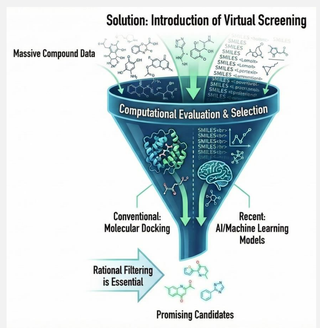
Benchmarking the Co-folding Model Boltz-2 and Generative Molecular Design for Affinity and Novelty in Histone Methyltransferase Inhibitors

Kazuyoshi Ikeda[1], Yugo Shimizu[1], Hitomi Yuki[2], Tomohiro Sato[2], Teruki Honma[1,2]

1 HPC- and AI-driven Drug Development Platform Division, RIKEN Center for Computational Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan.
2 RIKEN Center for Biosystems Dynamics Research, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan.

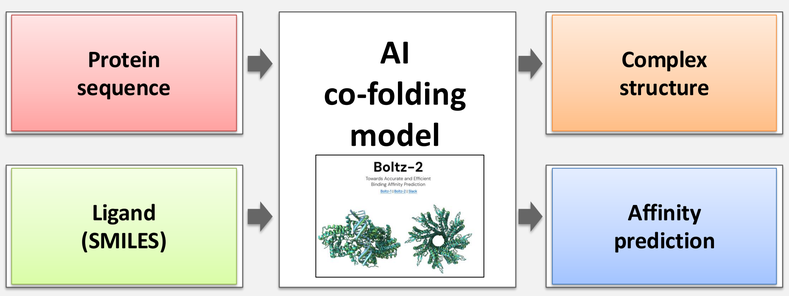
Background

- In early drug discovery, the **chemical space is vast**, but only a limited number of compounds can be experimentally tested.
- Therefore, virtual screening has become widely adopted. In addition to **traditional molecular docking**, the use of **AI-based prediction** models is becoming increasingly common.
- Highly accurate methods are required because poor prediction performance leads to wasted experimental resources on false positives and the risk of overlooking promising candidates (false negatives).



AI Co-folding Model & Objectives

- Recent advances in **co-folding models** (e.g., *AlphaFold 3*) enable direct prediction of **protein–ligand complex structures** from protein sequences and ligand structures.
- ***Boltz-2*** predicts both **complex structures and binding affinities**, with reported accuracy comparable to free-energy perturbation (FEP) methods.
- Here, we benchmark an AI-based co-folding method on an in-house **histone methyltransferase inhibitor dataset**, evaluating both throughput and affinity prediction accuracy across implementations (including **NVIDIA NIM**).



https://github.com/jwohlwend/boltz

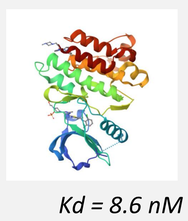
- We further integrate Boltz-2 with **generative molecular design** to enable scalable, high-precision virtual screening and computational lead optimization on HPC.

ex. Target：*BCR-ABL1 Kinase*

MENFQKVEKIGEGTYGVVYK…
Protein: Amino Acid Sequence

Nc1nccc(-c2cccnc2)n12)cc1Nc1nc….
Ligand: SMILES

Boltz-2 Prediction
Using GPU



Kd = 8.6 nM

Dataset & Parameter Settings

Dataset

- 728 small-molecule inhibitors with experimentally measured IC_{50} values
- Target: a histone methylation enzyme (histone methyltransferase)
- Used as a real-world benchmark for affinity prediction and ranking

Boltz-2 implementations compared

① Original version of Boltz-2 v2.2.0 (https://github.com/jwohlwend/boltz)
② NVIDIA NIM version Boltz-2 Release 1.3.0 (https://docs.nvidia.com/nim/bionemo/boltz2/latest/)
- NVIDIA GPU optimized implementation

- Comparison of ① and ② was performed using parameter settings that matched the default settings of the original

Parameter Settings

setting	Sampling_steps	Without_potentials	Sampling_steps_affinity	Diffusion_samples_affinity
①② Original version default	200	True	200	5
③ NIM default	50	False	200	5
④ Lightweight	50	False	50	1
NIM setting range	10-1000	True/False	10-1000	1-10
overview	Noise Reduction Iterations The more, the greater the variety and quality	Whether to use	Number of iterations to use for	Number of diffusion processes used in

- Boltz-2 calculations were performed using an NVIDIA A100 GPU.
- MSA was pre-generated using the original Boltz-2 algorithm.
- Sequences were input as monomers.

Benchmarking Protocol

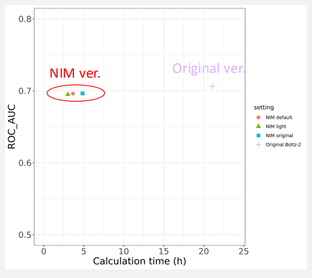
- **Compile experimental data:** Collect in-house inhibitory activity data (IC_{50}) for histone methyltransferase inhibitors and convert to pIC_{50}.
- **Run Boltz-2 inference:** Predict protein–ligand complex structures and binding affinities.
- **Test multiple implementations/settings:** Evaluate multiple Boltz-2 configurations, including NVIDIA NIM-based deployments, to assess robustness and throughput under realistic HPC settings.
- **Quantify prediction performance:** Compute correlation metrics (e.g., Pearson and/or Spearman) between predicted affinities and experimental pIC_{50} values.
- **Compare with docking baselines:** Perform docking-based scoring and compare affinity correlations to assess improvement over conventional docking.

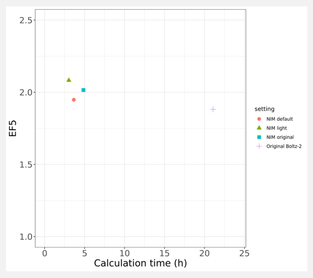
Results: Affinity Prediction

Discrimination accuracy

Labels for discrimination performance tests:
- $IC_{50} \leq 1$ μM (301 molecules) → Active
- $IC_{50} > 1$ μM (427 molecules) → Inactive

Computation time is per GPU or per CPU , does not include MSA calculation time

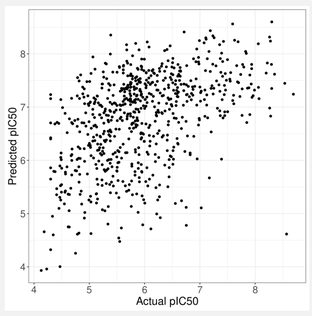




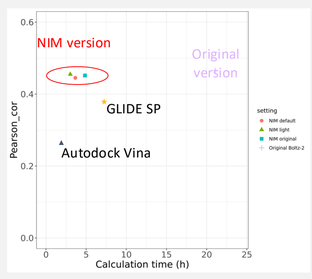
- NIM version for both discrimination and regression .
- There was almost no difference in prediction accuracy when changing parameter settings.

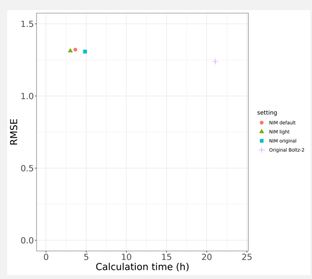
Regression accuracy

Perform regression using $pIC_{50} = -\log_{10}(IC_{50})$



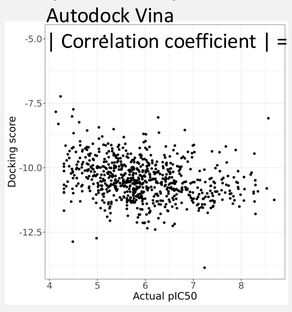
- Original Boltz-2
  - Pearson correlation coefficient = 0.466
- NIM lightweight setup is the most efficient as it is fast.



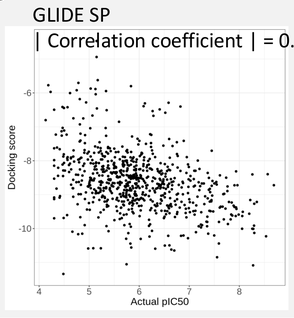


Comparison: Physics-based docking methods:

Autodock Vina



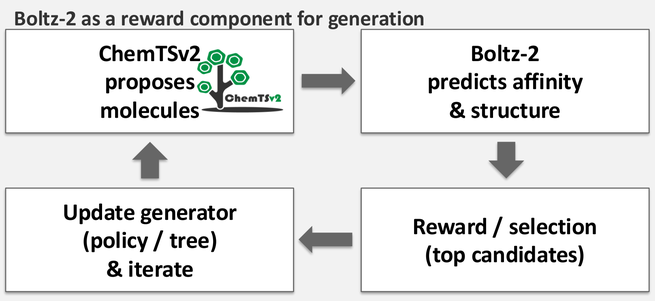
GLIDE SP



Boltz-2-guided Molecular Design (ChemTSv2)

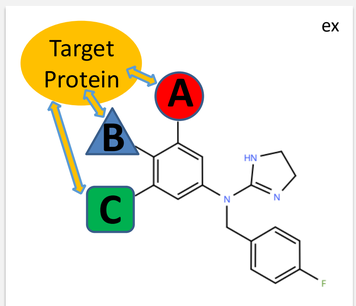
- Molecular generator: ChemTSv2 explores chemical space and proposes new structures.
- Reward function incorporates Boltz-2 predicted affinity (and novelty) for the target.
- Iterative optimization yields high-activity small-molecule candidates.
- Approach is readily scalable for large design campaigns on HPC.

Boltz-2 as a reward component for generation

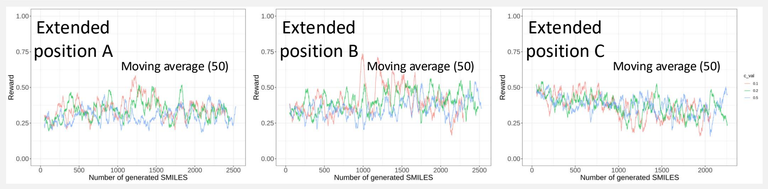


Results: ChemTS + Botlz score reward

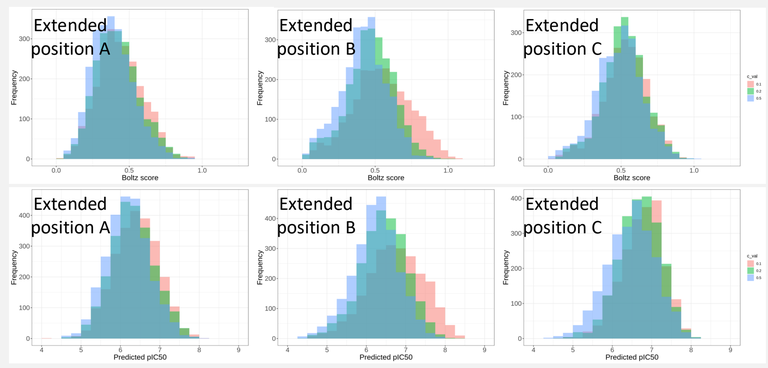
We used ChemTSv2 [2] (MCTS + RNN) for de novo molecular generation.
- Starting scaffold: an active compound (predicted pIC_{50} = 6.0, Boltz score = 0.36).
  - Grow the molecule from three carbon positions (A, B, C) by extending substituents.
- Activity reward: Boltz score mapped to 0–1 (≤0.2 → 0, ≥0.75 → 1, linear in between).
- MW reward: 0–1 score (MW ≤ 600 → 1, MW ≥ 610 → 0, linear in between).
  - Final reward: geometric mean of activity and MW rewards.
- RNN: trained on ChEMBL 220k compounds.
- C value: 0.1, 0.2, 0.5.
- Filters: remove radicals; apply PubChem rules; SA ≤ 3.5; ring size ≤ 7.



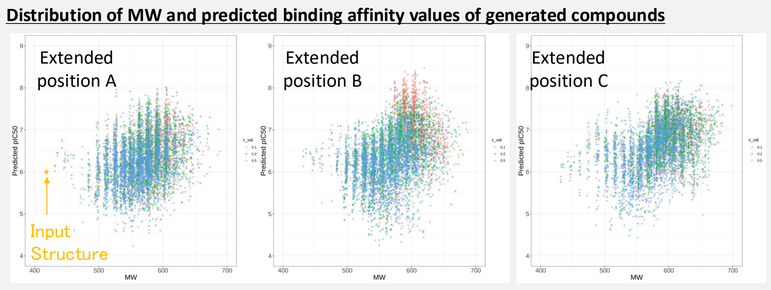
Current Reward Trends



Predicted value distribution of generated compounds



Distribution of MW and predicted binding affinity values of generated compounds



- In all structural generation, highly active compounds with predicted pIC_{50} values of 7 or higher were generated.
- In B, a compound with a large molecular weight was generated that was predicted to be highly active.
- In A and C, compounds with smaller molecular weights were generated, although their predicted activity values were not as high.

Conclusions

- Boltz-2 provides improved affinity prediction compared with docking on the histone methyltransferase dataset.
- Boltz-2 + ChemTSv2 enables efficient design of highly active, novel candidates.
- HPC-friendly workflows support large-scale virtual screening and design.
- Next: extend to additional targets/datasets, add synthesizability & ADMET constraints, and quantify uncertainty.

Acknowledgements & Contact

References :

[1] Passaro, S. *et al* . Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv* , doi:10.1101/2025.06.14.659707 (2025).
[2] Ishida, S., *et al* . ChemTSv2: Functional Molecular Design Using de Novo Molecule Generator. *WIREs Comput . Mol. Sci.* , e1680 (2023).

Acknowledgments :

- This study is part of AMED-BINDS project.
- This research was carried out using NVIDIA NIM.

Conflict of interest disclosure :

- There are no conflicts of interest to disclose in relation to this presentation.

Contact:

- kazuyoshi.ikeda@riken.jp