On the rise of AI technologies for virtual screening

Authors: Marco Cecchini[1,*] and Hryhory Sinenka[1]

Affiliations:

[1]Institut de Chimie de Strasbourg, UMR7177, CNRS, Université de Strasbourg, 67081, Strasbourg Cedex, France

Summary:

AI foundational models for predicting protein-ligand interactions and binding affinities have started to emerge. We challenged Boltz-2 on a difficult dataset constructed on ten ultra-large virtual screening hit lists of pharmacologically relevant targets with *in vitro* binding assays. We show that Boltz-2 is the best classifier, with a success rate twice that of any other rescoring strategy. Ligand classifications by Boltz-2 are straightforward, accurate, efficient and robust, opening to million-compound accurate rankings on commodity resources.

Virtual screening (VS) remains a cornerstone of modern drug discovery, enabling the efficient identification of bioactive compounds from vast chemical libraries.[1,2] Classical virtual-screening approaches, such as structure-based molecular docking, pharmacophore modeling, and ligand-based similarity searches, have long been used to prioritize potential hits for experimental validation. However, these traditional methods often struggle to capture receptor flexibility, solvent effects, and entropic contributions, thus limiting their predictive power for binding affinity estimation. To address these limitations, rigorous physics-based approaches, mostly based on alchemical free energy simulations, have been developed to predict absolute (ABFE) or relative (RBFE) binding free energies with high accuracy.[3–5] Despite their proven reliability, these approaches remain computationally demanding and are impractical for large-scale screening.

Recently, the accuracy of non-rigorous methods for rescoring docking results, i.e., those that can predict the free energy of binding within ca. 15 minutes on commodity resources, was evaluated on a dataset of ten ultra-large virtual screening hit lists of pharmacologically relevant targets (ULVSH) that underwent *in vitro* binding assays.[6] Therein, ligand discrimination between actives and inactives is particularly challenging because: i. virtual hits have been selected in docking campaigns with exceptionally high hit rates (i.e., there are no obvious non-binders); ii. active and inactive compounds occupy similar portions of the chemical space and cannot be distinguished by chemical similarity; iii. neither actives nor inactives share specific interaction patterns; and iv. the dataset is formed of mostly transmembrane proteins, whose modeling and simulation is notoriously challenging.[5] Interestingly, the work of Sindt *et al.*

[*] Corresponding author: mcecchini@unistra.fr

showed that none of the rescoring functions tested on this dataset could reliably distinguish true binders from experimentally inactive compounds.[6]

Recent advances in artificial intelligence (AI) have begun to transform the drug discovery landscape by offering scalable and accurate alternatives for modeling protein-ligand interactions. AI foundational models for co-folding, such as AlphaFold3[7] and Boltz-2,[8] can now reliably predict the 3D structure of the protein-ligand complex starting from 2D representations of the protein (FASTA sequence) and the ligand (SMILES format). Moreover, the introduction of an affinity-prediction module in Boltz-2 was shown to achieve accuracy comparable to rigorous physics-based methods, while operating approximately three orders of magnitude faster.[8] And, using a subset of protein targets from the FEP+ benchmark, Boltz-2 affinity predictions were shown to be on par with ABFE calculations started from X-ray structures.[9] These advances provide evidence of a paradigm shift in drug discovery, which opens to scalable and accurate virtual screening workflows alternative to docking with non-rigorous rescoring.

We challenged the Boltz-2 predictive power on the ULVSH dataset. This dataset contains 943 virtual hits (427 true positives, 516 false positives) that were identified in ten ultra-large screening campaigns targeting seven GPCRs, one kinase, one membrane receptor, and one transporter.[6] Starting from the MOL2 coordinates of the ligands available from Ref.[10] and the FASTA sequence of the proteins from the PDB, we developed a simple automatic procedure to prepare input files, run Boltz-2 and collect binding affinity predictions for all ligands of the ULVSH library. By running on a single GPU card of the latest generation (RTX 4500 Ada Generation), Boltz-2 affinity predictions per ligand took approximately 100 sec, and the ULVSH library could be processed in < 1 day. Following Sindt *et al.*, the classification ability of Boltz-2 (i.e., to distinguish true binders from inactive compounds) was evaluated by measuring ROC-AUC values per target and comparing with the performances of eight popular rescoring strategies, including empirical rescoring, machine learning, single-point energy calculations based on polarizable force fields or semi-empirical quantum mechanics, and end-point free-energy simulations. To construct ROC curves for Boltz-2, the ligands were ranked based on `affinity_probability_binary` values as recommended for hit identification (see the Boltz-2 documentation[11]); we note that ranking ligands based on the `affinity_pred_value` yields similar results (**Table S1**).

The results are given in **Table 1**. Remarkably, Boltz-2 predictions outperform all the rescoring strategies tested by Sindt *et al.* [6] with an average ROC-AUC of 0.70, which corresponds to a significant increase in classification accuracy. Moreover, if one considers a classification successful when the ROC-AUC is greater than 0.7, which is arbitrary but reasonable as

highlighted by Sindt *et al.*,[6] Boltz-2 largely outperforms all the other strategies with a success rate that is more than twice that of the best method in the pool (**Table 1**). Last, when repeated on a different machine with a different graphics card (i.e., RTX A4500), the Boltz-2 calculations yielded very similar results with absolute ROC-AUC deviations per target below 0.04 units (**Table S2**) and even with slightly better performances (i.e., 83 sec/ligand on RTX A4500 vs 96 sec/ligand on RTX 4500 Ada). Nonetheless, the Boltz-2 predictions were poor in two out of ten cases (i.e., CNR1 and MTR1A); note that without these two targets, the average Boltz-2 ROC-AUC would increase to 0.77.

Table 1. *ROC-AUC results from docking, empirical scoring (HYDE), machine learning (Δvina, Gnina), end-point free energy methods (MMPBSA, MMGBSA), polarizable force-field (GFN-NN) and semi-empirical quantum mechanics (PM6) calculations, and Boltz-2 are given. The average ROC-AUC, along with the standard error over the whole ULVSH dataset, is given at the bottom. Classifications were considered successful when the ROC-AUC was greater than 0.7 (green highlighting). The corresponding ROC curves are shown in* ***Figure S1****.*

Target	Docking	HYDE	Δvina	Gnina	MMPBSA	MMGBSA	GFN-FF	PM6	Boltz-2
ADA2B	0,63	0,61	**0,78**	**0,76**	0,56	0,56	0,58	0,62	0,65
CASR	0,64	0,58	**0,71**	**0,75**	**0,80**	**0,75**	0,58	**0,79**	**0,70**
CNR1	0,52	0,55	0,67	0,68	0,50	0,51	0,57	0,55	0,40
CBR2	0,60	0,57	0,53	0,53	0,54	0,52	0,54	0,50	**0,79**
DRD3	0,52	0,60	**0,79**	**0,71**	**0,78**	**0,74**	**0,85**	**0,90**	**0,81**
DRD4	0,54	0,62	0,53	0,52	0,50	0,54	0,50	0,57	**0,74**
MTR1A	0,63	0,56	0,67	0,64	0,64	0,64	0,63	0,63	0,45
ROCK1	0,68	0,57	0,65	0,62	0,64	0,64	0,51	0,60	**0,84**
SC6A4	0,67	0,61	0,64	0,58	0,53	0,60	0,57	0,68	**0,86**
SGMR2	0,55	0,66	0,58	0,59	0,51	0,55	0,61	0,58	**0,77**
avg	**0,60**	**0,59**	**0,66**	**0,64**	**0,60**	**0,60**	**0,59**	**0,64**	**0,70**
std err	0,02	0,01	0,03	0,03	0,04	0,03	0,03	0,04	0,05
hit rate	**0/10**	**0/10**	**3/10**	**3/10**	**2/10**	**2/10**	**1/10**	**2/10**	**7/10**

To identify potential limitations in our Boltz-2 affinity predictions and possibly correct for them, we re-examined the two unsuccessful targets (CNR1 and MTR1A) by testing several hypotheses. First, since Boltz-2 affinity predictions come with some variance in independent calculations, our analysis was repeated 20 times per ligand, and ROC-AUC values were computed using the mean, the best, or the worst prediction to benchmark Boltz-2 in an ensemble framework. Second, because high-resolution structures in complex with ligands were available for all targets (see PDB ID in Ref.[6]), affinity predictions were recomputed by providing Boltz-2 with experimental coordinates as templates (template conditioning). Third, since most of the targets in ULVSH are multi-domain and the ligand-binding site is always located within one of them, the calculations were repeated by providing single-domain FASTA sequences (corresponding to the 3D coordinates used for docking) to remove potential sources of noise. Fourth, to increase the confidence of the predictions, the analysis was

repeated by increasing the quality of the inference parameters (i.e., `recycling_steps=6`, `sampling_steps=1000`, `sampling_steps_affinity=1000`, and `diffusion_samples_affinity=10)` at the price of increased computation. Fifth, under the hypothesis that the true solution could have been missed by Boltz-2 because of premature convergence, the calculations were repeated by modulating the stochastic character of the search via the `step_scale` and `diffusion_samples` parameters, with the former controlling the diversity of the proposed solutions, while the latter setting the number of independent samples used for prediction. The results in **Table S3** show that no methodological variant produced a performance improvement relative to the default. One way of affecting the stochastic character of the search (`step_scale`=5 and `diffusion_samples`=20) seems to improve the classification ability for MTR1A (ROC-AUC of 0.66), but it has little to no effect on CNR1. We conclude that Boltz-2 affinity predictions on the ULVSH dataset with default parameters are robust, efficient, and close to optimum. Last, to make sure that the ULVSH compounds were not mis-docked during co-folding, the 3D models predicted by Boltz-2 were compared with the PDB coordinates of existing co-crystal structures (see PDB ID in Ref.[6]). As shown in **Figure S2**, all compounds were correctly docked in the crystallographic ligand-binding site.

The results obtained on the ULVSH dataset indicate that Boltz-2 is the best ligand classifier among a pool of popular rescoring strategies of different nature and complexity. In our hands, Boltz-2 provided affinity predictions at a rate of ca. 1000 ligands per day on a single GPU card. Considering a mid-sized computer cluster equipped with 20 GPUs of the latest generation, this makes 20K ligands per day, or equivalently half a million ligands per month. Albeit these numbers are still far from the ultra-large chemical space that is currently accessible *on demand*,[12,13] Boltz-2 outperforms typical docking calculations in terms of accuracy, even when these latter are rescored via computationally more intensive non-rigorous free energy methods. Based on the accuracy, robustness, and output rate, Boltz-2 and similar appear to effectively fill the gap between ultra-large library screenings and lead optimization in a novel and elegant fashion. In addition, AI-based strategies do not require knowledge of the 3D structure of the protein or the binding mode of the ligand and can be applied to targets that are not structurally characterized. Whether these methods would be able to provide accurate affinity results for proteins that were not used for training remains an open question. Last, in our hands, Boltz-2 failed in two out of ten classification problems (i.e., CNR1 and MTR1A). Although the reasons for it are unclear, it provides evidence that AI methods for virtual screening may fail. Therefore, AI-based strategies need to be further developed, improved, accelerated, and possibly complemented with accurate, albeit more expensive numerical

strategies like ABFE in new and intelligent ways. We note that developments in this direction have already started to emerge.[9]

By redefining the practical limits of virtual screening, the integration of AI-driven technologies like Boltz-2 is revolutionizing the field of drug discovery. For many relevant targets, these methods open to the unprecedentedly accurate exploration of vast chemical spaces with limited resources and technical requirements. Whether this will lead to a more efficient development of pharmaceuticals is currently unclear, but it is exciting to see how fast the field is evolving and growing.

Supporting Information: Boltz-2 classification performances using different predictors, runtime parameters or graphics cards; ROC curves for all targets; and structural models of the protein-ligand complexes predicted by Boltz-2 (PDF).

Data availability. Raw data to reproduce the ROC curves based on Boltz-2 predictions for all targets of the ULVSH dataset, to analyse the Boltz-2 classifications based on different sets of runtime parameters, and to explore the Boltz-2 structural models for all protein-ligand complexes of the ULVSH dataset are available at Zenodo [https://zenodo.org/uploads/17804167].

Acknowledgments

This work received funding from the European Union's Horizon 2020 Research and Innovation Program under Marie Sklodowska-Curie Grant Agreement 956314 [ALLODD].

References

(1) Zhu, H.; Zhang, Y.; Li, W.; Huang, N. A Comprehensive Survey of Prospective Structure-Based Virtual Screening for Early Drug Discovery in the Past Fifteen Years. *Int. J. Mol. Sci.* **2022**, *23* (24), 15961. https://doi.org/10.3390/ijms232415961.
(2) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229. https://doi.org/10.1038/s41586-019-0917-9.
(3) Fu, H.; Zhou, Y.; Jing, X.; Shao, X.; Cai, W. Meta-Analysis Reveals That Absolute Binding Free-Energy Calculations Approach Chemical Accuracy. *J. Med. Chem.* **2022**, *65* (19), 12970–12978. https://doi.org/10.1021/acs.jmedchem.2c00796.
(4) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57* (12), 2911–2937. https://doi.org/10.1021/acs.jcim.7b00564.
(5) Papadourakis, M.; Sinenka, H.; Matricon, P.; Hénin, J.; Brannigan, G.; Pérez-Benito, L.; Pande, V.; Van Vlijmen, H.; De Graaf, C.; Deflorian, F.; Tresadern, G.; Cecchini, M.; Cournia, Z. Alchemical Free Energy Calculations on Membrane-Associated Proteins. *J.*

Chem. Theory Comput. **2023**, *19* (21), 7437–7458. https://doi.org/10.1021/acs.jctc.3c00365.
(6) Sindt, F.; Bret, G.; Rognan, D. On the Difficulty to Rescore Hits from Ultralarge Docking Screens. *J. Chem. Inf. Model.* **2025**, *65* (11), 5553–5566. https://doi.org/10.1021/acs.jcim.5c00730.
(7) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.
(8) Passaro, S.; Corso, G.; Wohlwend, J.; Reveiz, M.; Thaler, S.; Somnath, V. R.; Getz, N.; Portnoi, T.; Roy, J.; Stark, H.; Kwabi-Addo, D.; Beaini, D.; Jaakkola, T.; Barzilay, R. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. June 18, 2025. https://doi.org/10.1101/2025.06.14.659707.
(9) Thaler, S.; Wu, Z.; Glass, W. G.; Bradshaw, R. T.; Tossou, P.; Wood, G. P. F. Boltz-ABFE: Free Energy Perturbation without Crystal Structures. arXiv 2025. https://doi.org/10.48550/ARXIV.2508.19385.
(10) The Ultralarge Virtual Screening Hit (ULVSH) Data Set, 2025. http://bioinfo-pharma.u-strasbg.fr/labwebsite/downloads/ULVSH.tgz.
(11) Boltz-2, 2025. https://github.com/jwohlwend/boltz.
(12) Liu, F.; Wu, C.-G.; Tu, C.-L.; Glenn, I.; Meyerowitz, J.; Kaplan, A. L.; Lyu, J.; Cheng, Z.; Tarkhanova, O. O.; Moroz, Y. S.; Irwin, J. J.; Chang, W.; Shoichet, B. K.; Skiniotis, G. Large Library Docking Identifies Positive Allosteric Modulators of the Calcium-Sensing Receptor. *Science* **2024**, *385* (6715), eado1868. https://doi.org/10.1126/science.ado1868.
(13) Corrêa Veríssimo, G.; Salgado Ferreira, R.; Gonçalves Maltarollo, V. Ultra-Large Virtual Screening: Definition, Recent Advances, and Challenges in Drug Design. *Mol. Inform.* **2025**, *44* (1), e202400305. https://doi.org/10.1002/minf.202400305.