

Derivatives of Salicylic Acid as Inhibitors of YopH in *Yersinia pestis*

Zunnan Huang¹, Yantao He², Xian Zhang²,
Andrea Gunawan³, Li Wu^{2,3}, Zhong-Yin
Zhang² and Chung F. Wong¹

¹Department of Chemistry and Biochemistry and Center for Nanoscience, University of Missouri-Saint Louis, One University Boulevard, St. Louis, MO 63121, USA

²Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 635 Barnhill Drive, Indianapolis, IN 46202, USA

³Chemical Genomics Core Facility, Indiana University School of Medicine, 635 Barnhill Drive, Indianapolis, IN 46202, USA

*Corresponding author: Chung F. Wong, wongch@umsl.edu

***Yersinia pestis* causes diseases ranging from gastrointestinal syndromes to bubonic plague and could be misused as a biological weapon. As its protein tyrosine phosphatase YopH has already been demonstrated as a potential drug target, we have developed two series of forty salicylic acid derivatives and found sixteen to have micromolar inhibitory activity. We designed these ligands to have two chemical moieties connected by a flexible hydrocarbon linker to target two pockets in the active site of the protein to achieve binding affinity and selectivity. One moiety possessed the salicylic acid core intending to target the phosphotyrosine-binding pocket. The other moiety contained different chemical fragments meant to target a nearby secondary pocket. The two series of compounds differed by having hydrocarbon linkers with different lengths. Before experimental co-crystal structures are available, we have performed molecular docking to predict how these compounds might bind to the protein and to generate structural models for performing binding affinity calculation to aid future optimization of these series of compounds.**

Key words: click chemistry, distance-dependent dielectric model, docking by simulated annealing, Generalized Born model, receiver operating characteristics curve, sensitivity and specificity of screening models

Received 10 March 2010, revised 6 May 2010 and accepted for publication 6 May 2010

Yersinia pestis can cause human diseases such as gastrointestinal syndromes and bubonic plague (1–3) and could be misused as a biological weapon (4). There is already evidence that blocking the

protein tyrosine phosphatase YopH of this bacterium can be an effective therapeutic strategy. For example, altering the gene of YopH to a non-functional one removed the bacterium's pathogenicity (5–7). Mutating the essential catalytic cysteine residue of YopH to alanine also abolished its protein tyrosine phosphatase activity and dampened the pathogenic effects of the bacterium (8,9). Consequently, potent and selective YopH inhibitors are expected to serve as novel anti-plague agents.

Several YopH inhibitors have already been identified over the last few years: Sun *et al.* (4) developed p-nitrocatechol sulfate (pNCS) and determined its co-crystal structure with YopH. Phan *et al.* (10) designed a hexapeptide mimic, Ac-DADE-F2Pmp-L-NH₂, of the protein's natural substrate (F2Pmp stands for difluoro-substituted phosphonomethylphenylalanine, which is a phosphotyrosine analog.) and determined its co-crystal structure with the protein. Liang *et al.* (11) identified aurintricarboxylic acid as a potent inhibitor of YopH and it displayed 6- to 120-fold selectivity in favor of YopH over a panel of mammalian protein tyrosine phosphatases. Tautz *et al.* (12) screened the DIVERSet™ library (ChemBridge, Inc. San Diego, CA, USA) of drug-like compounds and identified furanyl salicylate compounds as potent inhibitors of YopH. Hu and Stebbins (13) performed molecular docking and 3D-QSAR studies to rationalize the binding of derivatives of α -ketocarboxylic acids and squaric acid to YopH and to provide 3D-QSAR models to guide future refinement of this class of compounds.

In spite of these encouraging developments, the search for additional drug leads remains vital as many factors can prevent existing drug leads from passing through a series of stringent preclinical and clinical evaluations to become successful drugs. In this regard, most YopH inhibitors reported in the literature display unfavorable pharmacological properties and are not cell permeable. Moreover, multidrug-resistant strains of *Yersinia pestis* can emerge (14,15). To develop YopH inhibitors that carry sufficient polar and non-polar interactions with the active site and yet possess favorable pharmacological properties, we decided to capitalize our previous findings that the natural product salicylic acid can serve as a pTyr surrogate (16) and that naphthyl and polyaromatic salicylic acid derivatives exhibit enhanced affinity for protein tyrosine phosphatase relative to the corresponding single ring compounds (11,16). Therefore, in this work, we synthesized a new class of benzofuran salicylic acids and found many of them to demonstrate μ M activity.

Our initial design principle assumed the benzofuran salicylic acid core to bind to the phosphotyrosine-binding pocket. By introducing an additional chemical entity, linked to the core by a flexible hydrophobic linker, we hoped to target a neighboring pocket simulta-

neously to improve potency and selectivity. This article presents two series of these compounds differing by having different length of the linker connecting the two chemical moieties (B and D series shown in Figure 1).

To investigate whether these compounds are likely to bind the way that we expected, we performed molecular docking using a flexible ligand-flexible protein model we developed recently. The method improved docking by going beyond the rigid-protein approximation to account for induced-fit effects so that it could dock a wider range of ligands properly to a protein. The model used molecular dynamics simulation as a sampling tool. However, instead of running simulations at a constant temperature, it employed a simulated annealing cycling protocol to improve sampling efficiency. The protein was not completely flexible but with harmonic constraints applied to the α carbons to keep its structure near a suitable reference structure such as one obtained from X-ray crystallography. However, all other atoms, including all the side chains, were unrestrained (17,18). Although not yet a completely flexible protein model, this model avoided artifacts resulting from non-optimal energy and solvation models by focusing on exploring the conformational space near a known experimental structure. Previously, we showed that this model successfully docked several small organic ligands to protein kinases and the protein tyrosine phosphatase YopH (17,18); a completely rigid-protein model, on the other hand, failed for some ligands studied (18).

In this application, we further leveraged this approach by performing docking in two stages to improve speed without significantly sacrificing reliability. As we were studying relatively similar compounds within a chemical series, we assumed docking several representative ligands in stage 1 could generate all the major docking modes accessible by every compound in the series. Then, in stage 2, we allowed each compound in the series to refine its structure around these major docking modes by performing less expensive docking on a focused subset of configurational space. We then applied the resulting docking poses to compute binding affinity to check further whether they yielded results consistent with experimental IC₅₀. In the future, one can use the best resulting structural models for performing binding affinity calculation on new derivatives to suggest new compounds that are worthwhile to make and test experimentally.

Methods

Experimental

Library synthesis

Figure 2 depicts a focused library-based strategy for the acquisition of potent and selective YopH inhibitors that are capable of bridging both the active site and an adjacent peripheral site. The library contains (i) a benzofuran salicylic acid core to engage the active site, and (ii) 2 alkyl linkers of 2 and 4 methylene unit to tether the pTyr surrogate to (iii) a structurally diverse set of 20 amines, aimed at capturing additional interactions with adjacent pockets surrounding the active site. The benzofuran salicylic acid core **1** was prepared from a commercially available compound 4-hydroxysalicylic acid

that, upon regioselective bromination, afforded 5-bromo-4-hydroxysalicylic acid (**2**). This compound was selectively protected in the presence of acetone and trifluoroacetic anhydride/trifluoroacetic acid to furnish dioxanone **3**, which then reacted with CH₃I to give the methylation product **4**. Compound **4** was coupled with phenylacetylene in the presence of Pd(PPh₃)₄ to furnish **5**, which was then subjected to I₂-induced cyclization. Coupling of the iodination product **6** with ethynyltrimethylsilane gave compound **7**. Desilylation and deacetylation of **7** provided the core compound **1**.

To increase potency and selectivity, the strategically positioned alkyne in the benzofuran salicylic acid core was tethered to 40 azide-containing diversity elements (20 discrete amines with 2 alkyl linkers of 2 and 4 methylene length), using click chemistry or the Cu(I)-catalyzed [3 + 2] azide-alkyne cycloaddition reaction (Figure 2). The click chemistry offers an expedient way to connect two components together with high yield and purity under extremely mild conditions (19). More importantly, the cycloaddition reaction can be conducted in aqueous solution in the absence of deleterious reagents, thus allowing direct screening and identification of hits from the library. The azide-containing building blocks were synthesized in a one-pot procedure, in which alkyl or aryl amines were reacted with the acyl chloride linkers in N, N-Dimethylformamide (DMF), followed by S_N2 reaction with sodium azide to generate the corresponding azides. To construct the 40-member library, each azide was coupled with the alkyne containing core **1** in a mixed solvent of ethanol and water (7:3) and the click reaction was initiated by catalytic amount of Cu(I), which was generated by reacting CuSO₄ with sodium ascorbate. After 48 h, the products were collected by simple centrifugation. All products were assessed by LC-MS and determined to be at least 70–100% pure and were used directly for screening without further purification.

Screening of the salicylic acid library for YopH inhibitors

To screen the salicylic acid library for YopH inhibitors, the effect of each library member on the YopH-catalyzed *p*-nitrophenyl phosphate (*p*NPP) hydrolysis was determined. The YopH-catalyzed hydrolysis of *p*NPP in the presence of 10 μ M compound was assayed at 30 °C in a 200 μ L reaction system in a 96-well plate. Each reaction contained 2 μ L of 1 mM compound in DMSO (final concentration 10 μ M) and 198 μ L assay buffer (50 mM 3,3-dimethylglutarate, 1 mM EDTA, 1 mM DTT, pH 7.0 with an ionic strength of 0.15 M adjusted by addition of NaCl) containing 2 mM *p*NPP and 10 nM YopH. The PTP-catalyzed reaction was started by addition of the enzyme. As a control, 2 μ L of DMSO was used. The YopH-catalyzed hydrolysis of *p*NPP was measured by monitoring the absorbance at 405 nm of the product *p*-nitrophenol continuously, with a SpectraMAX 340 microplate spectrophotometer (Molecular Devices, Sunnyvale, CA, USA). The initial rate was obtained by calculating the slope of the product versus the time curve. Compounds that display significant inhibition at 10 μ M were subject to IC₅₀ measurements.

IC₅₀ measurement

The YopH-catalyzed hydrolysis of *p*NPP in the presence of inhibitor was assayed at 30 °C in a 200 μ L reaction system in the same

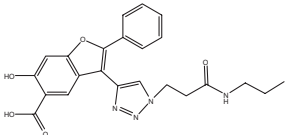
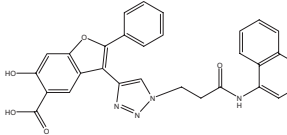
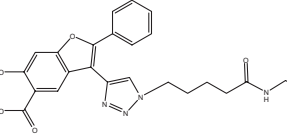
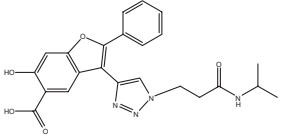
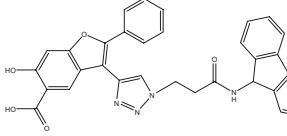
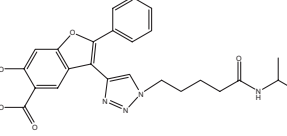
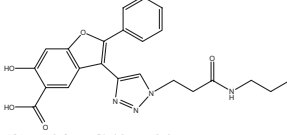
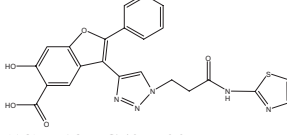
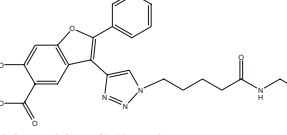
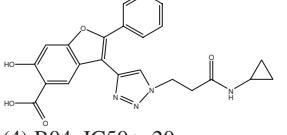
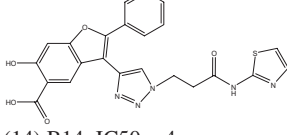
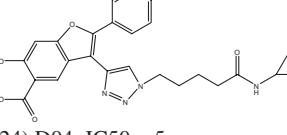
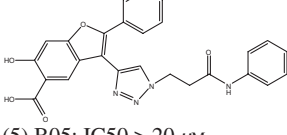
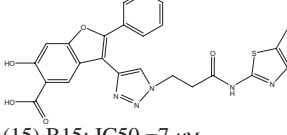
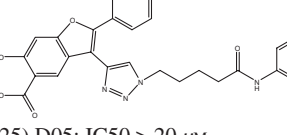
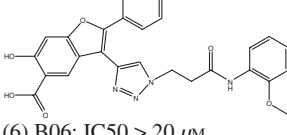
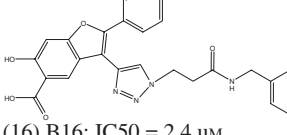
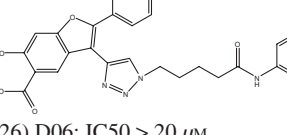
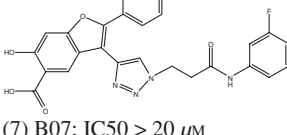
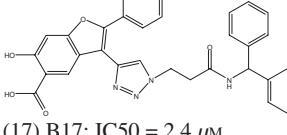
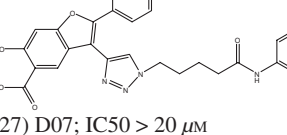
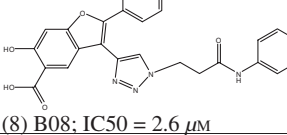
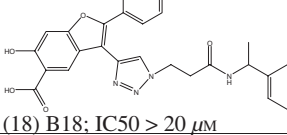
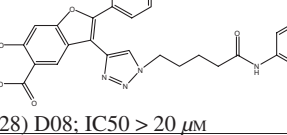
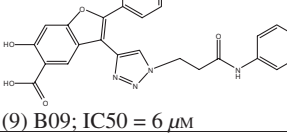
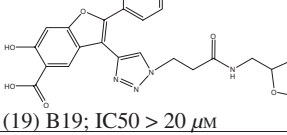
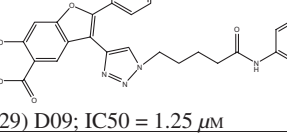
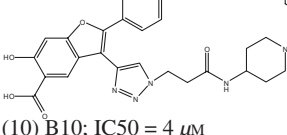
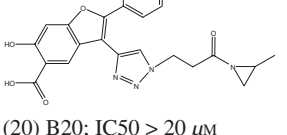
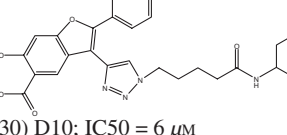
 (1) B01; IC ₅₀ > 20 μ M	 (11) B11; IC ₅₀ = 1.2 μ M	 (21) D01; IC ₅₀ > 20 μ M
 (2) B02; IC ₅₀ > 20 μ M	 (12) B12; IC ₅₀ = 3 μ M	 (22) D02; IC ₅₀ > 20 μ M
 (3) B03; IC ₅₀ > 20 μ M	 (13) B13; IC ₅₀ > 20 μ M	 (23) D03; IC ₅₀ = 3 μ M
 (4) B04; IC ₅₀ > 20 μ M	 (14) B14; IC ₅₀ = 4 μ M	 (24) D04; IC ₅₀ = 5 μ M
 (5) B05; IC ₅₀ > 20 μ M	 (15) B15; IC ₅₀ = 7 μ M	 (25) D05; IC ₅₀ > 20 μ M
 (6) B06; IC ₅₀ > 20 μ M	 (16) B16; IC ₅₀ = 2.4 μ M	 (26) D06; IC ₅₀ > 20 μ M
 (7) B07; IC ₅₀ > 20 μ M	 (17) B17; IC ₅₀ = 2.4 μ M	 (27) D07; IC ₅₀ > 20 μ M
 (8) B08; IC ₅₀ = 2.6 μ M	 (18) B18; IC ₅₀ > 20 μ M	 (28) D08; IC ₅₀ > 20 μ M
 (9) B09; IC ₅₀ = 6 μ M	 (19) B19; IC ₅₀ > 20 μ M	 (29) D09; IC ₅₀ = 1.25 μ M
 (10) B10; IC ₅₀ = 4 μ M	 (20) B20; IC ₅₀ > 20 μ M	 (30) D10; IC ₅₀ = 6 μ M

Figure 1: Chemical structure of ligands in the B series and the D series.

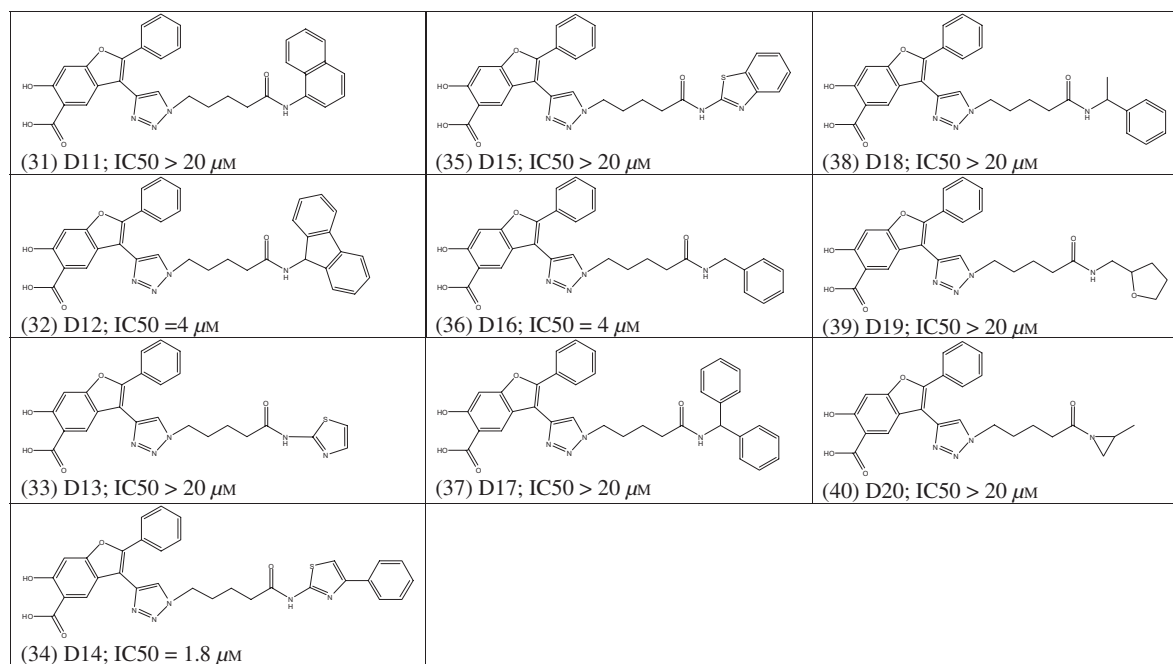


Figure 1: (Continued).

assay buffer described above. At various concentrations of the compound, the initial rate at fixed *p*NPP concentrations (equal to the corresponding *K_m* value for YopH) was measured by continuously following the production of *p*-nitrophenol as described above. The IC₅₀ value was determined by plotting the relative PTP activity toward *p*NPP versus inhibitor concentration and fitting to equation (1) using Kaleidagraph.

$$V_i/V_0 = IC_{50}/(IC_{50} + [I]) \quad (1)$$

In this case, *V_i* is the reaction velocity when the inhibitor concentration is [*I*], *V₀* is the reaction velocity with no inhibitor.

General Computational Strategy

The following steps summarize our strategy:

- Stage 1 docking: Performed extensive flexible ligand-flexible protein docking for a few representative compounds in a series to estimate how the ligands bound to YopH. This was the most expensive part of the simulation. Clustering the twenty lowest-energy structures afterward then generated several major docking modes for stage 2 docking.
- Stage 2 docking: Used each major docking mode from stage 1 to construct the structures of all the other compounds in the series and refined these structures by performing focused docking in which we only allowed the ligands and their surrounding protein residues to move; this significantly reduced computational cost in comparison to the less restrictive docking in stage 1. The focusing of the final docking of a similar series of compounds to the most relevant regions could also reduce statistical noise and systematic

errors in the subsequent computation of binding affinity to help distinguish actives from non-actives.

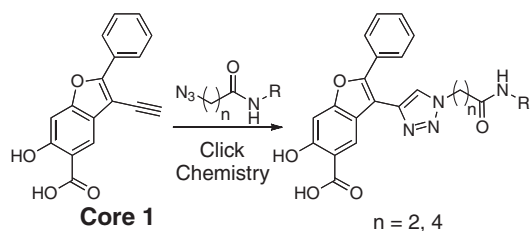
- Used the docking poses from stage 2 docking to estimate the binding affinity of each compound in the series employing either a distance-dependent dielectric model or the Generalized Born model termed GBMV (20–22).
- Checked which or which mixture of major docking modes gave the best agreement with experiment in terms of classifying the compounds into actives and non-actives – defined as compounds having IC₅₀ less than and greater than 20 μM, respectively – using several quantitative measures described below.
- Use the best docking models in the future to aid optimization of these series of compounds by attaching different functional groups to the core chemical skeleton and examine which modifications could lead to inhibitors with improved binding affinity.

Protein–ligand docking by simulated annealing cycling

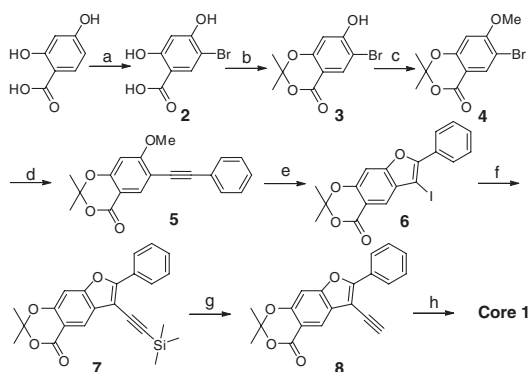
We performed molecular docking by using a molecular dynamics-based simulated annealing (SA) cycling protocol we published earlier (17,23). The method performed many short SA cycles in a molecular dynamics simulation to sample many energy minima of a protein–ligand system thoroughly. The lowest-energy poses then recommended how the ligand might bind to the protein.

We conducted the SA cycling simulations using the CHARMM param22 force field (24,25). In the docking simulations, we used a simple but inexpensive distance-dependent dielectric model with $\epsilon(r) = 4r$ where *r* was the distance between two atoms. In calculat-

A Library construction



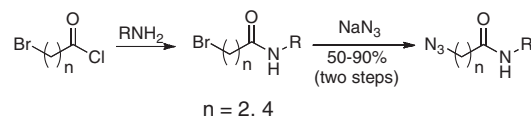
B Synthesis of Core 1



Conditions:

a) Br_2 , AcOH, r.t., 94%; b) TFAA, TFA, Acetone, r.t., 55%; c) MeI, K_2CO_3 , DMSO, r.t., 95%; d) Phenylacetylene, $\text{Pd}(\text{PPh}_3)_4$, CuI, Et_3N , DMF, 70°C , 72%; e) I_2 , NaHCO_3 , MeCN, 40°C , 40%; f) Trimethylsilylacetylene, $\text{Pd}(\text{PPh}_3)_2\text{Cl}_2$, CuI, Et_3N , DMF, r.t., 85%; g) TBAF, THF, r.t., 89%; h) LiOH, THF/ H_2O , 80°C , 71%.

C Synthesis of azides



D Chemical Structures of 20 amines

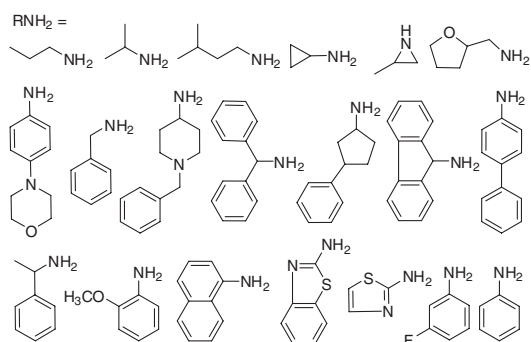


Figure 2: Synthetic scheme.

ing binding affinity, we further used the more sophisticated implicit-solvent GBMV model (20–22). During the simulation, we used a non-bonded cutoff distance of 14 Å, a switching function for the electrostatic interactions that began at 10 Å and ended at 12 Å, and a shifting function for the Lennard–Jones potential. We also used this $\epsilon(r) = 4r$ model in the energy minimization preceding each molecular dynamics simulation to relieve bad contacts. In recalculating the binding affinity with the GBMV (20,21) model, we used the

GBMV1 parameters of Chocholoušová and Feig (22) and we applied the same cutoff distances as described above for the $\epsilon(r) = 4r$ model.

Two-stage docking strategy

Stage 1 initial docking of representative ligands

For each of the two chemical series, B and D, studied here, we selected three compounds for thorough flexible ligand-flexible protein docking. We chose compounds B11, B16, and B17 for the B series and compounds D03, D09, and D14 for the D series (structures shown in Figure 1). For docking each ligand to YopH, we started the simulations from four different positions/structures. One lied inside the same pocket that pNCS binds [PDB entry 1PA9(4)]. Another was located on the surface of the protein near the binding pocket. At each location, we placed the ligands in two near anti-parallel orientations, thus generating four different starting structures to initiate docking runs. The structure in PDB entry 1QZ0 provided the starting structure for the YopH protein (10). The initial 3D structures of the ligands were prepared by using ChemSketch (26). We performed ten independent SA cycling simulations for each of the four starting structures, thus giving forty trajectories. Because each trajectory lasted 2 ns, the aggregate simulation time for each protein–ligand system covered 80 ns. In these simulations, we used a time step of 2 fs.

We allowed the protein to move in the docking simulations but with an appropriate restraint to prevent the protein from deviating too far away from the crystal structure in PDB code 1QZ0. We achieved this by applying the harmonic potential $F \times D^2$, where F was a force constant and D was the root-mean-square deviation (RMSD) of a dynamics snapshot from the crystal structure (an option in CHARMM) (27). Only the α -carbons were used to calculate the RMSD so that the side chains and the other backbone atoms were free to move. In our previous work, (17,18), we set $F = 1000 \text{ kcal/mol/\AA}^2$. In this work, we used two F 's in two different parts of the protein. We applied a smaller force constant of $100 \text{ kcal/mol/\AA}^2$ to the flexible WPD-loop containing nine residues spanning from Gly-352 to Val-360. The larger force constant $F = 1000 \text{ kcal/mol/\AA}^2$ was applied to the rest of the protein. We used a smaller force constant for the WPD-loop because it was more flexible than the other parts of the protein surrounding the phosphotyrosine-binding pocket (3,28). These restraints prevented sampling unrealistic structures because of the limitation of current force fields and solvation models but permitted larger conformational change of the protein to allow a larger range of compounds to dock properly to the protein. Prior to each molecular dynamics simulation, we performed 500 steps of steepest descent energy minimization on the protein–ligand complex to remove bad contacts.

Stage 2 docking of all compounds in each series

Using each selected major docking mode obtained from stage 1, we used the VEGA ZZ program (29) to generate the structures of all the compounds in a series by making the appropriate chemical modifications. We then refined these structures by performing SA

cycling simulations that allowed the ligands to move freely but restricted the protein to a larger extent than in stage 1 docking. In this stage 2 docking, we restrained the α carbons of protein residues within 5.0 Å of a docking ligand with $F \times D^2$ in which $F = 1000 \text{ kcal/mol/Å}^2$. Here, in calculating D , we used a low-energy pose obtained from stage 1 rather than from the crystal structure. The rest of the protein was held fixed so that stage 2 docking took significantly less computational time than stage 1. Table S1 shows the movable protein residues for the four representative low-energy poses for each compound series. In this stage, we ran only four trajectories for each ligand with each trajectory lasting 2 ns. As in stage 1, we also performed 500 steps of steepest descent energy minimization of the protein–ligand complex to remove bad contacts prior to running each molecular dynamics trajectory.

On a dual core-dual processor cluster node with 2.8 GHz Intel Xeon EM64T processors, it took ~28 h for each trajectory and ~1120 h for 40 trajectories performed for each ligand in stage 1 docking. On the other hand, stage 2 docking only took ~4–6 h for each trajectory and ~16–24 h per ligand with four trajectories. The simulation time varied slightly among different ligands.

Scoring of docking poses

To identify the best docking poses, we used the sum of the energy of the ligand and the interaction energy between the protein and the ligand with the $\epsilon(r) = 4/r$ model. We did not use the total energy because it was noisy to use from finite simulations (18). Removing the noisy components arising from the protein identified docking poses better in our previous docking of small organic ligands to protein kinases and to YopH (17,18,23).

Selecting representative docking modes from stage 1 for stage 2 docking

As shown in Results and Discussions below, stage 1 docking did not produce an unambiguous best docking mode for each series of compounds. We therefore selected several most likely docking modes to use for stage 2 docking and computed binding affinity using each docking mode to find out which mode was most consistent with measured IC₅₀.

To identify a small number of docking modes for stage 2 docking and binding affinity calculation, we first clustered the structures obtained from stage 1 by using a self-organizing neural net approach (30–33) implemented in CHARMM and accessible from the Multiscale Modeling Tools for Structural Biology (MMTSB) toolkit. The cluster algorithm optimizes cluster assignment by minimizing the distance between members and their centroid structure within each cluster and by requiring this distance to be within a user-predefined cluster radius. One does not need to specify the number of clusters as the algorithm determines the optimal number that satisfies the above criteria. In this work, we measured the distance between two ligand structures by calculating the RMSD between the Cartesian coordinates of their heavy atoms after we superimposed their cognate protein structures with the crystal structure. We set the cluster radius to 2 Å and we used the low-

Table 1: Number of clusters obtained from 40 2-ns protein–ligand trajectories for six ligands selected for stage 1 docking

Ligand	B11	B16	B17	D03	D09	D14
Small cluster	729	792	626	811	671	683
Large cluster	4	4	4	4	5	7

est-energy structure of each cluster to represent the structure of the cluster. After clustering, we selected the twenty lowest-energy clusters for each ligand. When graphically examining their three-dimensional structures, we found that some clusters gave visually similar structures. We therefore regrouped these twenty clusters into a smaller number of bigger clusters. This was performed by re-clustering the twenty clusters using a larger cluster radius of 5 Å. We also used the lowest-energy structure of each resulting cluster to represent the structure of each bigger cluster. Table 1 shows the number of clusters obtained from the first and second round of clustering for the six ligands selected in stage 1 docking.

Rigid-protein docking

For comparison, we also performed rigid-protein docking with Autodock v4.0.1 (34,35). To check the sensitivity of the results to the choice of protein structures, we used three different crystal structures [PDB entries 1QZ0 (10) (closed form), 1PA9 (4) (closed form), and 1YPT (3) (open form)]. In addition, for each protein–ligand structure obtained from stage 2 docking, we re-docked the ligand and recomputed its binding affinity using Autodock and carried out the performance analysis described below to check whether the flexible-receptor model gave better results.

Target preparation

Autodock prepared several potential grids for each protein structure. These grids contained the electrostatic potential and the van der Waals energies for the atom types C (aliphatic carbon), A (aromatic carbon), N (nitrogen), O (oxygen), SA (aromatic sulfur), H (polar hydrogen), and F (fluorine) on the ligand. For calculating electrostatics interactions, we used Kollman charges for the protein atoms and Gasteiger charges for the ligand atoms. Each grid measured $70 \times 70 \times 70$ with a grid spacing of 0.375 Å and it centered at the heart of four pockets comprising the phosphotyrosine-binding pocket and three neighboring secondary pockets.

Docking protocol

We selected the Lamarckian genetic algorithm of Autodock to search for the best scoring conformation. In addition, we set the docking parameters to: 100 docking runs, population size = 150, random starting position and orientation, maximum translation step size = 2 Å, maximum rotation allowed for each step = 35°, elitism = 1, mutation rate = 0.02, crossover rate = 0.8, local search rate = 0.06, and 25 million energy evaluations. We clustered the final structures using a 2.0-Å cutoff.

On a dual core-dual processor cluster node with 2.8 GHz Intel Xeon EM64T processors, it took ~28–51 h to dock each ligand to a rigid

protein model. This computation time nearly doubled that of stage 2 docking using the flexible-protein model.

Judging model performance

To help judge how well each model performed in classifying compounds into actives and non-actives, we calculated the following quantities:

Sensitivity and specificity

Drug designers (36,37) defined the sensitivity (Se) to be the ratio of the number of active molecules correctly predicted to the total number of actives present. In terms of the number of true positive, TP, and the number of false negatives, FN:

$$Se = \frac{N_{\text{correctly predicted actives}}}{N_{\text{total actives}}} = \frac{TP}{TP + FN} \quad (2)$$

Specificity (Sp), on the other hand, is the ratio of the number of non-active compounds correctly predicted to the total number of non-active molecules. In terms of the number of true negatives, TN, and the number of false positive, FP,

$$Sp = \frac{N_{\text{correctly predicted non-actives}}}{N_{\text{total non-actives}}} = \frac{TN}{TN + FP} \quad (3)$$

Accuracy (Acc)

It describes the percentage of molecules classified correctly into actives and non-actives (37,38):

$$Acc = \frac{N_{\text{correctly predicted actives}} + N_{\text{correctly predicted non-actives}}}{N_{\text{total}}} = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

Enrichment factor

The enrichment factor, EF, measures how well a model increases the fraction of actives identified relative to the fraction of actives in a database: (38).

$$EF = \frac{TP/(TP + FP)}{(\text{number of actives})/(\text{total number of compounds in a database})} \quad (5)$$

Receiver operating characteristic (ROC) curve

This curve plots Se versus 1-Sp obtained by using different cut-off values of a suitable quantity – calculated binding affinity in this application – to separate actives from non-actives (36,37). Se represents the ability of the model to pick out true positives. On the other hand, 1-Sp hints on the tendency of the model to produce false positives. Good models yield Se near unity and 1-Sp close to zero. If one plots Se versus 1-Sp for different cut-offs and calculates the area under the ROC curve (AUC), good models yield areas near unity, random models give an area of 0.5, and models performing worse than random produce areas smaller than 0.5.

Results and Discussions

Figure 3A–F plot $RMSD_{\text{heavy}}$ versus energy for the six YopH–ligand systems selected in stage 1 docking, each docking covered a total of 80 ns of simulation time. Here, $RMSD_{\text{heavy}}$ represented the RMSD of all heavy atoms of the ligand between a structure near a local minimum (a structure obtained below 5 K) and one that had the lowest energy. The plot for ligand D09 gave one deep-energy well significantly separated from the next lowest-energy well. On the other hand, the other ligands yielded two or more deep-energy wells with relatively similar energies, making it difficult to identify a single best docking mode. We therefore performed the clustering described above to identify several major docking modes for stage 2 docking that included all ligands within a series and used the resulting structures to perform binding affinity calculation to examine which model produced results most consistent with measured IC50.

The clustering described earlier produced four major docking modes denoted by B-Model I to IV for the B series and D-Model I to IV for the D series in Table 2. We labeled the structural models such that B-Model I was most similar to D-Model I, B-Model II was most similar to D-Model II, B-Model III was most similar to D-Model III, and B-Model IV was most similar to D-Model IV. However, remember from Figure 1 that ligands in the two series differed by the length of the linker connecting the two chemical moieties intended to target two different pockets. Therefore, one would expect minor structural differences between B-Model I and D-Model I, B-Model II and D-Model II etc. Also, recall that we obtained these models in two rounds of clustering. The first round clustered every structure near local energy minima obtained from the SA docking using a cutoff of 2 Å. The second round merged the twenty lowest-energy clusters obtained from round 1 into a smaller number of larger clusters using a larger cutoff distance of 5 Å. We found four clusters that were adopted by five of the six ligands and we chose these clusters as the four major docking mode used for stage 2 docking.

Table 2 shows how the twenty lowest-energy clusters obtained in round 1 were grouped to form the four clusters in round 2. For example, the notation 1st/11 in D-Model I for ligand D09 means that eleven of the twenty lowest-energy clusters obtained from round 1 were merged into one cluster and the lowest-energy structure came from the 1st small cluster obtained in round 1. (We labeled the clusters from 1st to 20th in increasing energy.) The structure of the 1st cluster was used to represent D-Model I and as a template to construct all the ligands for stage 2 docking.

Five (B11, B16, B17, D03, and D14) of these six ligands all yielded the above four docking modes. D09, on the other hand, did not assume Model IV. Instead, it adopted another docking mode in which the ligand was sharply bended at the linker with the chemical moieties on its two ends tightly packed against each other (shown in "Others" in Table 2). We did not include this docking mode in stage 2 docking because this docking mode occurred infrequently and at higher energies and were thus less likely to be the correct docking mode. Ligand D14 also took on a similar sharply

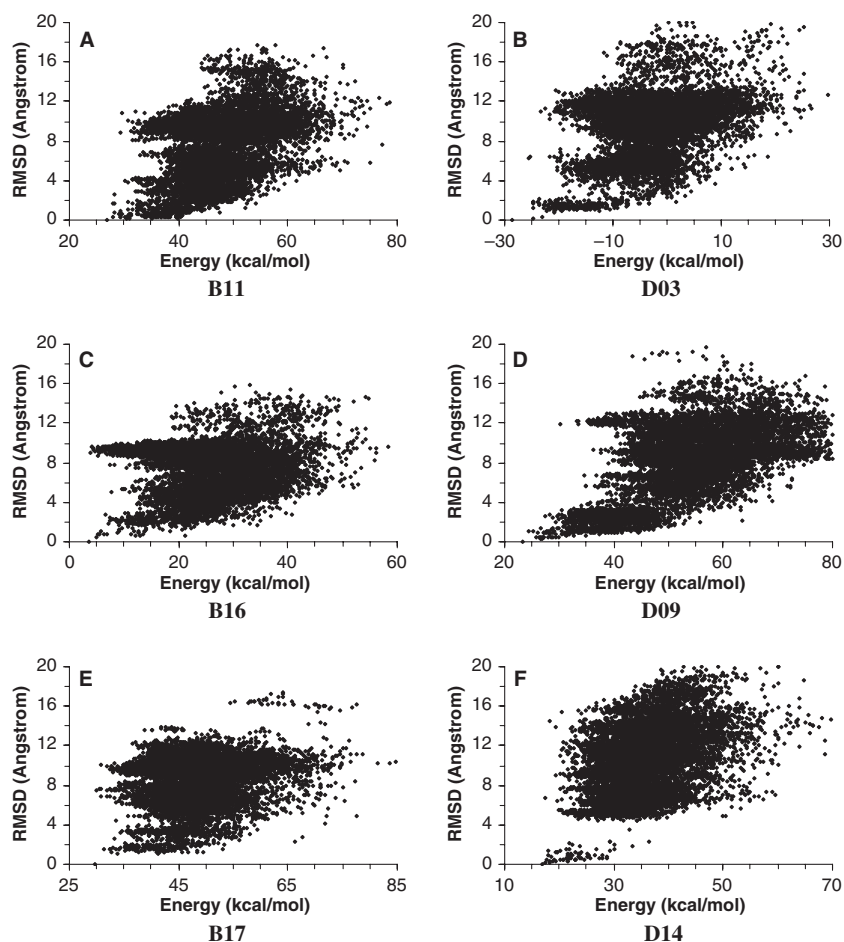


Figure 3: Plots of ligand $\text{RMSD}_{\text{heavy}}$ versus energy for docking six ligands to YopH. (A) ligand B11, (B) ligand B16, (C) ligand B17, (D) ligand D03, (E) ligand D09 system, (F) ligand D14.

Table 2: Twenty best structural clusters grouped into four larger clusters

Model	Ligand					
	B11	B16	B17	D09	D03	D14
B-Model I ^a / D-Model I	14th/1 ^b	2nd/6	1st/3	1st/11	1st/3	6th ^c /5
B-Model II/ D-Model II	1st/7	4th/4	2nd/9	14th/2	2nd/5	8th/5
B-Model III/ D-Model III	4th/8	1st/6	4th/7	6th/2	6th/4	14th/2
B-Model IV/ D-Model IV	11th ^c /4	11th/4	19th/1		3rd/8	3rd/2
Others				2 ^d /5		3 ^d /6

^aB-Model I, D-Model I, etc., denotes one of four major docking modes represented by a large cluster of structures.

^bThe first number identifies the smaller cluster, within a bigger cluster, which contained the lowest-energy structure. The second number tells how many small clusters (formed by using a 2-Å cutoff) were grouped to form the bigger cluster (formed by regrouping the twenty lowest-energy small clusters using a clustering cutoff of 5 Å).

^cIndicates that the second, rather than the first, lowest-energy structures were selected because they were closer to the lowest-energy structures obtained for the other two ligands.

^dRepresent sharply bent docking modes that was ignored in stage 2 docking.

Table 3: The major docking modes adopted by the lowest-energy structures of the six ligands used in stage 1 docking

Ligand	B11	B16	B17	D03	D09	D14
Model	B-Model II	B-Model III	B-Model I	D-Model I	D-Model I	D-Model I

bended docking mode but again was ignored for stage 2 docking for the same reason. Table 3 summarizes the docking mode adopted by the lowest-energy structure of each ligand. The three ligands in the B series took on different docking modes as their lowest-energy docking structures. On the other hand, the lowest-energy structures for the three ligands in the D series all accepted Model I.

Figure 4 displays the structures of these docking modes. In each model, two (for D-Model IV only) or three different ligands that docked similarly were overlaid. For example, B-model I in Figure 4A shows the lowest-energy structures from the 14th cluster for ligand B11, from the 2nd cluster for ligand B16, and from the 1st cluster for ligand B17. On the other hand, D-model I in Figure 4E displays the lowest-energy structures from the 1st cluster for ligand D09, the 1st cluster for ligand D03, and the 6th cluster for ligand D14. The figure shows that six binding modes (B/D-Model I, B/D-Model II, and B/D Model IV) had the salicylic acid core docked to the phosphotyrosine-binding pocket. They differed mainly in the

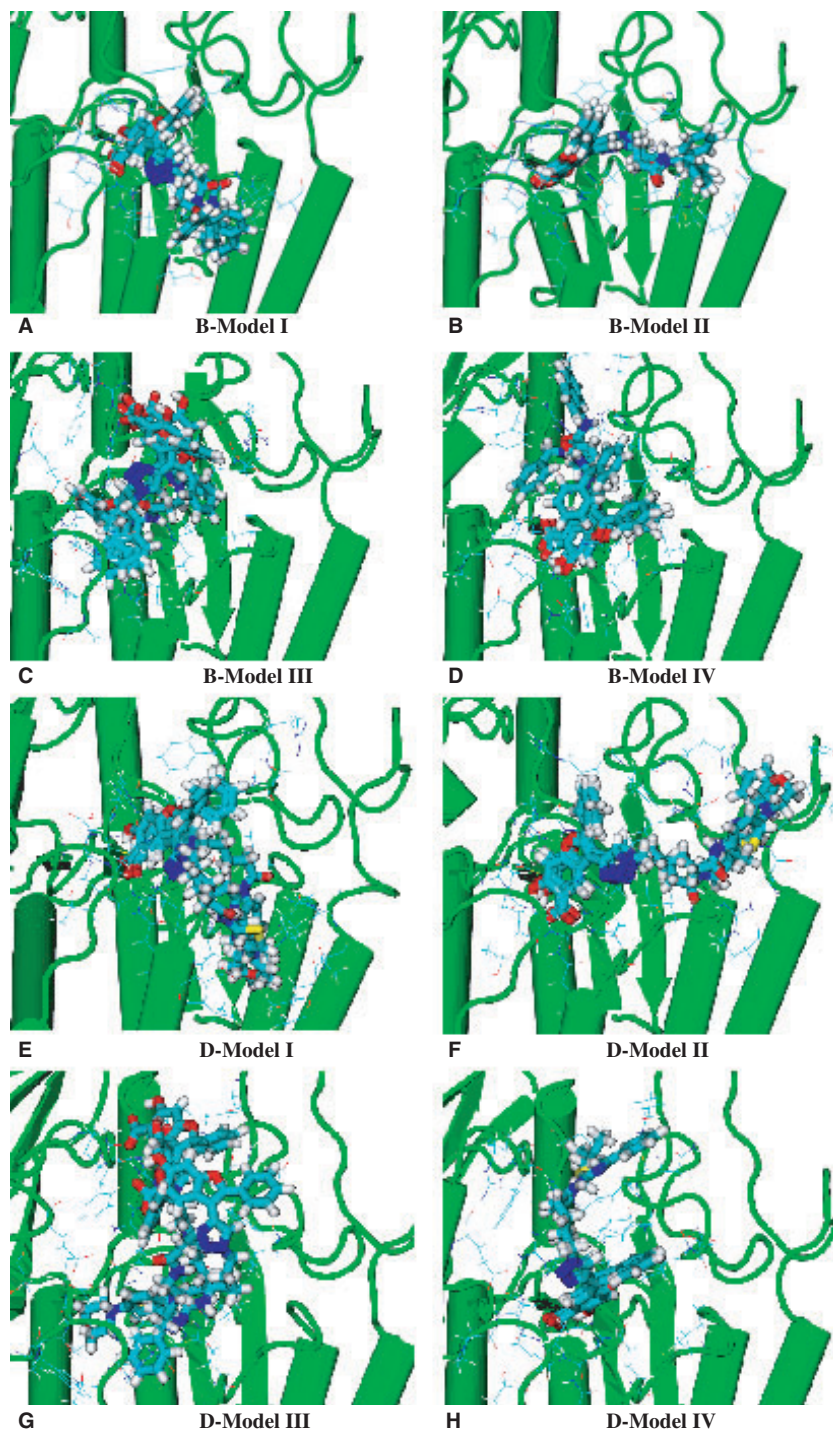


Figure 4: Four major docking modes identified from flexible ligand-flexible protein docking. (A) B-Model I. (B) B-Model II. (C) B-Model III. (D) B-Model IV. (E) D-Model I. (F) D-Model II. (G) D-Model III. (H) D-Model IV. Coloring scheme: protein in green, oxygen in red, nitrogen in blue, carbon in cyan, and sulfur in yellow. Structural representation: protein in cartoon, ligand in stick, and side chains (from Table S1) in line mode. The pictures were generated by Visual Molecular Dynamics (VMD) (39).

non-salicylic portion of the compounds. On the other hand, B/D Model III differed by having the non-salicylic acid core docked inside the phosphotyrosine-binding pocket, with the salicylic acid core exposed to the protein surface.

Figure 5 uses the surface representation to show the different ways that ligand B11 might fit into the protein. The four lowest-energy structures in column 2 of Table 2 are shown. These structures show that one end of the bidentate ligand always occupied the phospho-

hotyrosine-binding pocket. On the other hand, the other end of the ligand fitted into three different secondary pockets. A closer comparison of these structures revealed that these pockets changed somewhat depending on which ligand was bound to them – a flexible-protein model, as done here, is essential to capture such effects.

Stage 2 docking assumed that all the compounds within a series bound in one of these four docking modes with minor adjustment

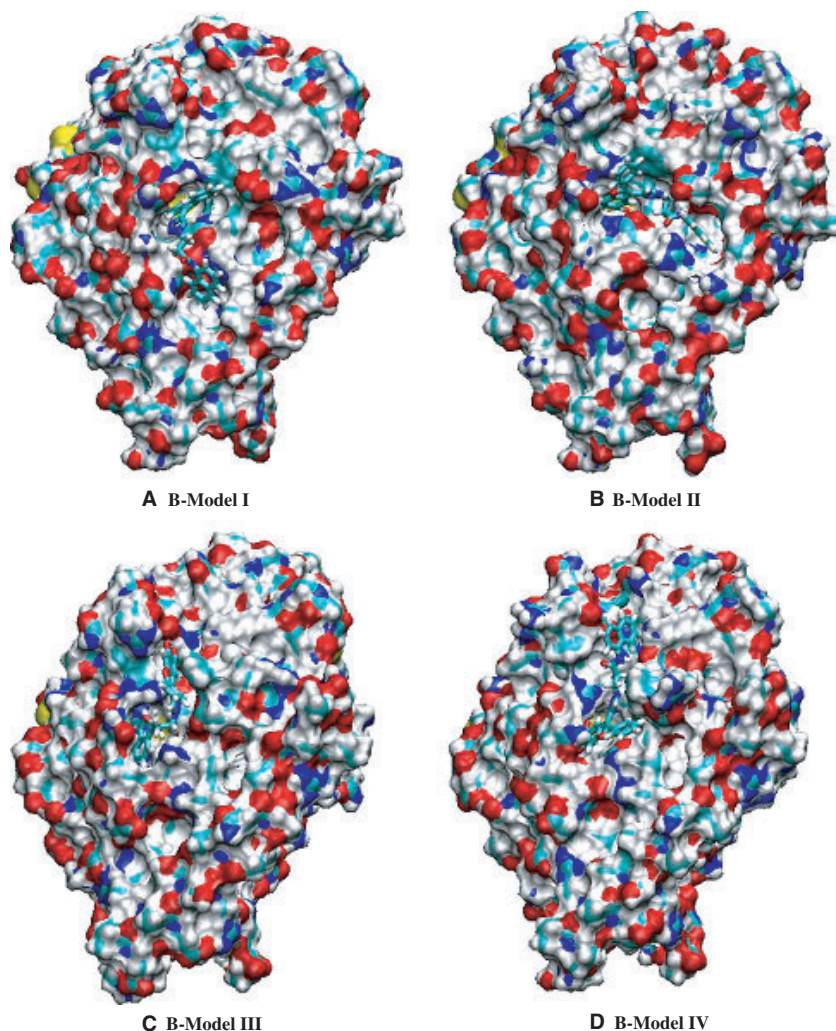


Figure 5: Four major docking modes from flexible ligand-flexible protein docking for compounds in the B series using a surface representation of the protein. (A) B-Model I. (B) B-Model II. (C) B-Model III. (D) B-Model IV. Ligands are displaced in stick mode. Coloring scheme: oxygen in red, nitrogen in blue, carbon in cyan, and sulfur in yellow. The picture was generated by VMD (39).

of the protein structures at the protein–ligand interface. D-Model I appeared to be the most likely docking mode for the D series, as it was the lowest-energy docking pose for all the three ligands in the D series on which extensive flexible ligand-flexible protein docking was performed in stage 1 (Table 3). On the other hand, the three ligands in the B series found different major docking modes as their lowest-energy docking pose, suggesting the possibility that not all compounds in the B series might bind with the same major docking mode.

To check this further, we used each structural model to perform binding affinity calculation to find out which model gave results most consistent with experimental IC₅₀. Table 4 gives the calculated binding affinity using the $\epsilon(r) = 4r$ or the GBMV solvation model for the twenty ligands in the B series and for each of the four different structural models: B-Model I–IV obtained in stage 1. The results also included a mixed structural model in which not all ligands within the series were required to dock to the same major docking mode. Instead, each ligand selected the docking mode that gave the most favorable binding affinity. In this preliminary evaluation of the performance of the different docking modes, we simply used one cutoff value of the binding affinity to

divide the compounds into actives and non-actives. Because there were nine known actives for the B series, we first classified the nine compounds with the most favorable predicted binding affinity as actives and the rest as non-actives. The sensitivity, specificity, accuracy, and enrichment factor in the table then indicate that B-Model I performed the best for both solvation models (with the best accuracy, for example). However, the mixed structural model was not far behind for both solvation models. B-Model IV also scored quite well for the $\epsilon(r) = 4r$ solvation model but did not score as well with the GBMV model. Based on these data, we therefore concluded that B-Model I was most likely, although the mixed structural model could be a possibility as well. For B-Model I, the salicylic acid core bound to the phosphotyrosine-binding pocket as intended.

Table 5 shows the corresponding results for the D series. Because the experimental results only showed seven actives, we classified the seven compounds predicted to have the most favorable binding affinity as actives. The results suggest that D-Model I performed the best for both solvation model. In contrast to the B series, the mixed structural model only scored well for the $\epsilon(r) = 4r$ solvation model but not for GBMV model. If we only trust docking models

Table 4: Binding affinity of ligands in the B series obtained from the $\epsilon(r) = 4r$ and the GBMV solvation models for the four major docking modes identified from stage 1 docking

Ligands	$\epsilon(r) = 4r$ model					GBMV model				
	I	II	III	IV	Mix ^a	I	II	III	IV	Mix ^a
B01	-66.08	-68.36	-67.93	-69.30	-69.30	-49.75	-41.72	-42.66	-53.80	-53.80
B02	-67.98	-68.76	-66.91	-69.16	-69.16	-47.59	-42.07	-41.31	-51.27	-51.27
B03	-69.95	-73.18	-72.38	-71.60	-73.18	-53.30	-44.88	-46.78	-48.23	-53.30
B04	-64.95	-65.61	-65.05	-69.58	-69.58	-50.50	-40.59	-43.15	-50.40	-50.50
B05	-69.88	-72.59	-70.78	-72.73	-72.73	-50.39	-42.57	-48.43	-50.70	-50.70
B06	-73.54	-76.71	-72.70	-75.12	-76.71	-51.48	-39.61	-48.07	-52.41	-52.41
B07	-70.89	-74.17	-70.89	-73.38	-74.17	-50.95	-44.44	-53.45	-52.49	-53.45
B08	-80.83	-74.16	-72.45	-79.39	-80.83	-54.00	-41.40	-47.71	-56.04	-56.04
B09	-86.87	-73.29	-78.64	-74.86	-86.87	-54.98	-44.79	-45.07	-56.20	-56.20
B10	-81.33	-74.26	-71.13	-78.41	-81.33	-57.49	-46.95	-39.28	-51.72	-57.49
B11	-75.68	-78.52	-74.27	-77.50	-78.52	-53.81	-45.54	-53.01	-50.95	-53.81
B12	-76.83	-74.03	-75.86	-76.85	-76.85	-54.54	-45.61	-46.00	-59.52	-59.52
B13	-69.81	-74.14	-75.09	-75.29	-75.29	-53.37	-45.20	-48.36	-51.17	-53.37
B14	-79.53	-76.35	-75.32	-79.78	-79.78	-54.17	-42.63	-55.08	-56.08	-56.08
B15	-78.79	-77.50	-73.73	-75.70	-78.79	-56.11	-48.59	-48.14	-50.97	-56.11
B16	-74.46	-76.23	-76.62	-76.02	-76.62	-52.20	-46.71	-57.20	-50.48	-57.20
B17	-77.91	-79.59	-76.44	-78.71	-79.59	-54.66	-43.51	-55.56	-51.19	-55.56
B18	-74.03	-73.87	-75.37	-77.30	-77.30	-51.86	-39.21	-54.03	-51.87	-54.03
B19	-69.75	-73.28	-71.39	-73.61	-73.61	-53.76	-42.13	-42.17	-50.77	-53.76
B20	-64.09	-64.14	-62.45	-66.51	-66.61	-47.76	-37.56	-36.28	-38.66	-47.76
TP/(TP + TN)	9/9	7/9	7/9	8/9	8/9	8/9	6/9	5/9	5/9	8/9
Se (%)	100	78	78	89	89	89	67	56	56	89
Sp (%)	100	82	82	91	91	91	73	64	64	91
Acc (%)	100	80	80	90	90	90	70	60	60	90
EF	2.22	1.73	1.73	1.98	1.98	1.98	1.48	1.23	1.23	1.98

^aIn the mixed model, ligands were allowed to select the docking mode that gave the most favorable binding affinity.

Bold: the nine ligands giving the most favorable binding affinity for each structural model.

TP, true positive; TN, true negative; Se, sensitivity; Sp, Specificity; Acc, accuracy; EF, enrichment factor.

that performed well with both solvation models, D-Model I was the best choice for the D series.

The above analysis relied on using one single cutoff value to classify compounds into actives and non-actives. To evaluate the models more thoroughly, we also generated ROC curves and calculated AUC which required repeating the calculations of sensitivity and specificity using different cutoff values for classifying compounds into actives and non-actives. Figure 6 shows these curves for different B Models. Table 6 gives the corresponding AUC. From these ROC plots and AUC values, we further confirmed that B-Model I performed the best for both solvation models. However, the mixed structural model had AUC almost as good. Therefore, our results did not significantly favor B-Model I over the mixed model. In future lead optimization, one may want to use both models to obtain consensus scoring and only suggests new derivatives that score well with both B-Model I and the mixed model for experimental testing.

Likewise, Figure 7 shows the ROC curves and Table 6 presents the corresponding AUC for compounds in the D series. This time, only D-Model I performed best with both solvation models. Therefore, this might be the most likely docking mode adopted by compounds in the D series.

Figure 8 shows how the ligand interacted with the protein in the lowest-energy structures for B-Model I and D-Model I as Model I was among the highest performance structural models for both series of compounds. Table S2 gives key protein-ligand interactions for these two predicted structures. The salicylic ring docked into the phosphotyrosine-binding pocket for both docking structures. In B-Model I (Figure 8A), the hydroxyl group formed hydrogen bonds with the main-chain NH groups of three P-loop residues (Arg-404, Ala-405 and Gly-406). One oxygen of the carboxylic group interacted with the guanidinium side chain of Arg-409 and the other oxygen hydrogen-bonded with the amino side chain of Gln-450 and the backbone NH group of Gln-357. The nitrogens in the triazole-ring interacted with the side-chain oxygen of Gln-357 and formed two hydrogen bonds with the terminal amino group of Lys-447. In addition, the oxygen and nitrogen from the amide hydrogen-bonded with the side-chain NH group of Arg-205 and the main-chain carbonyl group of Gln-446. One phenyl ring of the ligand situated in a pocket formed inside a bundle of three α -helices ($\alpha 1$, $\alpha 6$, and $\alpha 7$) and the other phenyl ring protruded into solution. The latter phenyl ring might be removed in the future to reduce the size of the compounds so that other more useful functional groups can be introduced elsewhere.

The structure of D-Model I demonstrated with ligand D09 (Figure 8B) had the phenylmorpholine ring tightly bound inside the

Table 5: Binding affinity of ligands in the D series obtained by using the $\epsilon(r) = 4r$ and the GBMV solvation models for the four major docking modes identified from stage 1 docking

D-Series ligands	$\epsilon(r) = 4r$ implicit model					GBMV implicit model				
	I	II	III	IV	Mix ^a	I	II	III	IV	Mix ^a
D01	-74.16	-73.98	-71.58	-76.34	-76.34	-55.71	-55.87	-44.90	-65.22	-65.22
D02	-74.30	-76.32	-68.61	-76.05	-76.32	-54.21	-54.86	-49.46	-71.45	-71.45
D03	-79.29	-75.19	-75.34	-79.26	-79.29	-54.90	-54.52	-50.88	-71.94	-71.94
D04	-71.30	-75.45	-66.67	-74.34	-75.45	-61.99	-53.63	-48.55	-65.15	-65.15
D05	-75.46	-77.03	-77.65	-77.26	-77.65	-54.97	-52.22	-53.53	-65.88	-65.88
D06	-78.47	-78.31	-78.92	-80.92	-80.92	-58.26	-59.44	-53.43	-69.02	-69.02
D07	-77.10	-77.45	-73.22	-78.66	-78.66	-61.00	-51.38	-49.14	-69.16	-69.16
D08	-86.62	-77.26	-83.61	-84.12	-86.62	-59.30	-52.73	-58.06	-68.04	-68.04
D09	-93.19	-79.02	-82.65	-79.48	-93.19	-64.11	-53.73	-50.25	-62.47	-64.11
D10	-90.97	-79.97	-76.31	-84.04	-90.97	-62.40	-53.84	-55.71	-71.16	-71.16
D11	-81.01	-80.60	-80.25	-81.79	-81.79	-63.64	-56.57	-52.17	-68.55	-68.55
D12	-82.27	-77.31	-78.87	-80.79	-82.27	-62.48	-53.74	-54.49	-68.98	-68.98
D13	-74.28	-79.96	-73.60	-77.40	-79.96	-59.33	-56.48	-48.11	-66.77	-66.77
D14	-85.38	-84.20	-84.75	-83.37	-85.38	-62.85	-58.98	-48.63	-66.94	-66.94
D15	-85.38	-84.42	-83.60	-87.25	-87.25	-59.40	-58.84	-52.85	-71.53	-71.53
D16	-79.63	-77.41	-78.88	-81.96	-81.96	-62.25	-54.19	-61.03	-68.41	-68.41
D17	-82.93	-82.49	-83.42	-85.34	-85.34	-59.82	-54.61	-60.20	-71.98	-71.98
D18	-78.20	-79.81	-78.29	-80.97	-80.97	-59.19	-50.64	-63.65	-70.45	-70.45
D19	-81.75	-75.44	-77.40	-79.79	-81.75	-59.16	-51.03	-55.37	-67.62	-67.62
D20	-69.63	-72.09	-62.05	-73.18	-73.18	-61.09	-45.48	-36.44	-57.23	-61.09
TP/(TP + TN)	4/7	2/7	2/7	3/7	4/7	6/7	1/7	3/7	2/7	2/7
Se (%)	57	29	29	43	57	86	14	43	29	29
Sp (%)	77	62	62	69	77	92	54	69	62	62
Acc (%)	70	50	50	60	70	90	40	60	50	50
EF	1.63	0.82	0.82	1.22	1.63	2.45	0.41	1.22	0.82	0.82

^aIn the mixed model, not all ligands were assumed to adopt the same major docking mode. Instead, the binding affinity was obtained from the mode that yielded the most favorable binding affinity.

Bold: the seven ligands with the most favorable binding affinity for each structural model.

TP, true positive; TN, true negative; Se, sensitivity; Sp, Specificity; Acc, accuracy; EF, enrichment factor.

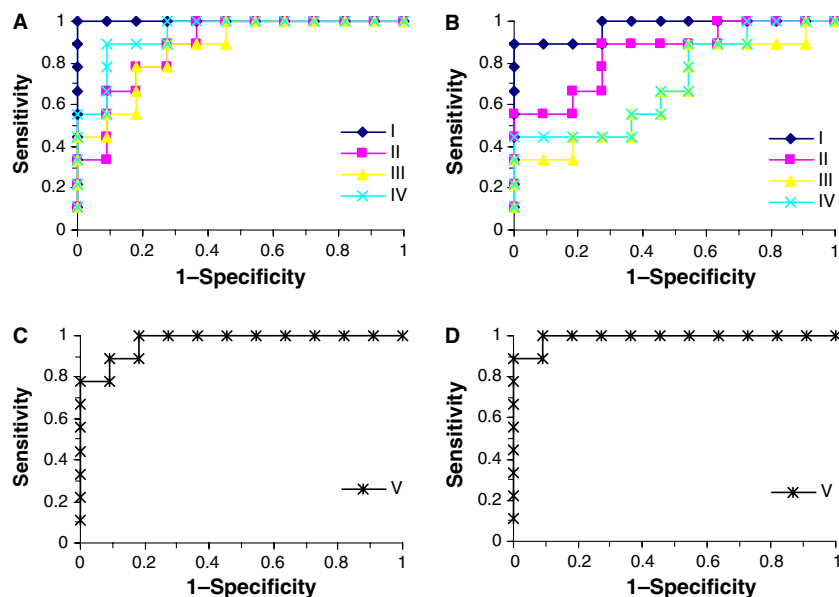


Figure 6: Receiver operating characteristics curves obtained by using four major docking modes and two solvation models for compounds in the B series. (A) B-Model I, II, III, and IV with the $\epsilon = 4r$ solvation model, (B) B-Model I, II, III, and IV with the GBMV model, (C) mixed structural model with the $\epsilon = 4r$ solvation model, (D) mixed structural model with the GBMV solvation model.

bundle of three helices: $\alpha 1$, $\alpha 6$, and $\alpha 7$. The hydroxyl group and the carboxylate groups of the salicylic acid core formed extensive hydrogen bonds with the side chains of Arg-409, Lys-447 and the

backbone NH groups of Gln-357, Arg-404 and Ala-405. Similar to B-Model I, the oxygen in the furan ring hydrogen bonded with NH2 of Arg-404. On the other hand, the three nitrogens of the triazole-ring

Table 6: Area under curve (AUC) for receiver operating characteristics (ROC) plots

B or D Model	B Series		D Series	
	$\epsilon(r) = 4r$	GBMV	$\epsilon(r) = 4r$	GBMV
Model I	1	0.97	0.71	0.81
Model II	0.88	0.85	0.52	0.53
Model III	0.87	0.67	0.55	0.52
Model IV	0.94	0.71	0.55	0.45
Mixed	0.97	0.99	0.69	0.45

formed hydrogen bonds with the side chains of Gln-357 and Lys-447. The amide interacted with Arg-205 and Gln-446 while the morpholine ring interacted with the side chains of Asp-448 and Arg-437.

Table 7 summarizes the AUC values obtained from the eight different rigid-protein docking models using Autodock described above. The best AUC values were smaller than those obtained

Table 7: Area under curve (AUC) for ROC plots from rigid-protein docking using Autodock

B/D Model	B Series	D Series
Model I	0.94	0.60
Model II	0.90	0.62
Model III	0.72	0.42
Model IV	0.57	0.51
1QZ0	0.75	0.37
1PA9	0.89	0.48
1YPT	0.66	0.52
Mixed	0.77	0.49

ROC, receiver operating characteristics.

from the flexible-receptor models, suggesting that the rigid-protein model did not perform as well as the flexible-protein model. It is therefore better to use the best docking modes obtained from flexible-receptor docking to guide future lead optimization.

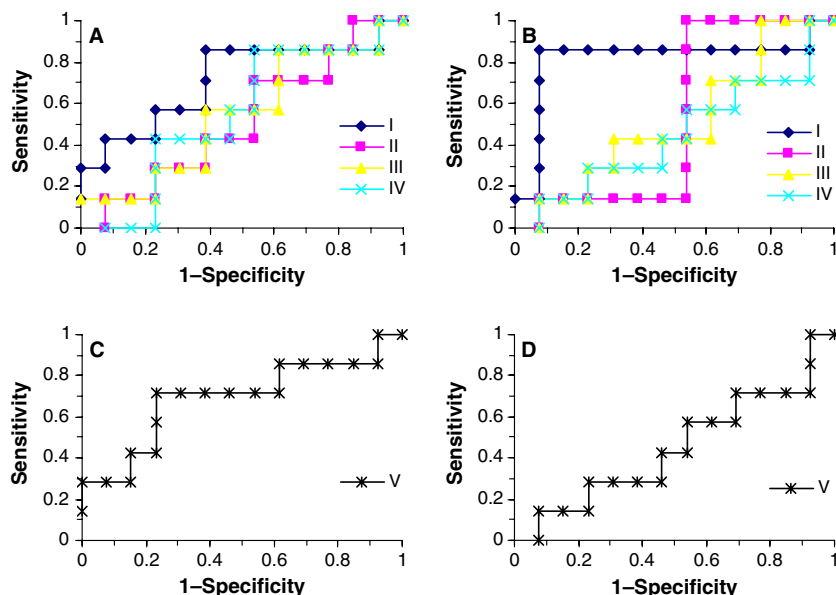


Figure 7: Receiver operating characteristics curves obtained by using four major docking modes and two solvation models for compounds in the D series. (A) D-Model I, II, III, and IV with the $\epsilon = 4r$ solvation model, (B) D-Model I, II, III, and IV with the GBMV model, (C) mixed structural model with the $\epsilon = 4r$ solvation model, (D) mixed structural model with the GBMV solvation model.

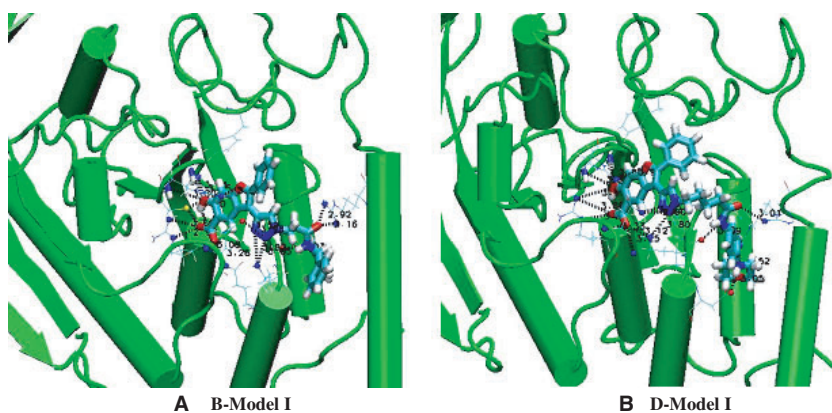


Figure 8: Key interactions occurring at protein-ligand interface. (A) B-Model I (YopH-B17 system). (B) D-Model I (YopH-D09 system). Coloring scheme: protein in green, oxygen in red, nitrogen in blue, carbon in cyan, and sulfur in yellow. The picture was generated by VMD (39).

Conclusions

In this work, we have developed two series of forty compounds derived from the salicylic acid core and found 16 to have micromolar inhibition activity against YopH. The initial design strategy relied on introducing two chemical moieties, linked together by a flexible hydrocarbon chain, to target two pockets in the active site of the protein. We used the salicylic acid moiety to target the pocket that phosphotyrosine bind and tried different chemical entities on the opposite end to target a nearby secondary pocket.

To predict how these compounds bind to YopH, we have started with flexible ligand-flexible protein docking using two different solvation models: $\epsilon(r) = 4r$ and GBMV. The docking suggested four possible docking modes, three had the salicylic acid core bound to the phosphotyrosine-binding pocket and one had the other end of the ligands bound instead. As the docking models yielded similar energy for these docking modes, it was difficult to single out the best docking mode. We therefore used each one of these docking modes to perform binding affinity calculations to examine which docking modes gave results most consistent with experimental IC50. We also considered a mixed structural model in which not all ligands were required to bind to the same binding mode. Instead, the docking mode – of four major docking modes identified from flexible-receptor docking – that yielded the most favorable binding affinity was taken. Different performance analysis such as calculating accuracy and area under receiver operating characteristics curve suggested that compounds in the B series might prefer a binding mode (B-Model I) in which all ligands bound with the salicylic acid core situated in the phosphotyrosine-binding pocket. However, the mixed model that allowed different ligands to take on different ones of the four possible major binding modes was also possible. On the other hand, compounds in the D series preferred a binding mode similar to B-Model I – i.e., D-Model I – in which all ligands bound in roughly the same way to the protein with minor adjustments at the protein–ligand interface.

We had also performed rigid-receptor docking using Autodock but the performance was not as good as the molecular dynamics-based flexible-receptor docking. The two best docking models found for the B series (B-Model I and the mixed model) and the best model found for the D series found from flexible-receptor docking might thus be the best to use in the immediate future to guide future optimization of these two series of compound, before experimental structures of protein–ligand complexes are available.

Acknowledgments

This research was supported by a Research Board Award from the University of Missouri System, and the National Institutes of Health (Grants AI071991 and CA126937). We also thank the University of Missouri Bioinformatics Consortium and the University of Missouri-Saint Louis Information Technology Services for providing computational resources.

References

1. Bulter T. (1985) Textbook of Medicine. Philadelphia, PA: W. B. Saunders.
2. Guan K.L., Dixon J.E. (1990) Protein tyrosine phosphatase activity of an essential virulence determinant in *Yersinia*. *Science*;249:553–556.
3. Stuckey J.A., Schubert H.L., Fauman E.B., Zhang Z.-Y., Dixon J.E., Saper M.A. (1994) Crystal structure of *Yersinia* protein tyrosine phosphatase at 2.5 Å and the complex with tungstate. *Nature*;370:571–575.
4. Sun J.-P., Wu L., Fedorov A.A., Almo S.C., Zhang Z.-Y. (2003) Crystal structure of the *Yersinia* protein-tyrosine phosphatase YopH complexed with a specific small molecule inhibitor. *J Biol Chem*;278:33392–33399.
5. Bölin L., Wolf-Watz H. (1988) The plasmid-encoded Yop2b protein of *Yersinia pseudotuberculosis* is a virulence determinant regulated by calcium and temperature at the level of transcription. *Mol Microbiol*;2:237–245.
6. Michiels T., Cornelis G.R. (1988) Nucleotide sequence and transcription analysis of yop51 from *Yersinia enterocolitica* W22703. *Microb Pathog*;5:449–459.
7. Straley S.C., Bowmer W.S. (1986) Virulence genes regulated at the transcriptional level by Ca²⁺ in *Yersinia pestis* include structural genes for outer membrane proteins. *Infect Immun*;51:445–454.
8. Bliska J.B., Clemens J.C., Dixon J.E., Falkow S. (1992) The *Yersinia* tyrosine phosphatase: specificity of a bacterial virulence determinant for phosphoproteins in the J774A.1 macrophage. *J Exp Med*;176:1625–1630.
9. Bliska J.B., Guan K.L., Dixon J.E., Falkow S. (1991) Tyrosine phosphate hydrolysis of host proteins by an essential *Yersinia* virulence determinant. *Proc Natl Acad Sci USA*;88:1187–1191.
10. Phan J., Lee K., Cherry S., Tropea J.E., Burke, T.R. Jr., Waugh D.S. (2003) High-resolution structure of the *Yersinia pestis* protein tyrosine phosphatase YopH in complex with a phosphotyrosyl mimetic-containing hexapeptide. *Biochemistry*;42:13113–13121.
11. Liang F., Huang Z.H., Lee S.-Y., Liang J., Ivanov M.I., Alonso A., Bliska J.B., Lawrence D.S., Mustelin T., Zhang Z.Y. (2003) Aurintricarboxylic acid blocks *in vitro* and *in vivo* activity of YopH, an essential virulent factor of *Yersinia pestis*, the agent of plague. *J Biol Chem*;278:41734–41741.
12. Tautz L., Bruckner S., Sareth S., Alonso A., Bogetz J., Bottini N., Pellicchia M., Mustelin T. *et al.* (2005) Inhibition of *Yersinia* tyrosine phosphatase by furanyl salicylate compounds. *J Biol Chem*;280:9400–9408.
13. Hu X., Stebbins C.E. (2005) Molecular docking and 3D-QSAR studies of *Yersinia* protein tyrosine phosphatase YopH inhibitors. *Bioorg Med Chem*;13:1101–1109.
14. Galimand M., Guiyoule A., Gerbaud G., Rasoamanana B., Chanteau S., Carniel E., Courvalin P. (1997) Multidrug resistance in *Yersinia pestis* mediated by a transferable plasmid. *N Engl J Med*;337:677–680.
15. Guiyoule A., Gerbaud G., Buchrieser C., Galimand M., Rahalison L., Chanteau S., Courvalin P., Carniel E. (2001) Transferable plasmid-mediated resistance to streptomycin in a clinical isolate of *Yersinia pestis*. *Emerg Infect Dis*;7:43–48.

16. Sarmiento M., Wu L., Keng Y.-F., Song L., Luo Z., Huang Z., Wu G.Z., Yuan A.K., Zhang Z.Y. (2000) Structure-based discovery of small molecule inhibitors targeted to protein tyrosine phosphatase 1B. *J Med Chem*;43:146–155.
17. Huang Z., Wong C.F. (2007) A mining-minima approach to exploring the docking pathways of p-nitrocatechol sulfate to YopH. *Biophys J*;93:4141–4150.
18. Huang Z., Wong C.F., Wheeler R.A. (2008) Flexible Protein-Flexible Ligand Docking with Disrupted Velocity Simulated Annealing. *Proteins: Struct Funct Bioinform*;71:440–454.
19. Kolb H.C., Finn M.G., Sharpless K.B. (2001) Click chemistry: diverse chemical function from a few good reactions. *Angew Chem Int Ed Engl*;40:2004–2021.
20. Lee M.S., Feig M., Salsbury F.R. Jr, Brooks C.L. III (2003) New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J Comput Chem*;24:1348–1356.
21. Lee M.S., Salsbury F.R. Jr, Brooks C.L. III (2002) Novel generalized Born methods. *J Chem Phys*;116:10606–10614.
22. Chocholoušová J., Feig M. (2006) Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations. *J Comput Chem*;27:719–729.
23. Huang Z., Wong C.F. (2009) Conformational selection of protein kinase A revealed by flexible-ligand flexible-protein docking. *J Comput Chem*;30:631–644.
24. MacKerell A.D. Jr, Bashford D., Bellott M., Dunbrack R.L. Jr, Evanseck J.D., Field M.J., Fischer S. *et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*;102:3586–3616.
25. MacKerell A.D. Jr, Feig M., Brooks C.L. III (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem*;25:1400–1415.
26. Advanced Chemistry Development in (2006) ACD/ChemSketch Freeware. Toronto, ON.
27. Brooks B.R., Bruccoleri R.E., Olafson B.D., States D.J., Swaminathan S., Karplus M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*;4:187–217.
28. Hu X., Stebbins C.E. (2006) Dynamics of the WPD Loop of the *Yersinia* protein tyrosine phosphatase. *Biophys J*;91:948–956.
29. Pedretti A., Villa L., Vistoli G. (2004) VEGA – an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. *J Comput Aided Mol Des*;18:167–173.
30. Carpenter G.A., Grossberg S. (1987) ART 2: self-organization of stable category recognition codes for analog input patterns. *Appl Opt*;26:4919–4930.
31. Feig M., Karanicolas J., Brooks C.L. III (2004) MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J Mol Graph Model*;22:377–395.
32. Karpen M.E., Tobias D.J., Brooks C.L. III (1993) Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry*;32:412–420.
33. Pao Y.H. (1989) Adaptive Pattern Recognition and Neural Networks. New York: Addison Wesley.
34. Morris G.M., Goodsell D.S., Halliday R.S., Huey R., Hart W.E., Belew R.K., Olson A.J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem*;19:1639–1662.
35. Goodsell D.S., Morris G.M., Olson A.J. (1996) Automated docking of flexible ligands: applications of AutoDock. *J Mol Recognit*;9:1–5.
36. Triballeau N., Acher F., Brabet I., Pin J.-P., Bertrand H.-O. (2005) Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J Med Chem*;48:2534–2547.
37. Kirchmair J., Markt P., Distinto S., Wolber G., Langer T. (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes. *J Comput Aided Mol Des*;22:213–228.
38. Jacobsson M., Lidén P., Stjernschantz E., Boström H., Norinder U. (2003) Improving structure-based virtual screening by multivariate analysis of scoring data. *J Med Chem*;46:5781–5789.
39. Humphrey W., Dalke A., Schulten K. (1996) VMD: visual molecular dynamics. *J Mol Graph*;14:33–38.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Protein residues allowed to move in Stage 2 protein-ligand docking.

Table S2. Key protein-ligand interactions for one representative compound in the B series and one in the D series.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.