

Supporting Information

Revealing the Limits of Covalent Docking and Advancing Affinity Prediction with Covalent-Aware Multi-Task Learning

Jiayang Leng^{1,2#}, Zhixuan Huang^{1,2#}, Lei Zheng^{3,4#}, Yuhan Yang⁵, Huilin Huang⁶, Yongjun Gao⁷,
Lizhen Huo^{1,2}, Yang Li⁸, Zhaoxi Sun^{2*} and John Z. H. Zhang^{2,3,4*}

¹*Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, Guangdong, China*

²*Faculty of Synthetic Biology, Shenzhen University of Advanced Technology, Shenzhen 518107, China*

³*Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning and NYU-ECNU Center for Computational Chemistry, NYU-Shanghai, 1555 Century Avenue, Pudong New Area, Shanghai 200062, China*

⁴*Department of Chemistry, New York University, NY, NY 10003, USA*

⁵*Guangdong Medical University, No. 1 Xincheng Avenue, Songshan Lake High-tech Industrial Development Zone, Dongguan 523808, China*

⁶*School of Life Sciences, South China Normal University, Guangzhou 510631, China*

⁷*Shenzhen MSU-BIT University, Shenzhen 518172, China*

⁸*School of Physics and Technology, University of Jinan, Jinan, Shandong Province, China*

*To whom correspondence should be addressed:

John Z. H. Zhang john.zhang@nyu.edu

Zhaoxi Sun z.sun@suat-sz.edu.cn

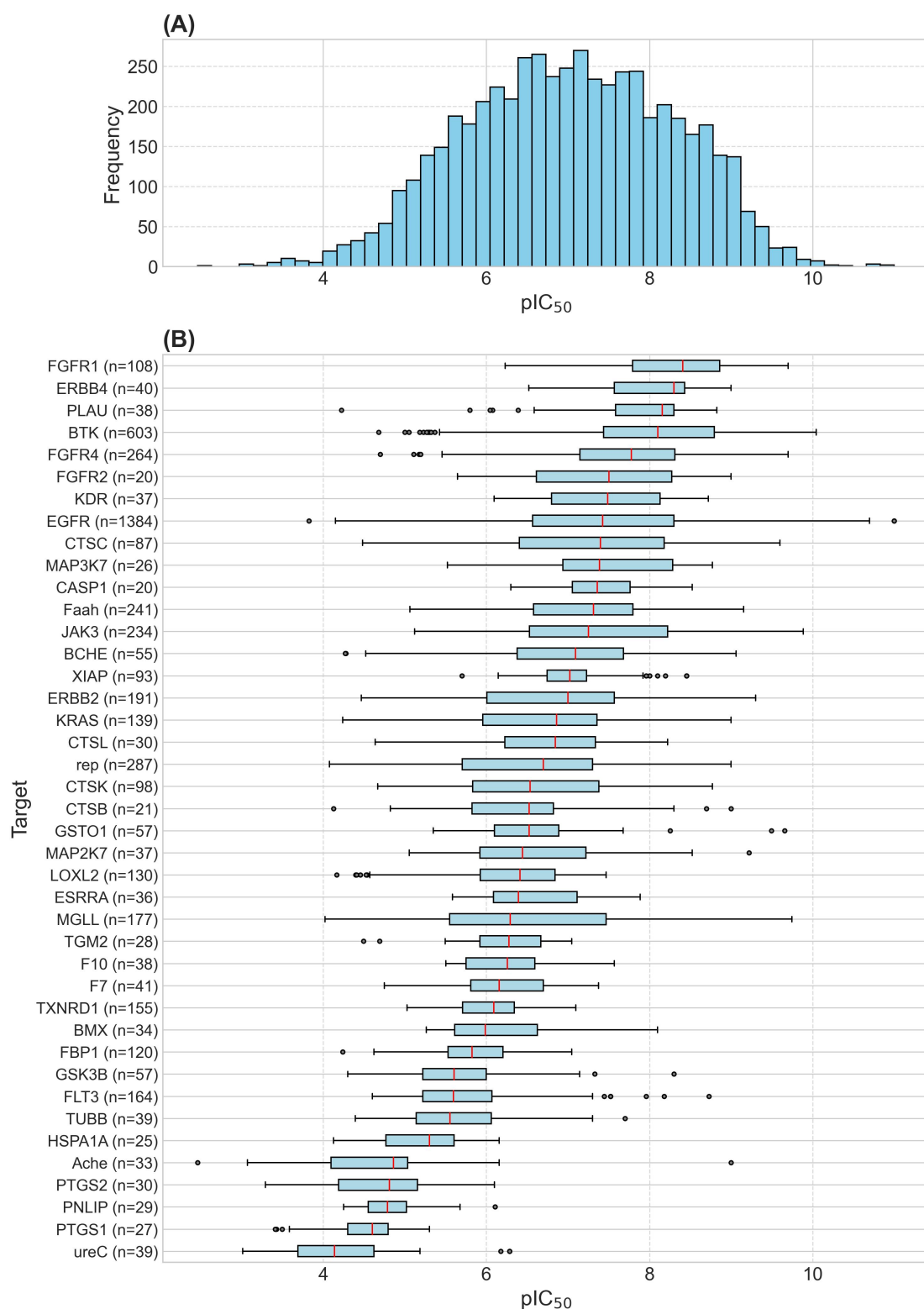


Figure S1. Distribution of experimental pIC₅₀ values in the curated covalent bioactivity dataset. (A) Histogram of experimental pIC₅₀ values for all ligand–target pairs in the dataset. (B) Box plots of pIC₅₀ values stratified by target. For each target, the horizontal box shows the interquartile range (25th–75th percentiles), the red line indicates the median, whiskers extend to 1.5× IQR, and dots denote outliers. The number of ligand–target pairs available for each target is given in parentheses.

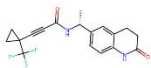
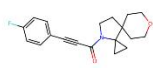
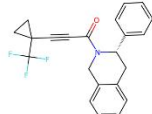
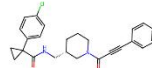
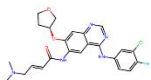
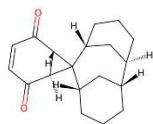
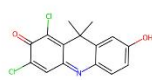
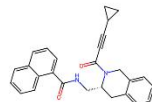
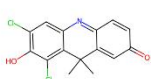
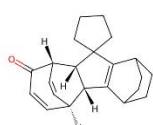
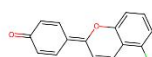
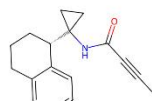
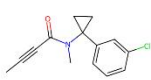
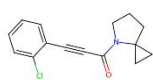
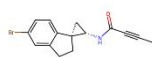
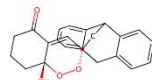
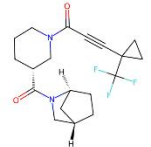
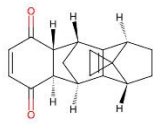
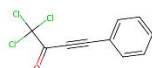
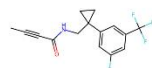
ZC1796043	ZC1143935	ZC2844564	ZC2379924
			
pIC ₅₀ : 8.94 QED: 0.81 Fsp ³ : 0.44 LogS: -4.01	pIC ₅₀ : 8.93 QED: 0.69 Fsp ³ : 0.53 LogS: -3.70	pIC ₅₀ : 8.89 QED: 0.67 Fsp ³ : 0.32 LogS: -5.33	pIC ₅₀ : 8.82 QED: 0.77 Fsp ³ : 0.36 LogS: -4.86
nan	ZC2070670	ZC2079301	ZC2380272
			
pIC ₅₀ : 8.68 QED: 0.46 Fsp ³ : 0.29 LogS: -5.44	pIC ₅₀ : 8.67 QED: 0.68 Fsp ³ : 0.79 LogS: -3.84	pIC ₅₀ : 8.65 QED: 0.74 Fsp ³ : 0.20 LogS: -4.46	pIC ₅₀ : 8.65 QED: 0.67 Fsp ³ : 0.26 LogS: -5.04
ZC2351847	ZC2825982	ZC2068819	ZC1114731
			
pIC ₅₀ : 8.62 QED: 0.73 Fsp ³ : 0.20 LogS: -4.57	pIC ₅₀ : 8.62 QED: 0.58 Fsp ³ : 0.68 LogS: -4.80	pIC ₅₀ : 8.58 QED: 0.71 Fsp ³ : 0.00 LogS: -4.01	pIC ₅₀ : 8.58 QED: 0.81 Fsp ³ : 0.47 LogS: -3.26
ZC1068907	ZC1115336	ZC1142304	ZC2156357
			
pIC ₅₀ : 8.57 QED: 0.74 Fsp ³ : 0.36 LogS: -3.28	pIC ₅₀ : 8.56 QED: 0.66 Fsp ³ : 0.40 LogS: -3.49	pIC ₅₀ : 8.55 QED: 0.79 Fsp ³ : 0.40 LogS: -3.51	pIC ₅₀ : 8.55 QED: 0.73 Fsp ³ : 0.35 LogS: -4.76
ZC1366749	ZC1118505	ZC1077148	ZC1651297
			
pIC ₅₀ : 8.53 QED: 0.67 Fsp ³ : 0.79 LogS: -3.68	pIC ₅₀ : 8.53 QED: 0.63 Fsp ³ : 0.67 LogS: -3.19	pIC ₅₀ : 8.53 QED: 0.51 Fsp ³ : 0.10 LogS: -3.57	pIC ₅₀ : 8.52 QED: 0.67 Fsp ³ : 0.40 LogS: -3.61

Figure S2. Additional EGFR covalent hits prioritized by CovMTL-DTA. Representative EGFR covalent inhibitors selected from the top-ranked compounds in the Michael-acceptor library. Approved drugs (e.g., Afatinib, Osimertinib) and their close analogs were excluded to highlight novel hits. For each compound, the 2D chemical structure, CovMTL-DTA predicted pIC₅₀, and key ADMET descriptors (QED, Fsp³, predicted solubility LogS) are shown.

Table S1. Curated dataset of covalent protein–ligand complexes used for docking and DTA modeling (separate CSV file).

Table S2. Per-complex RMSD values for GNINA on the covalent docking benchmark (separate CSV file).

Table S3. Per-complex RMSD values for CovDock on the covalent docking benchmark (separate CSV file).

Table S4. Per-complex RMSD values for AutoDock4 on the covalent docking benchmark (separate CSV file).

Table S5. Per-complex RMSD values for Boltz on the covalent docking benchmark (separate CSV file).

Table S6. Per-complex RMSD values for the intersection dataset on the covalent docking benchmark (separate CSV file).

Table S7. EGFR virtual screening results for the Michael-acceptor library (separate CSV file)