking calculations were executed in 48 parallel processes, assuming actual screening scenarios. Execution time measurements were performed under these parallelization conditions.

2.4 Comparison Methods

For screening performance evaluation, we compared Boltzina's performance with the following methods:

Boltz-2 Predictions using the original Boltz-2 with default settings.

AutoDock Vina Conventional molecular docking using the above settings.

GNINA Open-source software GNINA v1.3.2 [7, 29], which incorporates CNN-based scoring functions on top of AutoDock Vina and Smina [30]. GNINA ranked nine poses generated by AutoDock Vina using CNN VSScore (the product of CNN affinity and CNN score) [31], and selected the highest-scoring pose.

We also evaluated two parameters to examine each component's contribution to performance and effects on execution speed:

Boltzina (Cycle=1) Recycling in the trunk module reduced from the default five iterations to one (see Figure 1).

Boltzina (No Pose) Docking pose information masked by setting all ligand atom coordinates to the origin, eliminating initial pose dependency. Since docking pose information is not used, docking time is excluded from execution time.

For performance evaluation, we used Average Precision (AP) and Enrichment Factor at top percentiles, metrics specialized for hit discovery, similar to the Boltz-2 paper. Average Precision (AP) is a metric that increases when more true active compounds appear at the top, defined as the area under the precision-recall curve. Enrichment Factor (EF) is a metric that divides the proportion of active compounds contained in the top $K\%$ of candidates by the expected value from random selection, showing how efficiently actives can be found compared to random search.

2.5 Pose Selection Strategies

To evaluate the impact of using multiple docking poses, we tested the following strategies:

Best Pose Only Evaluating only the best (minimum-energy) pose from AutoDock Vina.

Top-N Best Score Selecting the highest Boltzina binding likelihood among the top N poses generated by AutoDock Vina.

Top-N Average Ranking by the average of the Boltzina predicted affinity scores over the top N poses.

Pose selection strategies were evaluated by randomly sampling 1,000 molecules per assay.

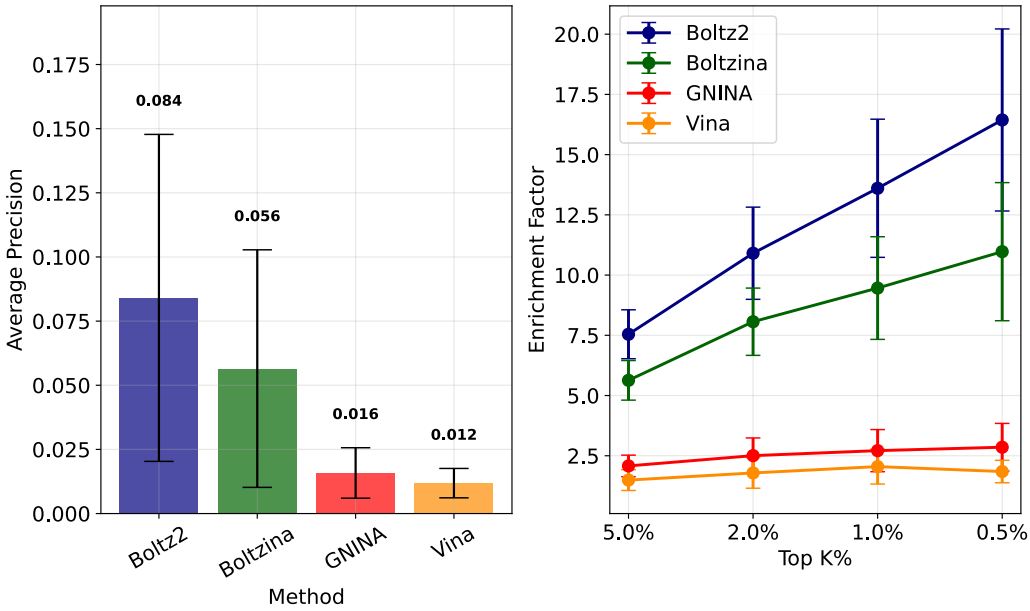

Figure 2: Comparison with existing methods on the MF-PCBA test set. a) Mean Average Precision across assays. b) Average Enrichment Factor at top $K\%$ ($K = 0.5\%, 1\%, 2\%, 5\%$).

2.6 Two-Stage Screening Experiment

To achieve an optimal balance between computational efficiency and prediction accuracy, we considered a hierarchical screening strategy [32] combining Boltzina and Boltz-2. In this strategy, methods with different computational costs are combined in stages. In the first stage, all compounds are rapidly screened using Boltzina; in the second stage, only promising compounds ranked highly are re-evaluated in detail with the more accurate Boltz-2. Specifically, the top $p\%$ of compounds were selected based on Boltzina's binding likelihood, these were rescored with Boltz-2, and final rankings were determined. Evaluation was performed under four conditions with p values of 50%, 20%, 10%, and 5%. At this time, estimated execution time was calculated as $T_{Boltzina} + \frac{p}{100} T_{Boltz2}$.

3 Results and Discussion

3.1 Comparison with Existing Methods

We evaluated the proposed method's screening performance using eight assays from the MF-PCBA dataset. As shown in Figure 2a, for Average Precision (AP), Boltz-2 showed the highest performance (mean AP 0.084), followed by Boltzina (mean AP 0.056). In contrast, the mean AP of existing GNINA and Vina methods was extremely low. Therefore, Boltzina achieved significant performance improvement compared to AutoDock Vina and GNINA.

The Enrichment Factor (EF) results shown in Figure 2b also exhibited similar trends. While GNINA showed slight improvement over AutoDock Vina through reranking, it was inferior to Boltzina's performance increase. Additionally, at the top 5%, the performance difference between Boltz-2 and Boltzina was relatively small, whereas at the top 0.5% there was a substantial difference between the two methods. This suggests that Boltz-2's precise calculations are essential for accurate ranking of a very small number of top compounds, while Boltzina is particularly effective for screening scenarios requiring medium-scale enrichment. Figure 3 shows ROC curves for each assay. While AutoDock Vina showed near-random performance for many targets, it improved significantly with Boltz-2 rescoring. Additionally, for some assays, the performance difference between Boltz-2 and Boltzina was small, indicating cases where Boltzina approached Boltz-2's performance.

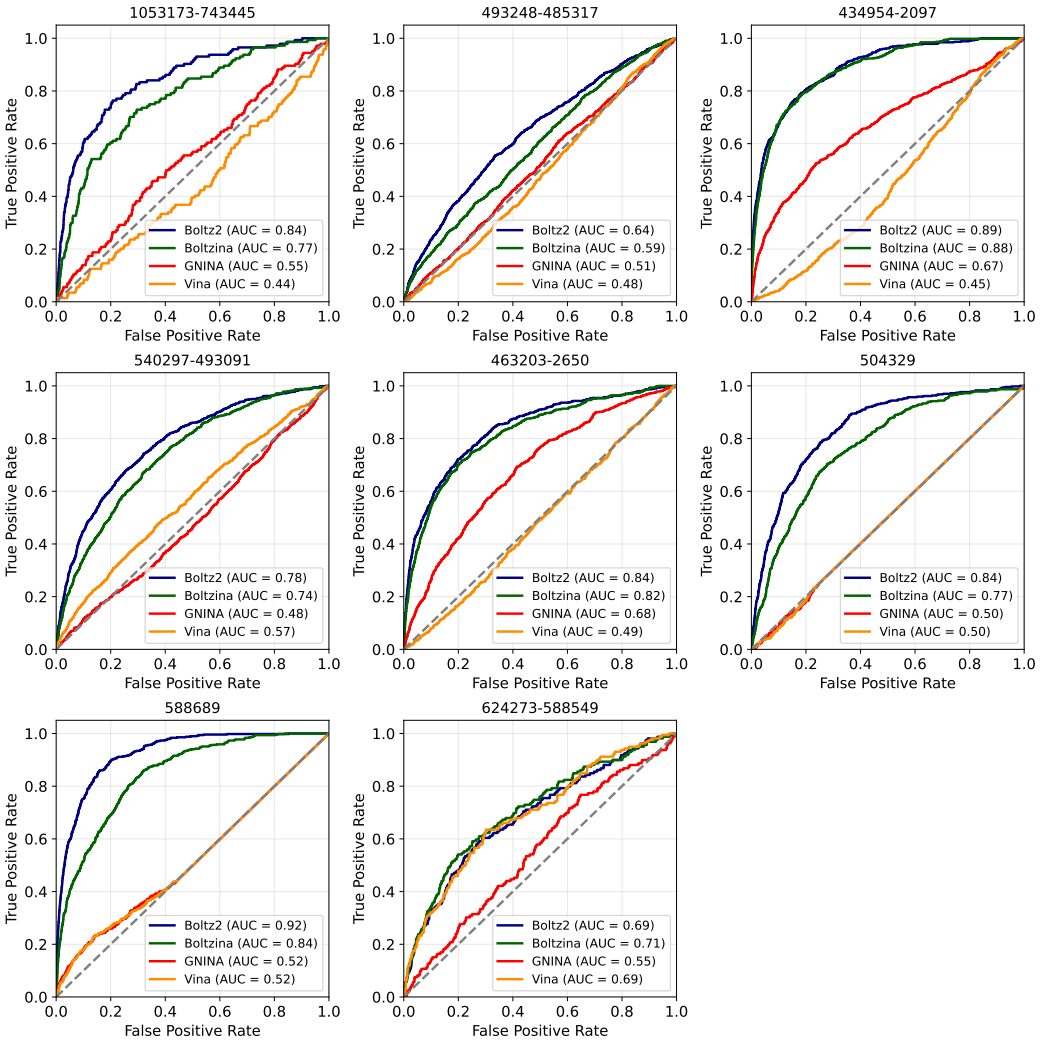

Figure 3: ROC curves for each assay in the MF-PCBA test set.

Next, Figure 4 shows comparisons of different parameters for Boltzina. For Boltzina (Cycle=1), the mean AP decreased from 0.056 to 0.048, but the decrease was limited, confirming that reducing the number of recycling iterations is effective for cutting computational cost without significantly compromising accuracy. On the other hand, for Boltzina (No Pose), even when ligand pose information was not provided, the mean AP was 0.043; while there was a negative impact on performance, it was more limited than expected. This result suggests that intermolecular interaction information obtained from the trunk module plays an important role in binding prediction. That is, the reason performance improves without depending on AutoDock Vina's pose accuracy is suggested to be the latent representations of intermolecular interactions learned by the trunk module.

Next, as shown in Figure 5a, Boltzina achieved a significant reduction in execution time. The average processing time per ligand was approximately 16.5 seconds for Boltz-2, compared to 2.3 seconds for Boltzina, which is 7.3× faster, and 1.4 seconds for Boltzina (Cycle=1), which is 11.8× faster. The main factors for this speedup are omission of structure prediction steps, improved parallel processing efficiency through increased batch size, and reduced recycling iterations. Notably, AutoDock Vina docking required approximately 0.8 seconds and became rate-limiting for Boltzina (Cycle=1). Therefore, the main bottlenecks for Boltzina are docking pose generation and trunk module processing. By adjusting these parameters according to compound library scale and usage, the balance between accuracy and speed could be further optimized. Overall, Boltzina combines practical

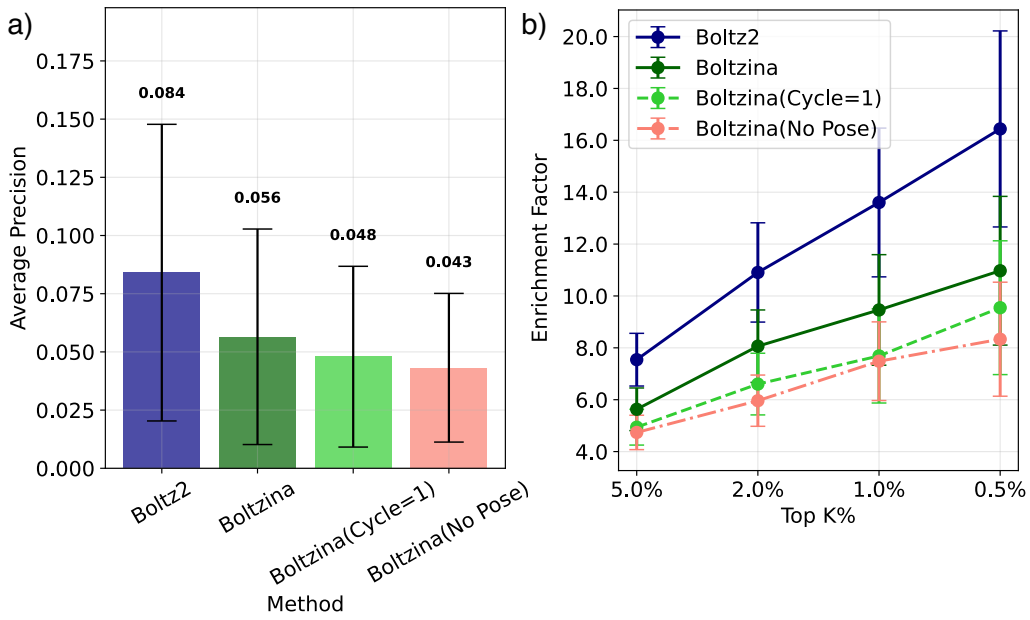

Figure 4: Comparison of Boltzina variants on the MF-PCBA test set. a) Mean Average Precision across assays. b) Average Enrichment Factor at top $K\%$.

computational efficiency with screening performance, making it a promising method especially for large-scale screening where time efficiency matters more than the need for high accuracy.

3.2 Pose Selection Strategies

Figure 5b shows the average ROC-AUC for each pose selection strategy. Top-5 Average strategy showed the highest ROC-AUC (0.778), improving performance compared to Top-3 Average and Best Pose Only strategies (0.746). Additionally, Top-5 Best Score strategy did not outperform the averaging strategy. If Boltz-2's binding likelihood could accurately identify the quality of individual poses, selecting the highest score should be superior; in practice, averaging was advantageous. This suggests that Boltz-2's binding likelihood may not sufficiently discriminate individual pose quality, and the improvement from averaging likely stems from reduced randomness. Such behavior is likely attributed to the fact that Boltz-2's binding likelihood is not trained to discriminate the quality of individual poses.

3.3 Two-Stage Screening Experiment

Finally, we conducted two-stage screening experiments combining rapid pre-screening with Boltzina and accurate binding prediction with Boltz-2. Figure 6 shows the relationship between estimated execution time and Average Precision for each method. Comparing Boltzina variants, Boltzina provided Pareto-optimal solutions in most cases. In particular, combining Boltzina (Cycle=1) with rescoring the top 5% by Boltz-2 outperformed Boltzina alone in accuracy per unit cost. Moreover, using the top $\sim$20% from Boltzina for two-stage screening achieved a mean AP exceeding 0.08 while being about three times faster than Boltz-2. These results show that two-stage strategies enable finer control of the trade-off between Boltzina and Boltz-2 according to application requirements. However, the optimal ratio depends on the prevalence of potential binders in the target library and thus should be chosen accordingly in practice.

4 Conclusion

We developed Boltzina, a pipeline for rapid and accurate virtual screening that predicts affinity using Boltz-2 with AutoDock Vina docking poses as input. Boltzina achieved a mean AP of 0.056, which

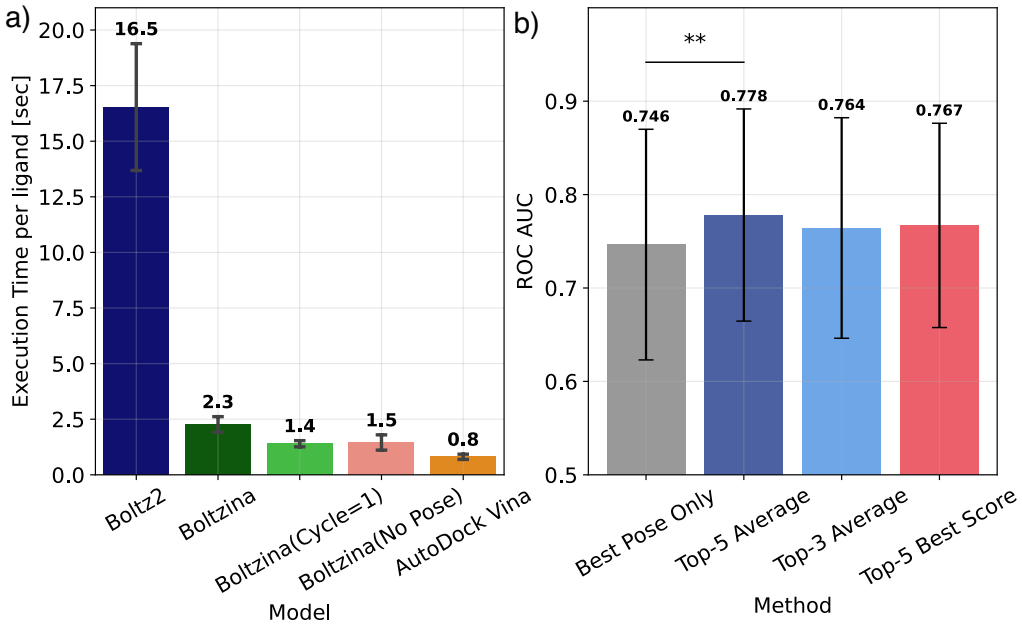

Figure 5: a) Execution time per ligand, computed from 1,000 ligands in the MF-PCBA test set using one GPU and 48 CPU cores. b) Average ROC-AUC for pose selection strategies using 1,000 ligands from the MF-PCBA test set. $**$ indicates a significant difference ($p < 0.01$) by the Wilcoxon signed-rank test.

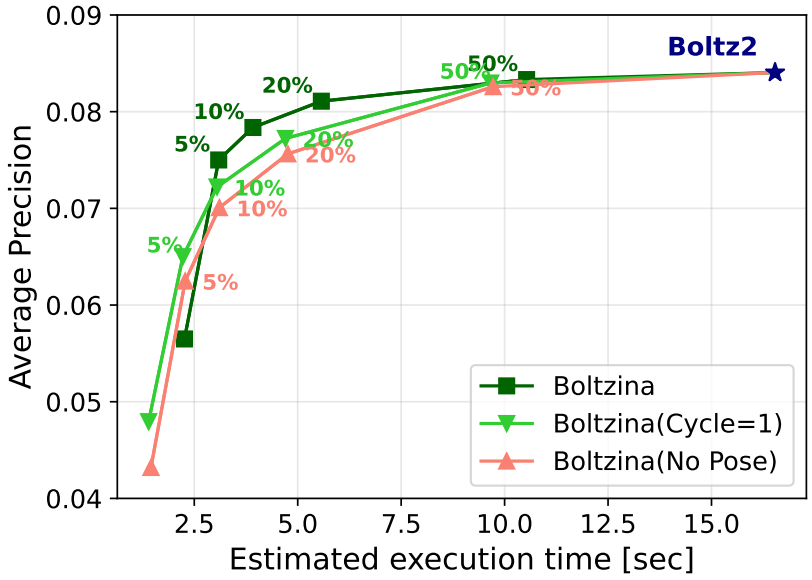

Figure 6: Two-stage screening: estimated execution time vs. mean Average Precision at different selection ratios. Estimated time was computed from the per-ligand measurements in Figure 5a.

was somewhat lower than Boltz-2's 0.084, but substantially outperformed conventional molecular docking methods Vina (mean AP 0.016) and GNINA (mean AP 0.012). Execution time achieved a 7.3× speedup over Boltz-2 in the standard setting, and up to 11.8× when recycling was reduced to a single iteration, enabling screening within realistic timeframes for library sizes that are difficult for Boltz-2 alone.

Experimental results indicate that intermolecular interaction representations generated by the trunk module largely govern binding prediction accuracy, allowing reasonable performance even without pose information; providing appropriate poses from AutoDock Vina further improves accuracy. Averaging multiple poses improved ROC-AUC and effectively reduced randomness associated with pose selection. Furthermore, two-stage screening that combines rapid screening with Boltzina and precise re-evaluation with Boltz-2 yielded Pareto-optimal solutions in both computational efficiency and accuracy.

Several limitations remain. We did not evaluate absolute affinity prediction or the physical validity of poses, and it remains unverified whether Boltzina can match Boltz-2 on these tasks. In addition, rigid docking with AutoDock Vina cannot account for protein flexibility, and pocket selection requires prior knowledge. Moreover, for molecules with a large number of heavy atoms that were excluded from this study, docking computational time increases dramatically [4]; for these corner cases, this represents a significant limitation as the proposed method cannot achieve computational speed improvements. Finally, applying the proposed method to ultra-large screening [22, 33] exceeding one billion compounds still presents challenges.

Overall, the proposed method bridges the gap between Boltz-2's high accuracy and the speed of conventional docking, substantially improving cost-effectiveness in virtual screening and contributing to more efficient drug discovery.

Acknowledgements

This study was partly supported by JSPS KAKENHI (JP23H04880, JP23H04887, JP24KJ1091), AMED BINDS (JP24ama121026), JST FOREST (JPMJFR216J), and JST ACT-X (JPMJAX25LB).

References

[1] Manasvi Saini, Nisha Mehra, Gaurav Kumar, Rohit Paul, and Béla Kovács. Molecular and structure-based drug design: From theory to practice. *Adv. Pharmacol.*, 103:121–138, 27 February 2025.

[2] E Lionta, G Spyrou, D Vassilatis, and Z Cournia. Structure-based Virtual Screening for drug discovery: Principles, applications and recent advances. *Curr. Top. Med. Chem.*, 14(16):1923–1938, 31 July 2014.

[3] Garrett M Morris, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30(16):2785–2791, 1 December 2009.

[4] Oleg Trott and Arthur J Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31(2):455–461, 30 January 2010.

[5] Ísak Valsson, Matthew T Warren, Charlotte M Deane, Aniket Magarkar, Garrett M Morris, and Philip C Biggin. Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. *Commun. Chem.*, 8(1):41, 8 February 2025.

[6] Azam Shirali, Vitalii Stebliankin, Ukesh Karki, Jimeng Shi, Prem Chapagain, and Giri Narasimhan. A comprehensive survey of scoring functions for protein docking models. *BMC Bioinformatics*, 26(1):25, 22 January 2025.

[7] Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. GNINA 1.0: molecular docking with deep learning. *J. Cheminform.*, 13(1):43, 9 June 2021.

[8] Hongjian Li, Kam-Heung Sze, Gang Lu, and Pedro J Ballester. Machine-learning scoring functions for structure-based virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 11(1), January 2021.

[9] Guy Durant, Fergus Boyles, Kristian Birchall, Brian Marsden, and Charlotte M Deane. Robustly interrogating machine learning-based scoring functions: what are they learning? *Bioinformatics*, 41(2):btaf040, 4 February 2025.

[10] Hongjian Li, Kwong-Sak Leung, Man-Hon Wong, and Pedro J Ballester. Improving AutoDock Vina using Random Forest: The growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Mol. Inform.*, 34(2-3):115–126, February 2015.

[11] Zechen Wang, Liangzhen Zheng, Yang Liu, Yuanyuan Qu, Yong-Qiang Li, Mingwen Zhao, Yuguang Mu, and Weifeng Li. OnionNet-2: A convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Front. Chem.*, 9:753002, 27 October 2021.

[12] Dejun Jiang, Chang-Yu Hsieh, Zhenxing Wu, Yu Kang, Jike Wang, Ercheng Wang, Ben Liao, Chao Shen, Lei Xu, Jian Wu, Dongsheng Cao, and Tingjun Hou. InteractionGraphNet: A novel and efficient deep graph representation learning framework for accurate protein-ligand interaction predictions. *J. Med. Chem.*, 64(24):18209–18232, 23 December 2021.

[13] Hui Zhu, Jincai Yang, and Niu Huang. Assessment of the generalization abilities of machine-learning scoring functions for structure-based virtual screening. *J. Chem. Inf. Model.*, 62(22):5485–5502, 28 November 2022.

[14] 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, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, page 2025.06.14.659707, 18 June 2025.

[15] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, Sebastian W Bodenstein, David A Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B Fuchs, Hannah Gladman, Rishub Jain, Yousuf A Khan, Caroline M R Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 8 June 2024.

[16] Zhiyi Wu, Gerhard Konig, Stefan Boresch, and Benjamin P Cossins. Optimizing absolute binding free energy calculations for production usage. *Journal of Chemical Theory and Computation*, 21(17):8330–8340, 26 August 2025.

[17] Gregory A Ross, Chao Lu, Guido Scarabelli, Steven K Albanese, Evelyne Houang, Robert Abel, Edward D Harder, and Lingle Wang. The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Commun. Chem.*, 6(1):222, 14 October 2023.

[18] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, page 2024.11.19.624167, 27 December 2024.

[19] David Buterez, Jon Paul Janet, Steven J Kiddle, and Pietro Liò. MF-PCBA: Multifidelity high-throughput screening benchmarks for drug discovery and machine learning. *J. Chem. Inf. Model.*, 63(9):2667–2678, 8 May 2023.

[20] Min Li, Zhangli Lu, Yifan Wu, and Yaohang Li. BACPI: a bi-directional attention neural network for compound-protein interaction and binding affinity prediction. *Bioinformatics*, 38(7):1995–2002, 28 March 2022.

[21] Mark McGann. FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.*, 51(3):578–596, 28 March 2011.

[22] Yuejiang Yu, Chun Cai, Jiayue Wang, Zonghua Bo, Zhengdan Zhu, and Hang Zheng. Uni-Dock: GPU-Accelerated Docking Enables Ultralarge Virtual Screening. *J. Chem. Theory Comput.*, 26 April 2023.

[23] Jiankun Lyu, John J Irwin, and Brian K Shoichet. Modeling the expansion of virtual screening libraries. *Nat. Chem. Biol.*, 19(6):712–718, 16 June 2023.

[24] João P G L M Rodrigues, João M C Teixeira, Mikaël Trellet, and Alexandre M J J Bonvin. pdb-tools: A swiss army knife for molecular structures. *F1000Res.*, 7:1961, 20 December 2018.

[25] Maxit suite. https://sw-tools.rcsb.org/apps/MAXIT/.

[26] Jonathan B Baell and Georgina A Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, 53(7):2719–2740, 8 April 2010.

[27] Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.*, 61(8):3891–3898, 23 August 2021.

[28] Keisuke Uchikawa, Kairi Furui, and Masahito Ohue. Leveraging AlphaFold2 structural space exploration for generating drug target structures in structure-based virtual screening. *Biochemistry and Biophysics Reports*, 43:102110, 1 September 2025.

[29] Andrew T McNutt, Yanjing Li, Rocco Meli, Rishal Aggarwal, and David Ryan Koes. GNINA 1.3: the next increment in molecular docking with deep learning. *J. Cheminform.*, 17(1):28, 2 March 2025.

[30] David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.*, 53(8):1893–1904, 26 August 2013.

[31] Ian Dunn, Somayeh Pirhadi, Yao Wang, Smmrithi Ravindran, Carter Concepcion, and David Ryan Koes. CACHE challenge #1: Docking with GNINA is all you need. *J. Chem. Inf. Model.*, 64(24):9388–9396, 23 December 2024.

[32] Ashutosh Kumar and Kam YJ Zhang. Hierarchical virtual screening approaches in small molecule drug discovery. *Methods*, 71:26–37, 2015.

[33] Gabriel Corrêa Veríssimo, Rafaela Salgado Ferreira, and Vinícius Gonçalves Maltarollo. Ultralarge virtual screening: Definition, recent advances, and challenges in drug design. *Mol. Inform.*, 44(1):e202400305, January 2025.

Appendix

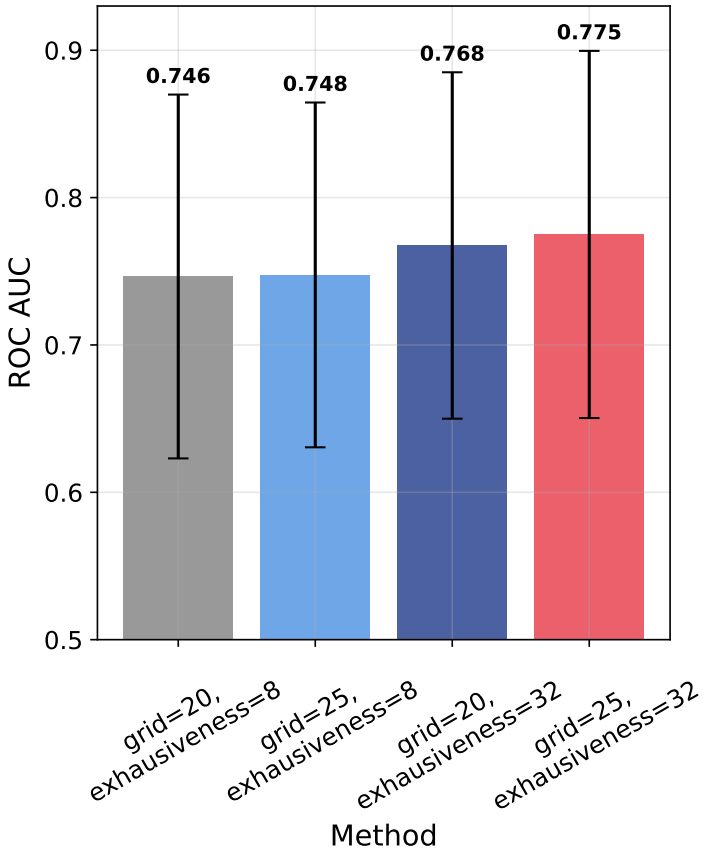

Figure 7: Average ROC-AUC for each AutoDock Vina parameter using 1,000 ligands from the MF-PCBA test set. Results for two parameters: grid size and exhaustiveness. No significant differences were found among all results by pairwise comparisons using the Wilcoxon signed-rank test.