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 Yang[5], Huilin Huang[6], Yongjun Gao[7], Lizhen Huo[1,2], Yang Li[8], 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

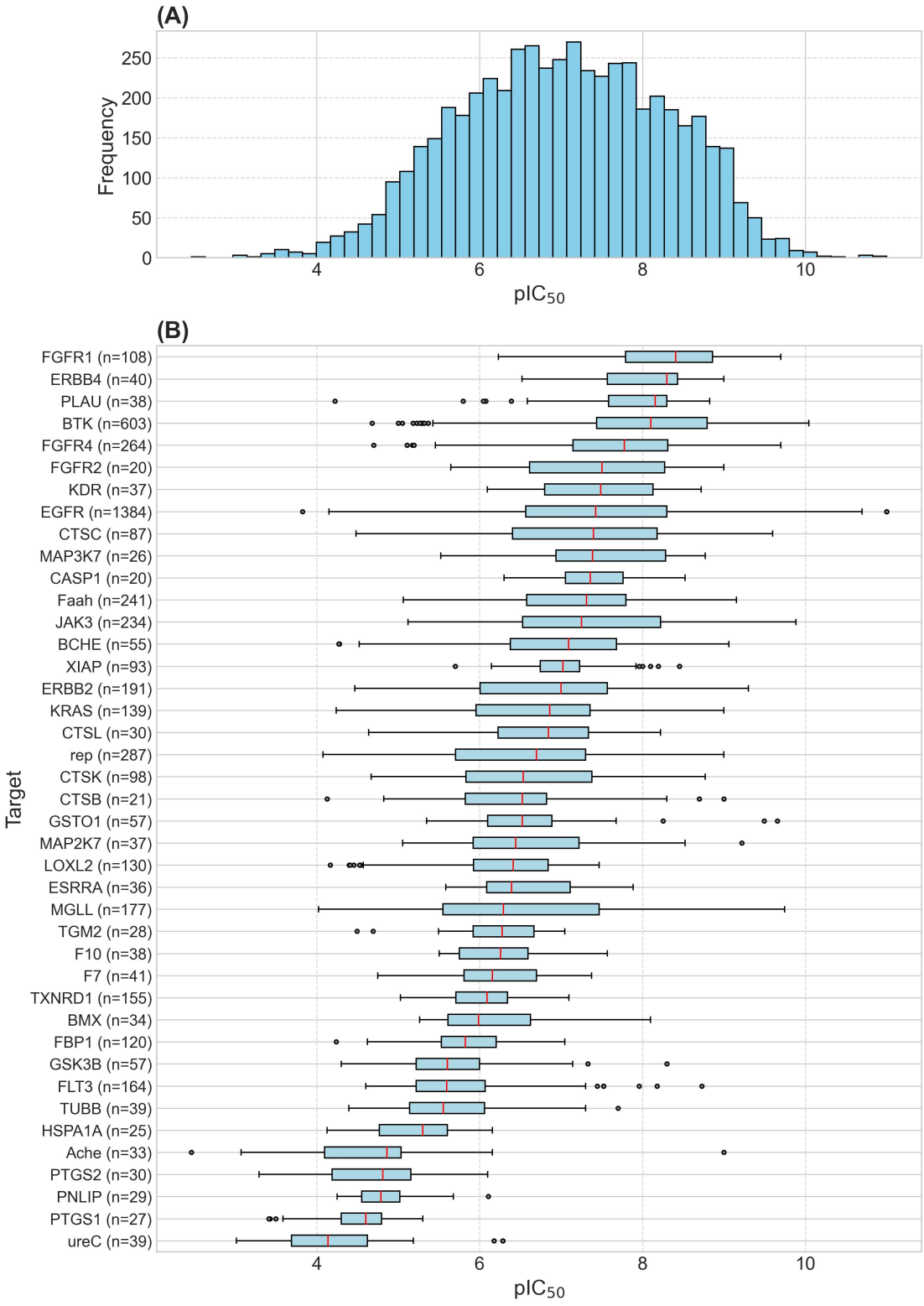

Figure S1. Distribution of experimental pIC_{50} values in the curated covalent bioactivity dataset. (A) Histogram of experimental pIC_{50} values for all ligand–target pairs in the dataset. (B) Box plots of pIC_{50} values stratified by target. For each target, the horizontal box shows the interquartile 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.

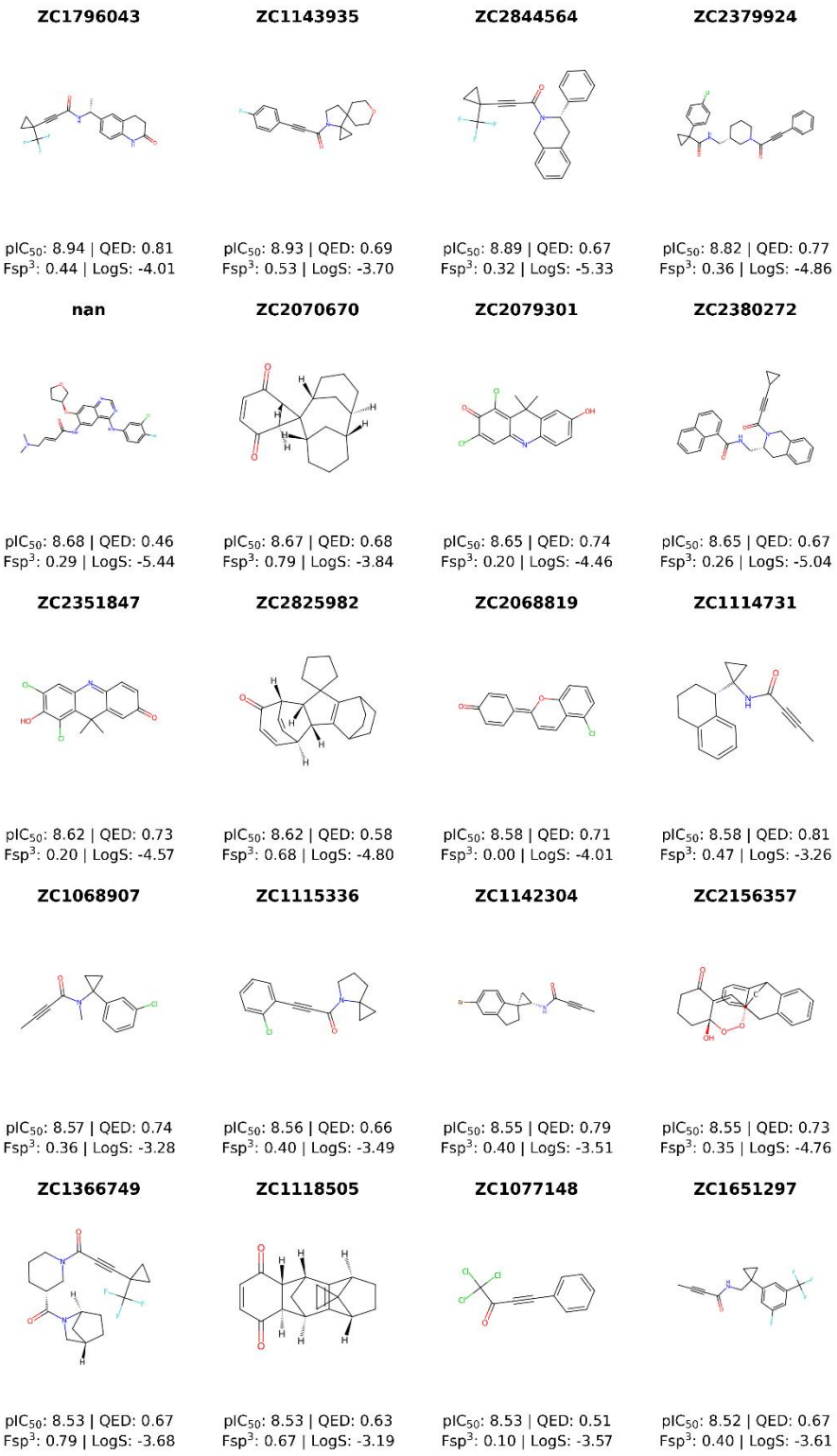

Figure S2. Additional EGFR covalent hits prioritized by CovMTL-DTA. Representative EGFR covalent inhibitors selected from the top-ranked compounds in the Michael-acceptor library. Approved drugs (e.g., Afatinib, Osimertinib) and their close analogs were excluded to highlight novel hits. For each compound, the 2D chemical structure, CovMTL-DTA predicted pIC_{50}, and key ADMET descriptors (QED, Fsp^3, 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)