Supporting Information for:

Identification of a Water-Coordinating HER2 Inhibitor by Virtual Screening Using Similarity-Based Scoring

Jiaye Guo,[†,#] Stephen Collins,[†,#] W. Todd Miller,[*,‡,∥] and Robert C. Rizzo[*,§,∥,⊥]

[†]Graduate Program in Biochemistry and Structural Biology, [‡]Department of Physiology and Biophysics, [§]Department of Applied Mathematics & Statistics, [∥]Institute of Chemical Biology & Drug Discovery, and [⊥]Laufer Center for Physical & Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States

*Corresponding authors e-mail: todd.miller@stonybrook.edu, rizzorc@gmail.com

[#] These authors contributed equally to this work

Introduction to Supporting Information

The information provided here supplements several sections in the main text. To supplement *Methods Part 1: HER2 Virtual Screen,* we present protocols for: (1) kinase domain homology modeling, and (2) expression and purification of HER2 kinase. To supplement *Methods Part 2: Protocols to Coordinate or Displace Bridging Waters*, we present protocols for: (1) solvated footprint reference preparation, and (2) geometric and energetic stability of water coordination and displacement. To supplement *Results Part 2: Development of Virtual Screen Protocols to Incorporate Bridging Waters*, we present the solvated footprint references derived for HIVPR and PARP1 systems used during the development of the COOR (coordination) and DISP (displacement) virtual screening protocols.

Methods Part 1: HER2 Virtual Screen

Kinase Domain Homology Model. To construct a homology model of HER2 in the fully active form for virtual screening, we used the previously reported x-ray structure of EGFR complexed with erlotinib (PDB 1M17)[1] as a template. The model was constructed by manually mutating residues to the desired HER2 sequence based on the EGFR template using the program Chimera[2] to maximize side chain overlap between new and old side chain positions. To make use of different knowledge-based scoring functions for compound prioritization (discussed further below), erlotinib from the EGFR template was prepared as a reference along with two coordinating water molecules.

To refine the constructed homology model complex, the AMBER16[3] suite of programs was employed to assemble and prepare the complex HER2 complex,

protonate the system as necessary and assign molecular modeling force field parameters, and equilibrate the structure through controlled energy minimization and short molecular dynamics simulations to relax the system prior to the virtual screen. The receptor and ligand were parameterized using the ff14SB[4] and GAFF[5] force fields respectively which as solvated in a periodic box (10 $Å^3$) of TIP3P[6] water molecules. Ligand partial atomic charges were based on the AM1BCC method.[7, 8] The refinement protocol consisted of five sequential steps that included energy minimization (max. 10,000 cycles), MD heating (100 ps), MD density equilibration (500 ps), MD equilibration (4 x 200 ps), and MD production (100 ns). Minimizations were executed on CPUs and all other steps were executed on GPUs using the AMBER Particle Mesh Ewald molecular dynamics (PMEMD) program.[9] Production runs were performed in the NPT ensemble (T = 298.15 K, P = 0.987 atm) using Langevin dynamics,[10] during which heavy atoms in the protein and the ligand were weakly restrained (0.1 kcal/mol-$Å^2$ restraint weight). The time step for the production run was 2 fs, SHAKE[11] was used to constrain bond stretching, and coordinates were written every 100 steps. Following the MD simulations of HER2 we retained 10,000 evenly spaced frames and selected a specific frame (frame #4859) for docking that (1) contained both bridging waters with the intended hydrogen bonding pattern, (2) had good Cα overlap with the centroid of MD trajectory, and (3) had no steric clashes.

To evaluate the overall quality of the final HER2 homology model (frame #4859), with that of the original EGFR template (PDB 1M17), we used the PROCHECK[12] software package to compute common stereochemical features. Overall, there was nothing unusual about the model compared to the template as shown in Table S1.

Specifically, the model had the majority (99.6%) of residues in allowed Ramachandran plot regions, favorable Morris classification scores (1, 1, 2),[13] a reasonable G-factor (-0.07), and all bond lengths (100%), most bond angles (95.2%), and most planar groups (94.9%) were within stereochemical limits.

Table S1. Stereochemical features of the HER2 homology model vs the EGFR template.

		HER2 Homology Model	**EGFR Template (**PDB 1M17)
Ramachandran Plot	Most favored[a]	89.0%	82.7%
	Additional allowed[b]	10.2%	16.5%
	Generously allowed[c]	0.4%	0.4%
	Disallowed[d]	0.4%	0.4%
Residue Stereochemistry	Morris classification[e]	1, 1, 2	1, 1, 2
Overall Properties	G-factors[f]	-0.07	0.23
	Main chain bond lengths[g]	100.0%	99.9%
	Main chain bond angles[h]	95.2%	97.9%
	Planar groups[i]	94.9%	98.0%

[a]Percent of residues in the most favored regions.
[b]Percent of residues in the additional allowed regions.
[c]Percent of residues in the generously allowed regions.
[d]Percent of residues in the disallowed regions.
[e]Overall assessment (on a scale of 1-4, 1 = best, 4 = worst) using Morris *et al.*[13] classification scheme.
[f]G-factor indicates the overall normality of the stereochemistry (larger = better).
[g]Percent of bond lengths within stereochemical limits.
[h]Percent of bond angles within stereochemical limits.
[i]Percent of planar groups within stereochemical limits.
Feature definitions taken from the PROCHECK operating manual (www.ebi.ac.uk/thornton-srv/software/PROCHECK).

Expression and Purification of HER2 Kinase Domain. The cDNA encoding the HER2 cytoplasmic domain (residues 703-1029) was cloned into the Xba1 and Xho1 restriction enzyme sites of the pFastbac HTb vector (Invitrogen). Recombinant baculovirus was generated using the Invitrogen Bac-to-Bac system. To produce protein, *Spodoptera frugiperda* (Sf9) cells in 1-liter spinner flasks were infected with baculovirus and harvested after 3 days by centrifugation. To purify HER2, cell pellets were suspended in lysis buffer (50 mM Tris pH 8.5, 100 mM NaCl, 5 mM 2-mercaptoethanol, 1mM phenylmethylsulfonyl fluoride, 1% NP-40, 5 µg/mL aprotinin, 5 µg/mL leupeptin).

Cells were lysed in a French pressure cell at 800 psi. Lysates were clarified by centrifugation at 10,000 x g for 10 minutes (4°C) then syringe filtered (0.8 μm). The clarified lysate was added to 2 mL of Ni-NTA agarose (Qiagen), and loaded onto a column. The resin was washed with 10 column volumes of a wash buffer A containing 20 mM Tris (pH 8.5), 10% glycerol, 500 mM NaCl, 10 mM imidazole and 5 mM 2-mercaptoethanol. The resin was subsequently washed with 10 column volumes of a wash buffer B containing 20 mM Tris (pH 8.5), 10% glycerol and 5 mM 2-mercaptoethanol, and 10 column volumes of wash buffer A. The protein was eluted in 1 mL fractions by a step gradient using elution buffers containing 20 mM Tris (pH 8.5), 10% glycerol, 5 mM 2-mercaptoethanol, and imidazole (at concentrations of 25, 50, 75, and 100 mM). The resulting fractions were analyzed by SDS-PAGE on 10% acrylamide gels and visualized by Coomassie Blue staining. Peak fractions were pooled and concentrated using an Amicon nitrogen concentrator (10 kDa NMWL membrane, Millipore). Protein concentrations were determined with the Bio-Rad dye binding assay.

Methods Part 2: Protocols to Coordinate or Displace Bridging Waters

Solvated Footprint Reference Preparation. Molecular models for HIVPR (1HPX[14], 2.00 Å) and PARP1 (1EFY[15], 2.20 Å) were relaxed using the same general setup and refinement procedures described above for HER2. Minor differences include a somewhat shorter MD production run (20 ns), with coordinates being written every 100 steps, to identify a suitable frame to construct the two references. For each MD frame, the five closest waters from the geometric center of the ligand were retained using CPPTRAJ.[16] The total interaction energy (ES + VDW) between each water

molecule and the rest of the complex (protein + ligand) was calculated and the single water with the most favorable interaction was retained. Then, the 1000 most favorable waters (one from each frame) were clustered into different water "sites" using the average linkage algorithm in CPPTRAJ with a distance (between oxygen atoms) cutoff of 2 Å.

For each water site that contained the desired bridging water, an individual MD frame was then identified to define a "specific reference" structure for the COOR or DISP protocols respectively. The procedure employed four distinct steps: (1) The similarity between the ES footprint (FPS_{ES})[17] of each water and the mean footprint of all waters in a site were calculated to identify one water (corresponding to one frame) with the highest overall similarity using DOCK6.8. (2) The identified frame was examined in terms of total interaction energy between the bridging water and other species (ligand + receptor) to verify it was not an outlier using box plots.[18] (3) The backbone RMSD between the candidate frame and the structural centroid over all 1000 frames (computed using CPPTRAJ clustering using the average linkage algorithm and a Cα-Cα distance cutoff of 1 Å) was examined to make sure it was less than 2 Å. (4) The ligand-receptor footprint was computed to confirm that the overall interactions in the identified frame were favorable. All candidate MD frames were rank-ordered by descending FPS_{ES} (step 1) and the first frame that met all four criteria listed above was used to generate the molecular references. Structures of the three species (receptor, ligand, bridging water) were then extracted from the relevant MD frame in MOL2 format minimized individually in Cartesian space using DOCK6.8, and assembled into the final COOR and DISP references as outlined in Figure S1.

Geometric and Energetic Stability of Water Coordination and Displacement. For each system (HIVPR, PARP1), we selected four representative compounds based on their categorization as coordinating or displacing (eq 1 and 2 main text) and subject to visualization in the binding site. For each compound, 20 ns MD simulations were then performed in triplicate with different random seeds to ascertain if the docked ligands would maintain their predicted coordination or displacement. The simulations employed a relatively weak protein backbone restraint (weight = 0.1 kcal/mol-Å^2) which reduced large protein backbone conformational changes however side chains were free to move.

For the COOR studies, positive controls were the cognate ligands KNI-272 (HIVPR, PDB 1HPX[14]) and NU1098 (PARP1, PDB 1EFY[15]) which coordinate bridging waters. For the DISP studies, positive controls employed ligands DMP323 (HIVPR, PDB 1QBS[23]) and 4AN (PARP1, PDB 2PAX[45]) which were previously reported as being rationally designed to displace bridging waters. As the coordination and displacement positive controls originated from different crystallographic structures of the same proteins, to eliminate potential noise arising from using different protein coordinates, ligand DMP323 from 1QBS was docked into 1HPX and ligand 4AN from 2PAX was docked into 1EFY followed by energy minimization of each complex in Cartesian space which yielded small RMSDs (DOCK6 Hungarian method[19]) of 0.83 Å, and 0.68 Å respectively.

Following the MD simulations, each trajectory was post-processed and the five most populated water sites were identified using the protocols described above. If water coordination simulations remain stable, we expected to see highly populated

water site(s) with similar molecular footprints to that of the COOR reference. In contrast, if water displacement simulations remain stable, we expected to see direct long-lived interactions between the ligand with residues in the binding site previously engaged by water, and similar molecular footprints to the DISP reference.

Results Part 2: Development of Virtual Screen Protocols to Incorporate Bridging Waters

Solvated Footprint References. As shown in Figure 3 (main text), for the COOR and DISP protocols, separate molecular references were required for similarity-based rescoring of the docked molecules. Here, for HIVPR, we selected MD frame #699 (Figure S1a) from a 20 ns simulation (see Methods for frame selection criteria) where the top populated water site matches the well-known "flap water"[20] labeled WAT301 in PDB 1HPX.[14] Figures S1b, c show the two solvated footprints derived from the ligand alone (COOR reference, magenta) or the ligand + water (DISP reference, blue) respectively. The footprints highlight the extensive VDW interactions made between the ligands and receptor versus the more sparse but specific ES patterns. In the COOR ES footprint in Figure S1b where the bridging water is included as part of the receptor, the peak labeled WAT (-1.99 kcal/mol) represents the ES interaction between ligand KNI-272 and the bridging water. In the DISP ES footprint in Figure S1c, where the bridging water is included as part of the ligand, the appearance of two new peaks (compare Figure S1c vs S1b) at positions Ile50 and Ile149 in the ES footprint (sum = 1.41 kcal/mol) reflect the geometry of the water bridge in Figure S1a.

For the PARP1 system, MD frame #3 (Figure S1d) was selected as the representative coordinates to generate COOR and DISP references as it contains a previously annotated bridging water labeled WAT52 in PDB 1EFY.[15] Here, the water mediates hydrogen bonding between a carboxylate oxygen in Glu327 and the amine hydrogen in ligand NU1098, which can be seen in the peak labeled WAT (-1.74 kcal/mol) in the COOR ES footprint show in Figure S1e and the significant increase in size of the ES peak at position Glu327 (-4.76 kcal/mol) in the DISP ES footprint in Figure S1f.

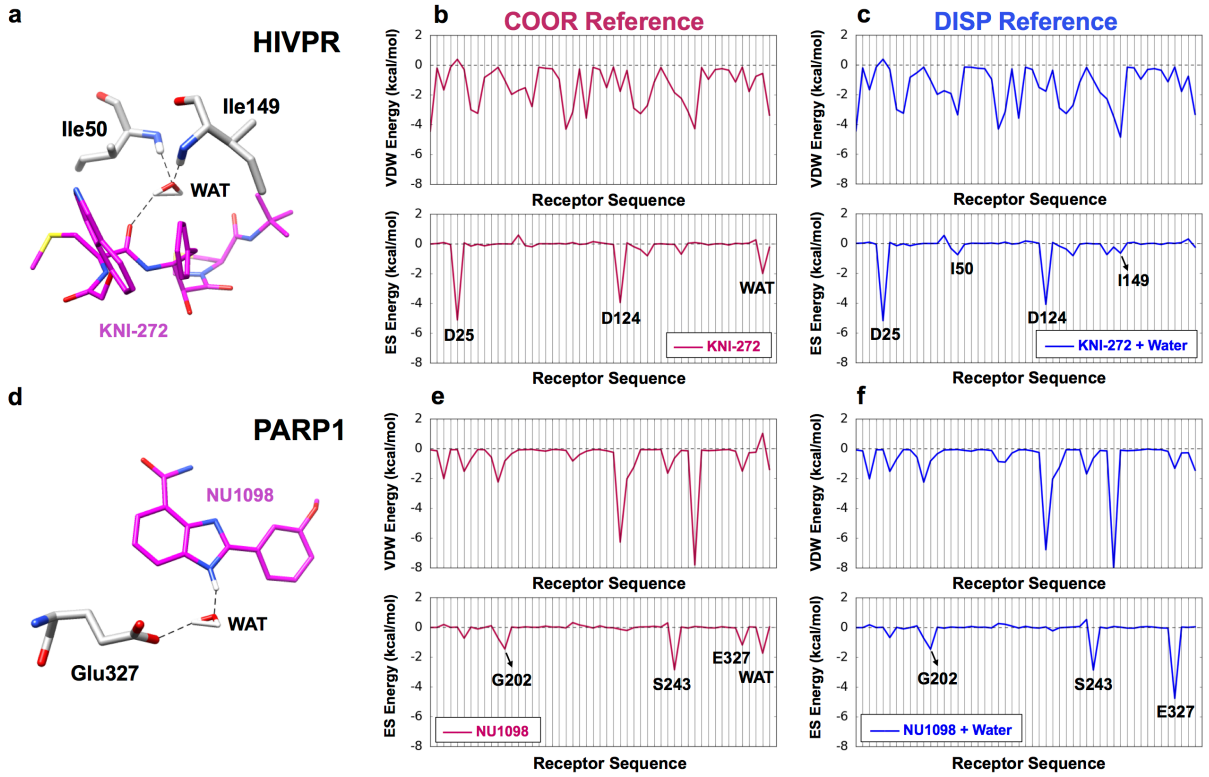

Figure S1. MD-based molecular references used to rescore docked molecules to HIVPR (**a-c,** MD frame #699) and PARP1 (**d-f,** MD frame #3). Panels **a** and **d** show water-mediated H-bonding as dashed lines with cognate ligands in magenta and key pocket residues in gray. Panels **b** and **e** show the solvated molecular footprints for the COOR references (cognate ligand alone) which include water as part of the receptor and panels **c** and **f** show the solvated molecular footprints for the DISP references (cognate ligand + water) which include water as part of the ligand.

References

[1] Stamos, J., Sliwkowski, M. X., and Eigenbrot, C. (2002) Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor, *J Biol Chem 277*, 46265-46272.

[2] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF chimera - A visualization system for exploratory research and analysis, *J Comput Chem 25*, 1605-1612.

[3] Case, D. A., Betz, R. M., Cerutti, D. S., T. E. Cheatham, I., Darden, T. A., Duke, R. E., Giese, T. J., Gohlke, H., Goetz, A. W., Homeyer, N., Izadi, S., Janowski, P., Kaus, J., Kovalenko, A., Lee, T. S., LeGrand, S., Li, P., Lin, C., Luchko, T., Luo, R., Madej, B., Mermelstein, D., Merz, K. M., Monard, G., Nguyen, H., Nguyen, H. T., Omelyan, I., Onufriev, A., Roe, D. R., Roitberg, A., Sagui, C., Simmerling, C. L., Botello-Smith, W. M., Swails, J., Walker, R. C., Wang, J., Wolf, R. M., Wu, X., Xiao, L., and Kollman, P. A. (2016) AMBER 2016, University of California, San Francisco.

[4] Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB, *J Chem Theory Comput 11*, 3696-3713.

[5] Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004) Development and testing of a general amber force field, *J Comput Chem 25*, 1157-1174.

[6] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of Simple Potential Functions for Simulating Liquid Water, *J Chem Phys 79*, 926-935.

[7] Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. (2000) Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method, *J Comput Chem 21*, 132-146.

[8] Jakalian, A., Jack, D. B., and Bayly, C. I. (2002) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation, *J Comput Chem 23*, 1623-1641.

[9] Darden, T., York, D., and Pedersen, L. (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems, *J Chem Phys 98*, 10089-10092.

[10] Feller, S. E., Zhang, Y. H., Pastor, R. W., and Brooks, B. R. (1995) Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method, *J Chem Phys 103*, 4613-4621.

[11] Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes, *J Comput Phys 23*, 327-341.

[12] Laskowski, R. A., Macarthur, M. W., Moss, D. S., and Thornton, J. M. (1993) Procheck - a Program to Check the Stereochemical Quality of Protein Structures, *J Appl Crystallogr 26*, 283-291.

[13] Morris, A. L., Macarthur, M. W., Hutchinson, E. G., and Thornton, J. M. (1992) Stereochemical Quality of Protein-Structure Coordinates, *Proteins 12*, 345-364.

[14] Baldwin, E. T., Bhat, T. N., Gulnik, S., Liu, B., Topol, I. A., Kiso, Y., Mimoto, T., Mitsuya, H., and Erickson, J. W. (1995) Structure of HIV-1 protease with KNI-272, a tight-binding transition-state analog containing allophenylnorstatine, *Structure 3*, 581-590.

[15] White, A. W., Almassy, R., Calvert, A. H., Curtin, N. J., Griffin, R. J., Hostomsky, Z., Maegley, K., Newell, D. R., Srinivasan, S., and Golding, B. T. (2000) Resistance-modifying agents. 9. Synthesis and biological properties of benzimidazole inhibitors of the DNA repair enzyme poly(ADP-ribose) polymerase, *J Med Chem 43*, 4084-4097.

[16] Roe, D. R., and Cheatham, T. E. (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data, *J Chem Theory Comput 9*, 3084-3095.

[17] Balius, T. E., Mukherjee, S., and Rizzo, R. C. (2011) Implementation and Evaluation of a Docking-Rescoring Method Using Molecular Footprint Comparisons, *J Comput Chem 32*, 2273-2289.

[18] Krzywinski, M., and Altman, N. (2014) Visualizing samples with box plots, *Nat Methods 11*, 119-120.

[19] Allen, W. J., and Rizzo, R. C. (2014) Implementation of the Hungarian algorithm to account for ligand symmetry and similarity in structure-based design, *J Chem Inf Model 54*, 518-529.

[20] Lam, P. Y. S., Jadhav, P. K., Eyermann, C. J., Hodge, C. N., Ru, Y., Bacheler, L. T., Meek, J. L., Otto, M. J., Rayner, M. M., Wong, Y. N., Chang, C. H., Weber, P. C., Jackson, D. A., Sharpe, T. R., and Ericksonviitanen, S. (1994) Rational Design of Potent, Bioavailable, Nonpeptide Cyclic Ureas as Hiv Protease Inhibitors, *Science 263*, 380-384.