­­­­Modeling and System preparation

System preparation, modeling, and initial docking calculations were performed using the Schrödinger Suite molecular modeling package (version 2015-4), using default parameters unless otherwise noted. The protein structure was obtained from an in-house repository and prepared using the Protein Preparation Wizard1. In this step, force field atom types and bond orders were assigned, missing atoms were added, tautomer/ionization states were assigned, water orientations were sampled, and ionizable residues (Asn, Gln, and His residues) have their tautomers adjusted to optimize the hydrogen bond network. A constrained energy minimization was then performed. All crystallographiclly resolved water molecules were retained.

Potential binding sites were explored and characterized using the SiteMap2,3 tool. Ligands with experimental activity and known inactives were docked into putative binding sites using Glide SP4,5 to evaluate enrichment of known actives. Best docking scores were for the ‘Ro’ series for the ‘Gossypol’ binding site described by Lan *et al*.6 compared to other putative pockets. The values however were only at a poor level around -4.6.

Since the receptor may not be in an optimal conformation to bind small molecule inhibitors, induced fit docking7 of ligand Ro 08-2750 was performed to this binding pocket. Induced fit docking results were validated with the metadynamics protocol described by Clark *et al*8. In these metadynamics simulations a biasing potential is applied to the ligand RMSD as collective variable. The resulting potential energy surface is evaluated towards how easy a ligand can move away from the initial binding mode. The underlying assumption is that a ligand pose which is closer to the real one has a higher energetic barrier to leave the pose than an incorrect pose. The pose ranked second using the induced fit docking score retrieved the best score from the metadynamics ranking protocol compared to the other induced fit docking poses. This receptor configuration was furthermore tested towards its suitability for a virtual screening by a Glide SP docking of known actives into this pocket. The docking scores using this receptor conformations were better (down to -6.2) compared to the initial protein conformation in the crystal structure. Furthermore, a WaterMap9,10 calculation was done for this receptor. The information about binding site hydration site energetics was used in conjunction with the virtual screening.

Virtual screening was performed with this modeled receptor conformation using Glide SP by docking the March 2016 collection of the eMolecules dataset. All ligand structures were prepared with LigPrep including a minimization with the OPLS3 force field11. One low energy ring conformation per compound was generated. Ionization states and tautomer forms were enumerated at pH 7.0 ± 2.0 with Epik12,13.

The top 5000 hits from virtual screening were filtered by applying filters according to Lipinski’s rule of five14, flagging REOS15 and PAINS

16.

In addition to the Glide SP docking score, a second ranking scheme was explored which uses energy information from WaterMap hydration sites. The docked ligands were ranked by a Pareto-ranking of docking scores and the number of WaterMap hydration sites identified with Δ*G* > 2 kcal/mol that overlap with the ligand pose. By doing this the exact energy of a hydration site, which can depend significantly on the investigated protein conformation, is not influencing the score too heavily. Instead only the fact if it is a high- or low-energy site is recognized. The underlying assumption is that tighter binding ligand molecules will likely displace also a higher number of high-energy water molecules from the binding site. At the same time protein-ligand interactions are not included in hydration site energetics but the docking score. In the Pareto ranking, compounds with a lower value for the docking score get better ranks as well as compounds with a higher number of displaced hydration sites. The resulting combined rank was used as a second score in addition to the SP score for the virtual screening results. The top 200 ranked hits from both lists were combined. Finally, a leader-follower clustering using dendridic fingerprints was performed using Canvas17,18 resulting in 243 unique cluster hits.

Induced Fit Docking of Ro-A6 and Ro-OH compounds

Induced Fit Docking (IFD) was performed against the receptor pose from the selected Ro 08-2750 pose, using Schödinger molecular modeling suite (version 2017-4). Poses for Ro-A6 and Ro-OH, the top and second scored poses respectively, were selected to most closely match the Ro 08-2750 pose.

Alchemical Free Energy Calculations

Absolute alchemical free energy calculations were carried out to validate the putative binding poses in a fully-flexible explicitly solvated system. The YANK GPU-accelerated free energy calculation code with the Amber family of forcefields was used for this purpose. Details follow.

*System preparation* *and modeling*. The top poses generated by induced fit docking, as described above, were selected as input protein and ligand poses. Because proteins and ligands were already prepared, they were simply run through the pdbfixer 1.4 command line tool with add-atoms and add-residues set to None to convert residue and atom names to be compatible with Amber tleap.

*Parameterization.* tleap (from the minimal conda-installable AmberTools 16 suite ambermini 16.16.0) was used to solvate the complex in a cubic box with a 12Å buffer of TIP3P water molecules around the protein. The system was parameterized using AMBER’s forcefield ff14sb19 and GAFF 1.820. Missing ligand parameters were determined using antechamber21. The ligand was assigned charges using the AM1-BCC 22,23 implementation in OpenEye (OEtoolkit 2017.6.1 through openmoltools 0.8.1).

*Minimization.* Minimization was performed using the implementation of the L-BFGS24 algorithm in OpenMM 7.1.125 with a tolerance of 1kJ/mol/nm.

*Production Simulation.* Production simulation was run using YANK 0.19.426 using OpenMMTools 0.13.4. In order to keep the ligand from diffusing away from the protein while in a weakly coupled state, it was confined to the binding site using a Harmonic restraint with an automatically-determined force constant (K = 0.33 kcal/mol/Å2). The restraint was centered on the following receptor residues using all-atom selection: 2, 4, 46, 76, 78, and 80. The ligand atoms were automatically determined. The calculation was performed using particle mesh Ewald (PME)27 electrostatics with default YANK settings with a real-space cutoff of 9Å. A long-range isotropic dispersion correction was applied to correct for truncation of the Lennard-Jones potential at 9Å. The system was automatically solvated with TIP3P28 solvent and four neutralizing Cl- ions, paramterized using the Joung and Cheaham paramters29. Production alchemical Hamiltonian exchange free energy calculations were carried out at 300 K and 1 atm using a Langevin integrator (VRORV splitting)30 with a 2 fs timestep, 5.0 ps-1 collision rate, and a molecular-scaling Monte Carlo barostat. Ro 08-2750 and Ro A6 were run for 10000 iterations (50 ns/replica) with 2500 timesteps (5 ps) per iteration, while Ro-OH was run for 15000 iterations (75 ns/replica) with 2500 timesteps (5 ps) per iteration. Complex configurations were stored for each replica once per iteration. Replica exchange steps were performed each iteration to mix replicas using the Gibbs sampling scheme described previously31. The alchemical pathway was automatically determined for each compound using the YANK autoprotocol protocol trailblazing feature.

*Absolute binding free energy estimates.* Absolute free energies (ΔG) of binding for each compound was estimated using MBAR32. Samples were reweighted to a cutoff of 16Å to correct the isotropic dispersion correction to a nonisotropic long-range dispersion. This correction is important to account for the heterogeneous density of protein. To remove the harmonic restraint bias, samples were reweighted to substitute a squared well restraint of radius 10Å

*Clustering analysis.* The fully interacting trajectory from YANK was extracted to a PDB file, discarding the following number of initial iterations, which came prior to equilibration33: 1500 for Ro 08-2750, 1600 for RoOH, and 1600 for RoA6. These trajectories were aligned in MDTraj34 using only protein backbone atoms. The small molecules were then sliced out and clustered on Cartesian coordinates using the MSMBuilder35 implementation of RegularSpatial clustering using a 1Å RMSD cutoff. For the most populated clusters for Ro 08-2750 and RoOH, cluster centers were selected and shown with 10 randomly sampled cluster members. RoA6 produced a large number of lowly populated clusters with highly heterogeneous binding poses, and were therefore not shown.

*Conformational heterogeneity analysis.* To investigate the conformational heterogeneity in the presence or absence of the ligand, the fully interacting thermodynamic state (corresponding to the holo protein bound to the ligand) and fully non-interacting state (corresponding to the apo protein free of ligand interactions) for all three ligands were extracted using a 4-frame skip, discarding the initial frames as above.

*Code availability:* All Schrödinger project files, YANK simulation inputs, and analysis scripts have been made publicly available (<https://github.com/choderalab/musashi>).

Author affiliations:

Daniel Cappel: Schrödinger GmbH, Q7, 23, 68161 Mannheim, Germany

Steven K. Albanese: 1) Gerstner Sloan Kettering Graduate School, Memorial Sloan Kettering Cancer Center, New York, NY 10065 2) Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York, USA,

Andrea Rizzi: 1) Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY, USA 2) Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Levi Naden: Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York, USA,

Grants to recognize:

JDC acknowledges support from the Sloan Kettering Institute and NIH grant P30 CA008748 and NIH grant R01 GM121505.

LN acknowledges support from Merck KGaA to support the development of open source tools for GPU-accelerated alchemical free energy calculations.

Conflict of interest statement:

JDC is a member of the Scientific Advisory Board for Schrödinger.

1. Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **27,** 221–234 (2013).

2. Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J Chem Inf Model* **49,** 377–389 (2009).

3. Halgren, T. New method for fast and accurate binding-site identification and analysis. *Chem Biol Drug Des* **69,** 146–148 (2007).

4. Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47,** 1739–1749 (2004).

5. Halgren, T. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **47,** 1750–1759 (2004).

6. Lan, L. *et al.* Natural product (-)-gossypol inhibits colon cancer cell growth by targeting RNA-binding protein Musashi-1. *Mol Oncol* **9,** 1406–1420 (2015).

7. Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A. & Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **49,** 534–553 (2006).

8. Clark, A. J. *et al.* Prediction of Protein-Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations. *J. Chem. Theory Comput.* **12,** 2990–2998 (2016).

9. Abel, R., Young, T., Farid, R., Berne, B. J. & Friesner, R. A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **130,** 2817–2831 (2008).

10. Young, T., Abel, R., Kim, B., Berne, B. J. & Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *PNAS* **104,** 808–813 (2007).

11. Harder, E. *et al.* OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **12,** 281–296 (2016).

12. Shelley, J. C. *et al.* Epik: a software program for pK( a ) prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* **21,** 681–691 (2007).

13. Greenwood, J. R., Calkins, D., Sullivan, A. P. & Shelley, J. C. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput. Aided Mol. Des.* **24,** 591–604 (2010).

14. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **46,** 3–26 (2001).

15. Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening—an overview. *Drug Discov. Today* **3,** 160–178 (1998).

16. Baell, J. B., chemistry, G. H. J. O. M.2010. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *ACS Publications*

17. Duan, J., Dixon, S. L., Lowrie, J. F. & Sherman, W. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* **29,** 157–170 (2010).

18. Sastry, M., Lowrie, J. F., Dixon, S. L. & Sherman, W. Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *J Chem Inf Model* **50,** 771–784 (2010).

19. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713 (2015).

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

21. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25,** 247–260 (2006).

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

23. Jakalian, A., Bush, B. L., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high‐quality atomic charges. AM1‐BCC model: I. Method. *J Comput Chem* **21,** 132–146 (2000).

24. Nocedal, J. American Mathematical Society. *Math. Comp.* **35,** 773–782 (1980).

25. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13,** e1005659 (2017).

26. Wang, K., Chodera, J. D., Yang, Y. & Shirts, M. R. Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics. *J. Comput. Aided Mol. Des.* **27,** 989–1007 (2013).

27. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N⋅log(N) method for Ewald sums in large systems. *J Chem Phys* **98,** 10089–10092 (1998).

28. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79,** 926–935 (1998).

29. Joung, I. S. & Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112,** 9020–9041 (2008).

30. Leimkuhler, B. & Matthews, C. Efficient molecular dynamics using geodesic integration and solvent-solute splitting. *Proc. Math. Phys. Eng. Sci.* **472,** 20160138 (2016).

31. Chodera, J. D. & Shirts, M. R. Replica exchange and expanded ensemble simulations as Gibbs sampling: simple improvements for enhanced mixing. *J Chem Phys* **135,** 194110 (2011).

32. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys* **129,** 124105 (2008).

33. Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.* **12,** 1799–1805 (2016).

34. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109,** 1528–1532 (2015).

35. Beauchamp, K. A. *et al.* MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **7,** 3412–3419 (2011).