

## *E-Novo*: An Automated Workflow for Efficient Structure-Based Lead Optimization

Bradley C. Pearce,<sup>\*,§</sup> David R. Langley,<sup>§</sup> Jia Kang,<sup>‡,§</sup> Hongwei Huang,<sup>†</sup> and Amit Kulkarni<sup>†</sup>

Bristol-Myers Squibb, Computer-Assisted Drug Design, 5 Research Parkway, Wallingford, Connecticut 06492, and Accelrys, 200 Wheeler Road, South Tower, Second Floor, Burlington, Massachusetts 01803

Received February 25, 2009

An automated *E-Novo* protocol designed as a structure-based lead optimization tool was prepared through Pipeline Pilot with existing CHARMM components in Discovery Studio. A scaffold core having 3D binding coordinates of interest is generated from a ligand-bound protein structural model. Ligands of interest are generated from the scaffold using an R-group fragmentation/enumeration tool within *E-Novo*, with their cores aligned. The ligand side chains are conformationally sampled and are subjected to core-constrained protein docking, using a modified CHARMM-based CDOCKER method to generate top poses along with CDOCKER energies. In the final stage of *E-Novo*, a physics-based binding energy scoring function ranks the top ligand CDOCKER poses using a more accurate Molecular Mechanics-Generalized Born with Surface Area method. Correlation of the calculated ligand binding energies with experimental binding affinities were used to validate protocol performance. Inhibitors of Src tyrosine kinase, CDK2 kinase,  $\beta$ -secretase, factor Xa, HIV protease, and thrombin were used to test the protocol using published ligand crystal structure data within reasonably defined binding sites. In-house Respiratory Syncytial Virus inhibitor data were used as a more challenging test set using a hand-built binding model. Least squares fits for all data sets suggested reasonable validation of the protocol within the context of observed ligand binding poses. The *E-Novo* protocol provides a convenient all-in-one structure-based design process for rapid assessment and scoring of lead optimization libraries.

### INTRODUCTION

A primary objective in Structure-Based Drug Design (SBDD) is the ability to assess protein–ligand binding energetics as a means for hit identification (virtual screening) and lead optimization (enhance desired drug properties). To achieve this one needs adequate sampling of ligand conformations and a way to accurately dock and score binding poses. It is becoming increasingly popular to use traditional ligand docking methods in combination with more accurate physics-based rescoring as a SBDD tool.<sup>1</sup> The physics-based implicit solvation models, molecular mechanics – Poisson–Boltzmann surface area (MM-PBSA) and molecular mechanics – generalized Born surface area (MM-GBSA) methods, compute relative and absolute ligand binding energies while requiring reasonable computer resources.<sup>2</sup> Many SBDD studies are focused on a congeneric series and are intended to navigate around existing patents or to simply model analogs for medicinal chemistry.<sup>3</sup> Experimental X-ray data often indicate analogues in a congeneric series position their scaffolds very similarly in the receptor or protein binding site.<sup>4,5</sup> These closely related binding poses mean a full docking experiment is unnecessary and actually increases the chance of getting a wrong pose. In these instances, developing a fast and easy way to enumerate the fragment-scaffold combination using the bound scaffold 3D coordi-

nates becomes essential. Previously, we had described an automated *de novo* SBDD workflow using Pipeline Pilot (PP) components as implemented within Discovery Studio (DS).<sup>6</sup> Within this protocol, AutoLudi was used as a *de novo* library generation and docking tool, CHARMM refinement provided top binding poses, and a physics-based scoring method was implemented. CHARMM (Chemistry at HARvard Macro-molecular Mechanics) provides a vast range of functionality for molecular mechanics and can be applied to diverse areas of research, including protein modeling, SBDD, structural biology, and nanotechnology.<sup>7</sup> Generally this earlier workflow did not perform well since AutoLudi lacked interaction site information resulting in missing some potential fragment candidates for library enumeration. Hence we developed a new workflow/protocol that can produce a fast enumeration of analogues, followed by core-constrained refinement and physics-based scoring. The key difference in this study is initial placement of the scaffold, close to the correct binding pose. To our knowledge this method of constraining the scaffold has not been reported in the literature. In order to distinguish from the previous automated *de novo* workflow, we named this new protocol “*E-Novo*” (*E*numerated *d*e *N*ovo Design).

While preparing this manuscript a group at Amgen published a related study.<sup>8</sup> The Amgen group evaluated Glide XP docking with MM-GBSA rescoring as a lead optimization tool. A diverse set of pharmaceutically relevant targets was used to validate their method that included CDK2, factor Xa, thrombin, and HIV-RT. The Amgen study demonstrated a high level of success correlating calculated binding energies

\* Corresponding author phone: (203)677-6904; e-mail: bradley.pearce@bms.com.

<sup>§</sup> Bristol-Myers Squibb.

<sup>‡</sup> Current address: Department of Translational Informatics & Department of Statistics, Yale University, 300 George Street, New Haven, CT 06510.

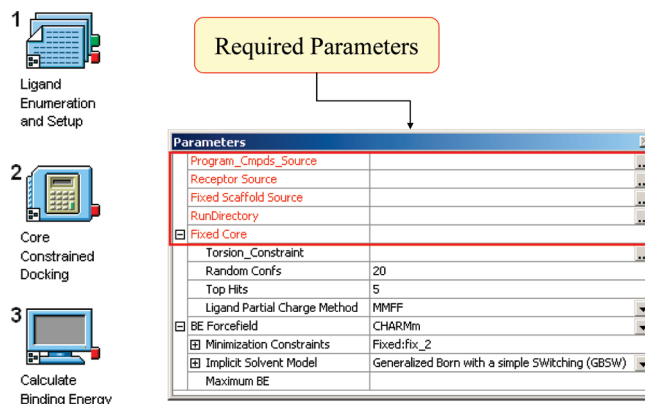
<sup>†</sup> Accelrys.

to experimental binding affinities within these examples. The author's method may provide an alternative to more rigorous but computationally expensive methods as free-energy perturbation and thermodynamic integration or even intermediate level methods such as empirically derived linear interaction energy models.<sup>9</sup> Key differences between the Amgen work and *E-Novo* include the docking method (Glide XP vs core-constrained CDOCKER), the minimization routine, and force fields (Embrace-OPLS vs CHARMM) along with some rescoring energy terms. In addition, the Amgen study evaluated differences in a single conformation versus conformer ensemble averages to represent each ligand in the unbound state for comparison. As a means for comparison and additional protocol validation we used the Amgen data sets of congeneric compounds.

*E-Novo* accomplishes these processes in three basic steps as an "all-in-one" protocol. The first step of ligand enumeration can be implemented through enumerating molecules from core and substituents with attachment points in Pipeline Pilot. The second step performs conformer generation, core-constrained docking, and scoring. In terms of core-constrained docking, the molecular docking algorithm CDOCKER is considered since it has been shown to be a viable research tool.<sup>10,11a</sup> CDOCKER is a CHARMM-based grid-enabled docking method that uses soft core potentials, molecular dynamics (MD) generated random ligand conformations, and pose refinement in the active site using a simulated annealing process. In the original work, CDOCKER treats the entire ligand as flexible during the initial docking phase. An alternate method to CDOCKER called SDOCKER was developed, which utilizes a 3D similarity template to better mimic the crystal structure pose.<sup>12</sup> The results from SDOCKER indicate docking accuracy improvements of 25–50% over CDOCKER for the three cases studied. The core-constrained docking method described here is a modification to the CDOCKER CHARMM script that allows the scaffold to be locked but the rest of the ligand molecule to be flexible during initial minimization and the MD conformer generation stage. The core is allowed to move during final refinement with simulated annealing. Current docking tools do a reasonable job at getting correct poses, but errors occur in a significant number of cases.<sup>13,14</sup> Constraining the core to the crystal structure coordinates helps prevent incorrect docking poses. In the third step, *E-Novo* uses CHARMM-based pose rescoring with the MM-PB(GB)SA method. An *E-Novo* workflow, designed for automated library enumeration, core-constrained docking, and potentially more accurate physics-based rescoring for structure-based lead optimization routines is described in this work. Validation studies carried out to examine the performance of the protocol are discussed. We expected *E-Novo* can reduce computation time on virtual screening of designed libraries to a practical level.

## METHODS

**(1.1). Protein and Fixed Scaffold Preparation.** The protein–ligand structures, which are typically determined by X-ray crystallography, NMR, homology modeling, or through hand built models based on photoaffinity labeling and/or mutagenesis data are imported into Discovery Studio (DS) and processed. The protein is split from any ligand, oligo-

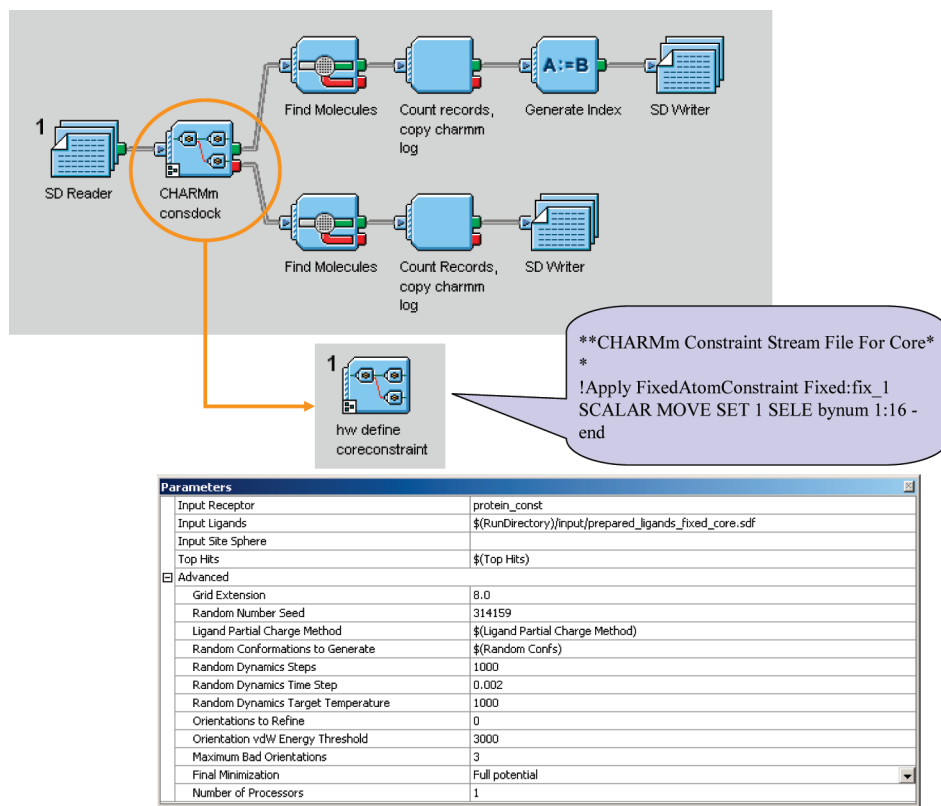


**Figure 1.** *E-Novo* lead optimization protocol and exposed top level parameters. Ligand preparation, core-constrained docking, and calculate binding energy protocol stages and required top level parameters.

meric chains, water molecules, and/or solvent. The apo protein is run through the "clean" command from the "Protein Modeling" DS toolbox, which standardizes the protein and includes adding hydrogens, fixing missing side chains, and adjusting residue protonation states to a user specified pH [7.0] and is saved as the receptor.pdb file. The ligand is extracted from the initial input pdb structure file and processed into a scaffold R-group query (RG) file.<sup>15</sup>

**(1.2). E-Novo Lead Optimization Protocol.** The "E-Novo Lead Optimization" protocol (*E-Novo*) contains the entire package for ligand/receptor preparation, CHARMM docking using CDOCKER and physics-based ligand rescoring. This protocol can be run after initial manipulations of the receptor as described in section 1.1, through either the Pipeline Pilot (PP) or DS client. The *E-Novo* PP protocol can be downloaded from the Accelrys Community Web site, where detailed software documentation and tutorials for PP and DS can also be obtained.<sup>16</sup> The *E-Novo* protocol contains three distinct subprotocols, labeled 1–3, as shown in Figure 1. The subprotocols run sequentially, saving appropriate files during run time, so that sections of the protocol can be rerun using differing downstream parameter sets, facilitating data optimization. At the top level, required parameters (Figure 1) include file sources for the program compounds, receptor protein, fixed core template, a designated run directory, and the core atoms to be held fixed during constrained docking. Program compounds may be from screening results or a generated virtual library or a combination of both. The program compounds should be in the form of an ISIS structure data file<sup>17</sup> having a molecule ID field and, if desired, an appropriate biological data field, typically expressed as pIC<sub>50</sub>, pEC<sub>50</sub>, or pK<sub>i</sub> values. The fixed scaffold core should be prepared in ISIS RG file format (section 1.1) and have atom numbering specified to that in the editable "Fixed Core" parameter. The fixed core parameter specification is "first atom number": "last atom number" and generally refers to the heavy atom count.

**(1.3). Ligand Enumeration and Setup (LES) Sub-protocol.** The *E-Novo* LES subprotocol prepares the ligands for subsequent docking. A 2D structure data file containing ligands of interest and optional biological data are processed to fully enumerate tautomer and ionization states. The "Enumerate Ionization States" component ionizes acidic and basic residues to a desired pH range, in these cases the pH



**Figure 2.** Core-constrained docking pipeline. Generates ligand conformers and performs core-constrained docking of prepared ligands from step 1 and outputs top scoring CDOCKER poses.

= 7. The ionization enumerator relies on a lookup table within PP and is intended to provide an estimate of possible ionization states. However, one has to examine this on an individual basis and make corrections or adjustments as necessary. These corrections can be included within the subprotocol as SMIRKS,<sup>18</sup> which specifies a particular molecular transformation or series of transformations. Both the ionization and tautomer enumerator can be turned on or off depending on the study at hand. The enumerated 2D molecules are converted into 3D structures and minimized for cleanup using the “clean force field” within PP.<sup>19</sup> A “Generate RGroups” component then fragments R-groups from the ligands based on the attachment points defined in the template RG file. The R-groups are then reattached to the scaffold core, in all possible combinations using an “RG reader” component. Importantly, these structures now have their cores aligned to the crystal structure geometry. By joining on canonical smiles generated from both the 2D enumerated structures and those generated from the RD reader, only unique desired molecules are retained. A structure data file is written into the input directory having the fully prepared ligands for downstream docking and scoring.

**(1.4). Core-Constrained Docking (CCD) Subprotocol.** The second CCD subprotocol (step 2, Figure 1) generates diverse conformations of the ligands from step 1 and performs core-constrained docking with the protein of interest. The subprotocol pipeline and its exposed parameters are shown schematically in Figure 2. This process occurs through a lengthy number of embedded subprotocols and scripts that the end-user can largely ignore. Core-constrained docking can be achieved by adding constraint command lines into the CDOCKER CHARMm script as illustrated in Figure 2.

A grid is defined using the “Grid Extension” parameter visible at the top level of the subprotocol and has been set to a default 8 Å distance from the ligand’s center of mass. The user can vary these parameters as shown in Figure 2.

Details of the modified CDOCKER process begin with creation of the CHARMm setup files and generation of the CDOCKER protein grid. Ligand partial atomic charges and atom types default to those of Momany-Rone force field<sup>20</sup> as implemented in CHARMm. The typed ligand is first run through an ABNR minimization stage.<sup>7</sup> Ligand conformations are then generated, in the absence of protein, through high-temperature molecular dynamics (MD) simulations, and a specified number (top level parameter, default = 20) of simulations are applied. The starting ligand conformation for each MD simulation is that of its predecessor. Conformations resulting from each MD simulation are then docked into the protein and minimized using steepest-descent (SD), preparing them for final refinement. It is only during this initial docking and conformer refinement phase where an energy grid for the ligand is imposed and nonbonded interactions involving vdW and electrostatic potentials are softened, allowing enhanced sampling of conformational space. The core is held fixed throughout the conformer generation and docking phase. With the core now unconstrained, the docked poses are then further refined in the receptor active site using a simulated annealing protocol and a full molecular mechanics minimization (SD + conjugate-gradient). During this entire process the protein is held rigid. A user specified number of top poses (top level parameter), based on the largest minus CDOCKER scores, are saved for the final rescoring step. Many of the advanced CHARMm parameters have been optimized and do not require changing from their default values.



**(1.5). Calculate Binding Energy (CBE) Subprotocol.**

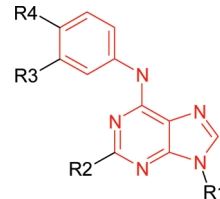
The CDOCKER docked ligands are rescored using a physics-based implicit solvation model as the final step in Figure 1. Within the CBE subprotocol step, the docked ligand poses are rank scored in terms of their energies of binding. For this study, top CDOCKER poses of neutral and/or charged ligands are rescored using Molecular Mechanics-Generalized Born with Molecular Volume (MM-GBMV) and/or Molecular Mechanics-Generalized Born with Simple Switching (MM-GBSW) methods in DS CHARMM, which approximates the binding energy.<sup>21–23</sup> Bound and unbound ligand–receptor energy terms contained within the calculated binding energy include three simulations: free ligand, apo protein, and protein–ligand complex. Solute entropy contributions are ignored in these calculations. Standard output includes the binding energy terms for the three simulations and the net calculated binding energy as shown in eq 1

$$\Delta\Delta G_{\text{Bind}} = \Delta G_{\text{Complex}} - \Delta G_{\text{Ligand}} - \Delta G_{\text{Protein}} \quad (1)$$

**(1.6). Inhibitors of Src Tyrosine Kinase, CDK2 Kinase, B-Secretase, Factor Xa, HIV Protease, and Thrombin.** The Src kinase data represent a scaffold having four points of diversity within a fairly narrow binding pocket.<sup>24</sup> The protein–ligand complexes 2BDJ.pdb and 2BDF.pdb were downloaded from the RCSB PDB<sup>25</sup> and prepared as described in section 1.1. For the Src kinase docking and MM-GBSA rescoring results we fixed the core at 1:16 (16 heavy atoms, Table 1 highlighted in red), generated 20 random conformations keeping the top 5 scoring CDOCKER poses, and used default ligand partial atom charges, CHARMM force field, and Generalized Born with simple SWitching solvent model. The tautomer enumerator was turned off in the first subprotocol since all core tautomers (3 for each) resulted in longer run times and poorer fitting data for the alternate tautomers.

Inhibitors of a related ATP binding site domain kinase, CDK2, were evaluated with three diversity points. In this case we looked at the congeneric series defined in Table 1 of the Amgen publication.<sup>8,26</sup> The protein–ligand complexes 1FVT.pdb and 1E9H.pdb were downloaded from the PDB and prepared as described in section 1.1. In addition, we made a similar hand-modeled adjustment as did the Amgen researchers to accommodate the correct protein state. For the CDK2 kinase docking and MM-GBSA rescoring results we fixed the core at 1:22, as highlighted in red in Supporting Information Table 2. The remaining parameters were identical to those described for Src kinase.  $\beta$ -Secretase represented a congeneric series of peptidomimetic inhibitors with two points of diversity.<sup>27</sup> The protein–ligand complex 2P4J.pdb was downloaded from the PDB and prepared as described in section 1.1. For the  $\beta$ -secretase docking and MM-GBSA rescoring results we fixed the core at 1:35, as highlighted in red in Supporting Information Table 3. The remaining parameters were identical to those described for Src kinase.

Another series for comparison to the results from the Amgen data is that described for factor Xa.<sup>8,28</sup> The protein–ligand complex 1FJS.pdb was downloaded from the PDB and prepared as described in section 1.1. For the factor Xa docking and MM-GBSA rescoring results we fixed the core at 1:26, as highlighted in red in Supporting Information Table 4. The remaining parameters were identical to those described for Src kinase.

**Table 1.** *E-Novo* Results From Src Tyrosine Kinase Inhibitors<sup>a</sup>


MOL_ID	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	MW	MM-GBSW Calculated $\Delta G$ kcal/mol	Experimental pIC <sub>50</sub>
1			Cl	H	388.9	-38.42	6.6
2			Cl	H	346.8	-38.38	5.8
3			Cl	H	424.9	-48.93	7.6
4			H	H	390.4	-48.51	7.6
5			H	H	399.5	-51.48	7.5
6			H		475.5	-53.45	9.3
7			H		470.4	-41.30	5.7
8			H		523.5	-41.88	6.6
9			H		494.4	-43.06	7.2

<sup>a</sup> Core-constrained scaffold (red highlight) with R-group enumeration, MOL\_ID, molecular weight (MW), MM-GBSW calculated  $\Delta G$  binding energies in kcal/mol, and experimental pIC<sub>50</sub> =  $-\log_{10}(\text{IC}_{50} \text{ M})$ .

A congeneric series of HIV protease inhibitors with two points of diversity was examined as an example of a protein binding site well-known for its plasticity.<sup>29</sup> The protein–ligand complexes 2PQZ.pdb, 2QNP.pdb, 2QNQ.pdb, 2PWR.pdb, 2PWC.pdb, and 2QNN.pdb were downloaded from the PDB and prepared as described in section 1.1. For this study we used the 2PQZ.pdb complex. It should be noted that all these ligands exhibit very similar binding poses, having a maximal rmsd of 1.3 Å between ligands 2PQZ and 2QNN. For the HIV protease docking and MM-GBSA rescoring results we fixed the core at 1:25, as highlighted in red in Supporting Information Table 5. The remaining parameters were identical to those described for Src kinase.

The final comparison with the Amgen study was a set of thrombin inhibitors having three points of diversity.<sup>30–32</sup> The protein–ligand complex 1ETT.pdb was downloaded from the PDB and prepared as described in section 1.1. For the thrombin docking and MM-GBSA rescoring results we fixed the core at 1:17, as highlighted in red in Supporting Information Table 6. The remaining parameters were identical to those described for Src kinase with the exception where the charged ionization state was used since this gave slightly better results than the neutral state.

**(1.7). RSV Validation Model.** Molecular Dynamic model of the RSV/RSV-17 complex: A model of the N-28 RSV 28 heptad repeat trimer/RSV-17 (RSV inhibitor) complex was constructed based on the previously described model

for the RSV N-terminus complex with BMS-433771 (**RSV-17**).<sup>33</sup> Details of this model are reported here. To build this model the N-28 heptad repeat trimer (amino acids 180–207) was excised from the 1G2C.pdb structure,<sup>34</sup> and three molecules of **RSV-17** were hand docked into the three hydrophobic cavities near A:Tyr:198, B:Tyr:198, and C:Tyr:198. The trimer complex was neutralized by the addition of 6 chloride ions and solvated in a  $46.2 \times 55.9 \times 60.4 \text{ \AA}^3$  box containing 4355 TIPS3<sup>35</sup> water molecules. The system was then minimized using the conjugate gradient algorithm implemented in NAMD<sup>36</sup> in a two step procedure. The system was minimized for 250 steps with the protein backbone atoms fixed. This was followed by an additional 250 steps of minimization with C $\alpha$  atoms harmonically restrained ( $k=1 \text{ kcal/mol/\AA}^2$ ). Heating was conducted by running MD for 125 ps with C $\alpha$  atoms restrained while increasing the system temperature in 10 degree increments every 500 timesteps (1 time step = 2 fs) until the system temperature reached 298 K. In the remaining heating phase, the system temperature was maintained by velocity rescaling at every 500 steps. This was followed by a 125 ps equilibration with all atoms free to move. The production simulation was carried out with no restraints for approximately 5 ns. Both equilibration and production simulations were run in the NPT ensemble using the Langevin piston method at 1 atm pressure and maintained at 298 K using heat bath coupling as implemented in NAMD. Equations of motion were integrated using the Verlet algorithm. The Amber94 parameter set<sup>37</sup> was used to describe the protein, ions, and water. The General Amber Force Field (GAFF)<sup>38</sup> with AM1\_BCC charges<sup>39</sup> was used to describe **RSV-17**. The medoid structure from the third nanosecond was used as the template for the *E-Novo* studies described in this paper. For the RSV docking and MM-GBSA rescoring results we fixed the core at 1:23, as highlighted in red in Supporting Information Table 7. The remaining parameters were identical to those described for Src kinase.

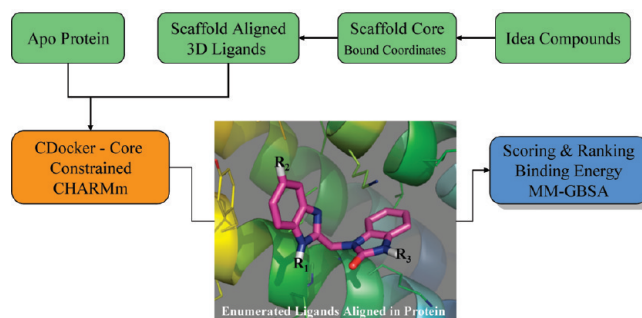
All of the relevant components used in this project are available with Discovery Studio 1.7 and Pipeline Pilot 6.02. The components were accessed and further customized through the Pipeline Pilot client.

## RESULTS AND DISCUSSION

**(2.1). *E-Novo* Workflow.** The *E-Novo* workflow for fast lead optimization using CHARMM-based components within Pipeline Pilot or Discovery Studio is shown schematically in Figure 3. The workflow consists of three basic steps: 1) ligand enumeration from 3D scaffold, 2) constrained docking using CDOCKER, and 3) MM-GBSA rescoring. Starting with a ligand-bound receptor, the complex is split into the apo protein and ligand, retaining structural 3D geometry. The receptor protein typically requires some preparation, which may include removal of extraneous ions, solvent molecules, repair of incomplete residues, addition of hydrogens, and adjustment of residue charges. In our experience the protein preparation is very important and depending on how this is done can lead to vastly different results. Removal of the attached groups from the ligand prepares the scaffold core as an R-group query (RG) template. The scaffold RG-template first defines a set of unique R-groups from idea/program compounds and second it prepares the generated

### Workflow Steps

1. Ligand enumeration from 3D scaffold
2. Constrained docking using CDOCKER
3. MM-GBSA for rescoring



**Figure 3.** *E-Novo* for fast lead optimization workflow.

conformers for core-constrained docking using defined 3D coordinates.

The second *E-Novo* step involves conformational sampling of the ligands and core-constrained docking using CDOCKER as implemented in CHARMM. Differences here reside mainly in modifications to the CDOCKER CHARMM scripts that allow the core to be constrained during both initial minimization and conformational sampling. This is a unique feature of *E-Novo* and was part of our design strategy. During this initial phase, conformational sampling is performed outside the protein using high-temperature MD simulations. The authors cited some of their earlier work where automated MD docking using soft-core potentials had shown greater efficiency and accuracy to Monte Carlo and genetic algorithm searching methods.<sup>11b</sup> Thus, we used the CDOCKER MD conformational search method. The default setup for a single MD simulation is 1000 steps at 2 fs per step and at 1000 K. The number of MD simulations is defaulted to 20 as a compromise in thoroughness and efficiency. In general we have found only slight improvements increasing to 30 simulations. In their original CDOCKER work the authors found little improvement in using greater than 20 replicas.<sup>11a</sup> Without rigorous torsional sampling it is conceivable that conformations are missed, which could reflect some of the observed outliers. We have not performed an extensive evaluation of this. Being outside the protein, no inherent restrictions are placed on the side chain conformations. This may result in some side chain clashes with protein residues. In order to avoid potential high energy clashes the conformers are core-constrained docked and minimized using steepest-descent (SD) in the receptor active site after imposing softened van der Waals and repulsive/attractive electrostatic terms within the defined grid. This should facilitate greater conformational sampling. In the final simulated annealing pose refinement stage, the core is unlocked, and a full molecular mechanics force field minimization is performed for each generated conformation. The top docked poses based on computed CDOCKER energy scores are carried into the final stage of *E-Novo*. The limitations of scoring functions within docking programs are well documented, and CDOCKER is no exception.<sup>14</sup>

The physics-based rescoring of top CDOCKER poses is performed in the final step as illustrated in Figure 3. As mentioned in the methods section 1.5, the binding energies are calculated in three simulations for that of free ligand,

apo protein, and ligand-protein complex. Components of the Gibbs free energy of binding in the bound state using MM-GBSA are shown in eq 2<sup>9</sup>

$$\Delta G_{bind} = \Delta G_{ele} + \Delta G_{np} + \Delta G_{conf} - T\Delta S_{solute} \quad (2)$$

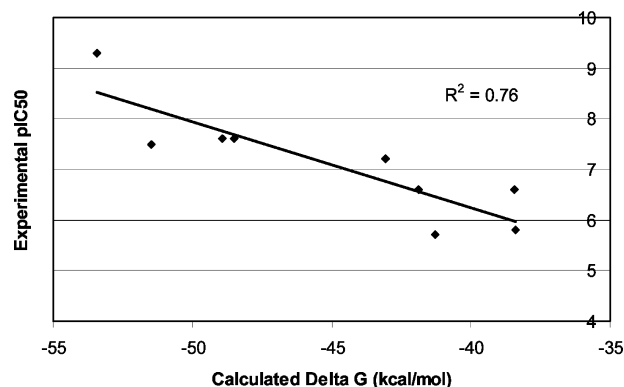
The  $\Delta G_{ele}$  electrostatic term incorporates both Coulombic interactions between ligand and receptor along with solute charges within the solvent. The  $\Delta G_{np}$  term contains the nonpolar contributions of van der Waals and hydrophobic interactions between ligand and receptor. The  $\Delta G_{conf}$  term evaluates internal strain realized upon binding of ligand with protein. Internal strain energy reflecting protein deformation upon ligand binding is omitted since the protein is held fixed in this protocol. The final term,  $-T\Delta S_{solute}$ , computationally describes the solute entropy that results in loss of rotational, translational, and vibrational freedoms with ligand binding and is omitted in our BE calculation. It is often assumed that within a congeneric series binding to the same protein, that the ligand entropy terms will be similar and effectively fall out when making relative comparisons. The MM-GBSA terms for the binding energy that *E-Novo* uses are a simplification of enthalpic and entropic terms, making the calculation an approximation of simple binding energy (BE) or  $\Delta G$ . Our objective was to make *E-Novo* practical for idea molecule selections, not a precise analytical tool for the rank scoring of ligands. We found that the CHARMM implemented MM-GBMV worked better for the RSV inhibitors and that the MM-GBSW worked better in all other examples. The MM-GBSW model is similar to the MM-GBMV model except that rather than calculating Born radii by analytic integration it uses a van der Waals-based surface with a smooth dielectric boundary.<sup>40</sup> Validation of the solvation model using empirical data is probably the best route if available. Examples of the MM-GBSW and MM-GBMV methods are listed in the Supporting Information tables.

To our knowledge this is the first publication where ligand preparation, conformer generation/core-constrained docking, and pose rescoring are performed in an all-in-one protocol to simplify structure-based lead optimization. Only three input files are required: the prepared apo protein, the prepared 3D scaffold, and a set of idea/program compounds for evaluation. Biological data are optional for model validation and can be included in the program input file. *E-Novo* is a reasonable method allowing for practical runtime computation of ligand docking within a congeneric series, enabling idea compound prioritization.

## (2.2). Protocol Validation Results and Discussion.

**Src Tyrosine Kinase.** Src tyrosine kinase is an enzyme that modulates numerous cellular signal transduction processes that are implicated in a number of disease states including cancer and osteoporosis. Crystal structures of two Src tyrosine kinase inhibitors, along with a series of analogues defining structure–activity relationships (SAR), having a reasonable pIC<sub>50</sub> bioactivity range, have been recently published by Dalgano and co-workers.<sup>24</sup> This represents a system of four R-groups as defined in the core shown in Table 1. This arrangement fixes both purine and aryl ring coordinates to those of the crystal-bound ligand within the classic ATP binding site as highlighted in red. This publication offered a case study where one has a narrowly defined binding pocket with good buried protein–ligand interactions.

## Correlation of Calculated Binding Energy with Experimental Src Kinase Inhibition Data

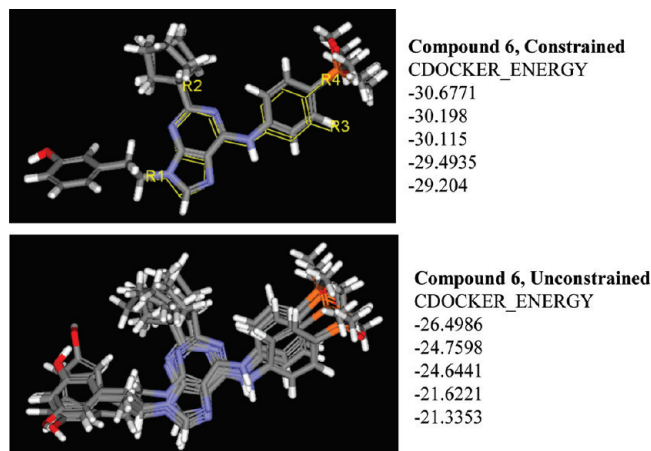


**Figure 4.** Correlation between MM-GBSW calculated binding energy and experimental pIC<sub>50</sub> data for Src kinase inhibitors. Correlation coefficient squared = 0.76.

The 2BDJ crystal structure is one protein chain with compound **6** (author nomenclature) bound. The 2BDF crystal structure represents two chains, both with compound **9** bound. These three ligand-bound systems provided comparisons. Using the same parameter sets with the three complexes, we found the 2BDJ and 2BDF chain B systems to be comparable and gave higher correlations of calculated  $\Delta G$  to the experimental pIC<sub>50</sub> than the 2BDF chain A system. Reported here are the results from the 2BDJ crystal structure. The 2BDJ.pdb was used as receptor and ligand source file and were processed and run in *E-Novo* as described in sections 1.1 and 1.6. An SAR table defining the compound core structure, MOL\_ID (using the published nomenclature), R-groups, molecular weight (MW), MM-GBSW calculated binding energy (BE) or  $\Delta G$ , and pIC<sub>50</sub> data is shown in Table 1. Within Supporting Information Table 2, nonpolar surface area (NPSA) [total surface area – polar surface area],<sup>16a</sup> CDOCKER binding energy scores in kcal/mol, and charged ligand MM-GBSW calculated  $\Delta G$  binding energies in kcal/mol are added for comparison. Compound **1** is Purvalanol A, which is an inhibitor of a related kinase CDK2, and was used to compare specificity differences. Compounds **2–6** represent a series of purine Src kinase inhibitors targeted for cancer metastases. Compounds **7–9** represent a series of related Src kinase inhibitors targeted for bone resorption and differ in having a water solubilizing phosphonomethylphosphinyl (PMP) side chain that points toward the solvent exposed domain of the protein.

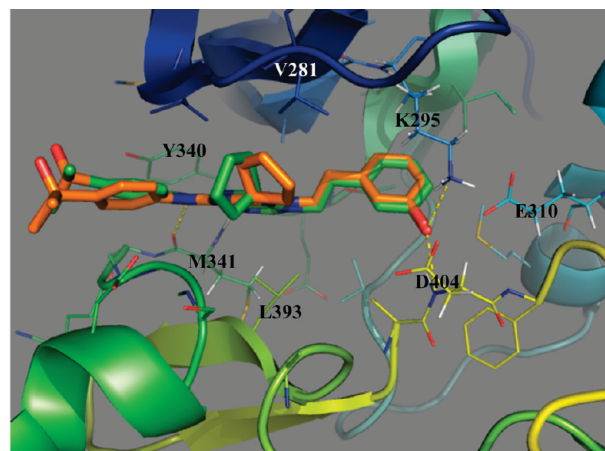
As can be seen from the data in Table 1 and the corresponding graphical representation in Figure 4, a reasonable correlation exists between the calculated BE and the experimental reported enzyme inhibition data. With the full data set a correlation coefficient squared ( $R^2$ ) of 0.76 was found using neutral ligands and the MM-GBSW rescoring function. Using charged ligands and the same scoring function an  $R^2$  of 0.69 was found. The higher correlation using neutral ligands is consistent with what Reynolds et al. had found with a series of  $\beta$ -secretase inhibitors.<sup>41</sup> Reynolds had carried out a systematic study of the relative contributions of van der Waals, Coulombic, and continuum-solvation contributions to ligand binding in the presence and absence of ligand charges using GBSA solvation models. With charged ligands they found large deviations in all these terms.



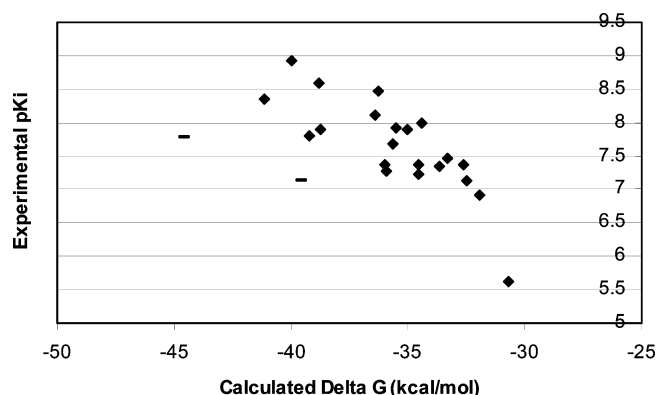


**Figure 5.** Constrained CDOCKER top poses of Src kinase compound **6** (top panel) and purine core (yellow) and unconstrained CDOCKER top poses compound **6** (bottom panel) and their CDOCKER Scores.

They found the van der Waals contribution best correlated with ligand binding and the Coulombic/solvation terms less correlated and sensitive to the protonation state of the ligand and protein. Their best correlations between calculated binding energies and experimental  $\beta$ -secretase binding affinities were found using neutral ligands. As controls, we observed no correlation of experimental  $IC_{50}$ s to either CDOCKER scores ( $R^2 = 0.09$ ) or molecular weight ( $R^2 = 0.04$ ). In the case of CDOCKER scores this is consistent with the results of Vieth et al.<sup>11</sup> where they concluded that CDOCKER optimizes the energy of the system to a similar extent for all the ligands regardless of pose. However, they found a fairly good Spearman Rho correlation with MW of the ligands. In our case, the MW did not correlate well and is likely a result of these ligands having differing specific protein residue interactions. However, reasonable correlation is observed between the calculated nonpolar surface area (NPSA) and experimental  $IC_{50}$ s ( $R^2 = 0.67$ ). This result shows the importance of the nonpolar energy contributions to the calculated binding energy in this case. For all other examples there was little to no correlation of NPSA to experimental binding affinities (Supporting Information tables). Displayed in Figure 5 are overlays of compound **6** showing the top 5 poses and associated CDOCKER energy values, done with both constrained and unconstrained docking. Also shown in the upper panel is an overlay of the scaffold (yellow) with the crystal structure geometry. As is evident from Figure 5, some movement of the core does occur during the final minimization stage, but the movement is far greater when the core is not constrained during initial docking. In particular, the low rmsd variation of the phenolic ring [ $R_1$ ] is apparent and reflects a strong H-bonding interaction with protein residues. Core variation for unconstrained docking is also evident in the greater variability and less negative CDOCKER energy values. The overlays in Figure 5 underscore the importance of core-constrained docking and reinforces our objective to minimize poor docking poses. Shown in Figure 6 is an overlay of the top MM-GBSW scoring CDOCKER pose of compound **6** (orange carbons) and its crystal structure bound conformation (green carbons). *E-*Novo** correctly establishes the classic purine H-bond network with Met341. In addition, H-bond donation of Lys295 and Asp404 (backbone NH) with the



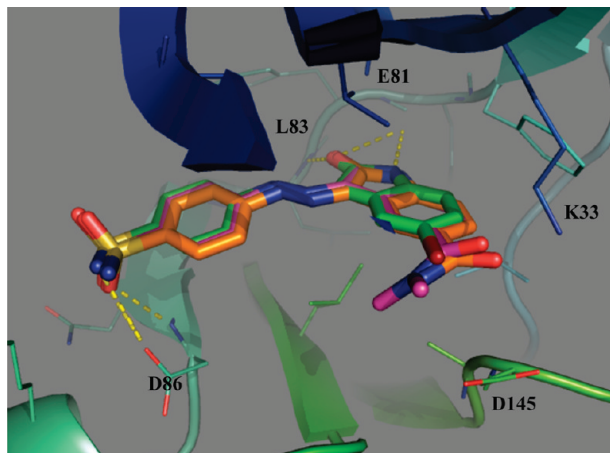
**Figure 6.** Top scoring Src kinase MM-GBSW CDOCKER pose (orange carbons) along with the crystal structure for compound **6** (green carbons), showing H-bonding of the purine nucleus to Met341 and phenolic OH to Lys295, Glu310, and Asp404.



**Figure 7.** Correlation between MM-GBSW calculated binding energy and experimental  $pK_i$  data for CDK2 kinase inhibitors. All data points, correlation coefficient squared = 0.36. Removal of (-) outliers,  $R^2 = 0.63$ .

phenolic hydroxyl along with the H-bond acceptor interaction with Glu310 is also evident. Variability in the conformation of the appended cyclopentane ring at  $R_2$  reflects weak hydrophobic interactions with protein residues Val281 and Leu393. The PMP residue is solvent exposed, and compound **9** exhibits a strong H-bonding network to Tyr340 via a crystal-bound water. These interactions are not explicitly captured by *E-*Novo** and could lead to some BE errors.

**CDK2 Kinase.** Similar to Src kinases, CDK2 kinases are implicated in cell regulatory cycles and are targets for pharmaceutical modulation. As a comparator to the Amgen data we used the oxindole-based series of CDK2 inhibitors described by Kuyper et al.<sup>8,26</sup> The *E-*Novo** results are shown in Supporting Information Table 3. The correlation of BE and experimental  $pK_i$  is shown in Figure 7. In all ligand docking poses *E-*Novo** places the oxindole core in the proper enzyme binding motif. Key H-bonding interactions include the indole NH with the backbone Glu81 C=O, oxindole C=O with the backbone NH of Leu83, sulfonamide S=O to the backbone NH of Asp86, and the sulfonamide NH to CO<sup>-</sup> of Asp 86 (Figure 8). Using all 23 data points the correlation coefficient squared is 0.36. There were 2 significant outliers with compounds **18** and **21**. Removing compounds **18** and **21** increases the  $R^2$  to 0.63, illustrating the tightness of data. From Figure 8 it is apparent that one carboxamide N-Me within compound **18** points into Asp145,

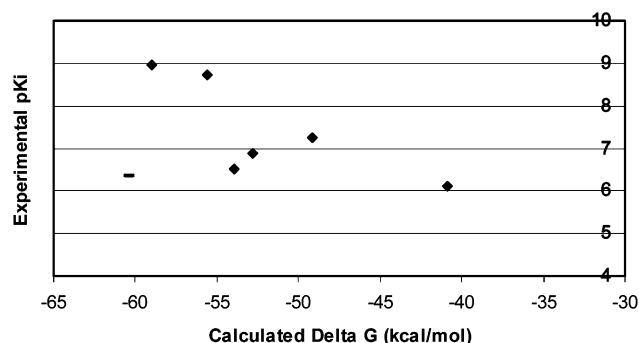


**Figure 8.** Top scoring CDK2 MM-GBSW CDOCKER poses (compound **18** purple carbons, compound **17** orange carbons) along with the crystal structure for compound **10** (green carbons), showing H-bonding of the oxindole nucleus to Glu81 and Leu83 along with H-bonding to Asp86. Proximity of Asp145 to N-Me of compound **18** is evident.

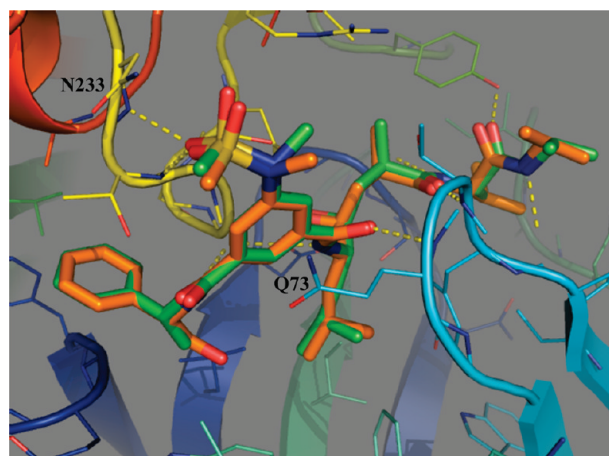
likely displacing solvated waters in this region and may explain the over estimation of BE. Interestingly, the outlier number **16** described by Amgen fits our data points fine. This compares favorably with the Amgen results where they reported an  $R^2 = 0.64$ . Compound **7** is the least active both experimentally and calculated and is a modest outlier. Compound **7** is an aromatic nitro compound, which in our experience poorly correlates using MMFF. The ability of nitro-aryl containing ligands to polarize in the presence of charged and/or polarizable protein residue side chains can lead to significant errors in atom partial charge estimations. In this case the nitro-aryl ring is adjacent to Lys33 and Phe80. It is likely that QM, QM/MM, or linear scaling semiempirical QM methods may contribute ligand-binding energy contributions accounting for specific polarization effects.<sup>42–44</sup> In this example, selecting the top 12/23 compounds by MM-GBSW rescoring would have provided 75% of the actual most actives for a library synthesis.

**$\beta$ -Secretase.**  $\beta$ -Secretase is an aspartyl protease that has been implicated in the pathogenesis of Alzheimer's disease. Ghosh and co-workers evaluated a congeneric series of hydroxyethylene dipeptide isosteres.<sup>27</sup> These molecules are highly flexible, so it was anticipated that core-constrained docking would be particularly beneficial. The *E-Novo* results are shown in Supporting Information Table 4. The correlation of BE and experimental  $pK_i$  is shown in Figure 9. In all ligand docking poses *E-Novo* places the hydroxyethylene isostere in the same pose as found in the crystal structure, maintaining key protein–ligand binding interactions, Figure 10. We did observe some ligand complexes where the sulfonamide was turned 120 degrees capturing an alternate H-bond to Gln73. There were 1–2 outliers in this series, where all data points gave a correlation coefficient of 0.21. Removing compound **5f** increased the  $R^2$  to 0.58 and removing **5e** further increased the  $R^2$  to 0.80. The crystal structure indicates a crystallographic water molecule in proximity to the  $\alpha$ -Me benzyl moiety of compound **5d**. A partial explanation for the overestimation of binding affinity for the hydroxymethyl analogue **5f** is a desolvation penalty not accounted for by *E-Novo*. Similarly with **5e** a desolvation penalty may be in play along with rotation of the sulfonamide

**Correlation of Calculated Binding Energy with Experimental  $\beta$ -Secretase Inhibition Data**



**Figure 9.** Correlation between MM-GBSW calculated binding energy and experimental  $pK_i$  data for  $\beta$ -secretase inhibitors. All data points, correlation coefficient squared = 0.21. Removal of (–) outlier,  $R^2 = 0.58$ .



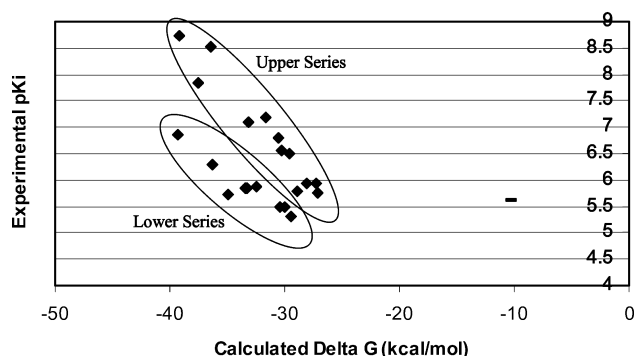
**Figure 10.** Top scoring  $\beta$ -secretase MM-GBSW CDOCKER pose for compound **5f** (orange carbons) along with the crystal structure for compound **5d** (green carbons). Both show extensive protein residue H-bonding networks.

moiety. In addition, the MMFF parametrization assigns  $sp^2$  hybridization to the sulfonamide nitrogen apparent in Figure 10, which being planar could introduce additional errors given its central role in protein binding.

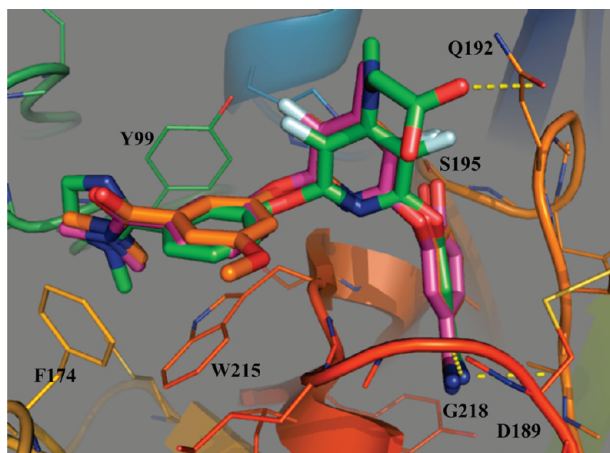
**Factor Xa.** A key enzyme involved in blood coagulation is factor Xa making it an attractive drug discovery target.<sup>28</sup> Significant synthetic efforts have been accomplished toward this target with hundreds of analogues prepared. However, in order to maintain comparisons to the Amgen data we used their compound series from Table 3.<sup>8</sup> It should be noted that this series used a slightly different template within the central aryl ring of the **1FJS** ligand, which may influence protein structure. The *E-Novo* results are shown in Supporting Information Table 5. The correlation of BE and experimental  $pK_i$  is shown in Figure 11. The drug binding pocket is intriguing from a modeling perspective as shown in Figure 12. Shown in Figure 12 is the crystal bound **1FJS** ligand (green carbons), the top scoring MM-GBSW CDOCKER pose for compound **15** (orange carbons), and the most potent compound **18** (purple carbons), which arguably rescues equivalent to compound **15**. The basic benzamidine moiety forms a salt bridge to Asp189 and H-bonds to the backbone C=O of Gly218 in the S1 pocket. The **1FJS** ligand has a hydroxyl moiety para to the amidine that engages a key H-bond donation with Ser195, forming part of the catalytic



### Correlation of Calculated Binding Energy with Experimental Factor Xa Inhibition Data



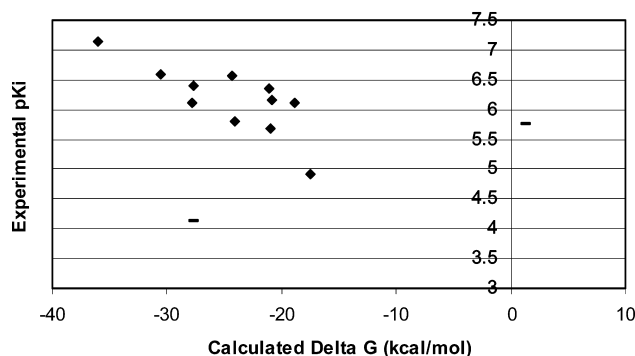
**Figure 11.** Correlation between MM-GBSW calculated binding energy and experimental  $pK_i$  data for factor Xa inhibitors. All data points, correlation coefficient squared = 0.29. Removal of (–) outlier,  $R^2 = 0.43$ . The data further delineate into an “upper” and “lower” series, which have correlation coefficients squared of 0.93 and 0.91, respectively.



**Figure 12.** Top scoring factor Xa MM-GBSW CDOCKER pose for compound **15** (orange carbons), most potent compound **18** (purple carbons) along with the **1FJS** ligand crystal structure (green carbons). Amidine salt bridge to Asp189 and H-bond to Gly218 backbone C=O. Unique **1FJS** H-bonding interactions to Gln192. Polarizable hydrophobic pocket defined by residues Phe174, Trp215, and Tyr99.

triad. In addition, a nest of crystallographic water molecules is found in this pocket. Bridging the S1 and S3/S4 binding pockets is a 1,3-diaryloxy-pyridyl ring that was held constant in this series. The S3/S4 binding pocket is a highly polarizable hydrophobic region. The **1FJS** ligand shows multiple near perfect edge-to-face and face-to-face  $\pi$ -stacking interactions with aromatic residues Phe174, Tyr99, and Trp215. In the **1FJS** ligand there is a 2-Me-3,4-dihydroimidazol-1-yl ring that is orthogonal to the aromatic ring setting up these near perfect interactions. Compound **18** is that reported in Figure 8 of the Amgen publication and is the only ligand-protein pose where direct comparison could be made to the poses in Figure 12. Core-constrained CDOCKER and GlideXP provide very similar poses. With *E-*Novo** there were 1–2 outliers in this series, where all data points gave a correlation coefficient squared of 0.29. Removing compound **5** increased the  $R^2$  to 0.43. However, of greater interest is a clear delineation of two series, call them a lower and upper, shifted by approximately 4 kcal/mol (Figure 11). The lower series exhibits an  $R^2$  of 0.91, and if compound 10 is

### Correlation of Calculated Binding Energy with Experimental HIV Protease Inhibition Data

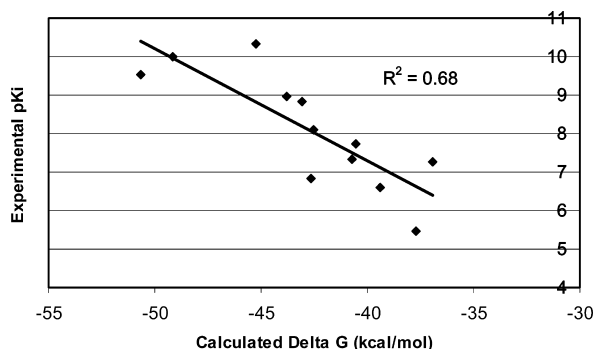


**Figure 13.** Correlation between MM-GBSW calculated binding energy and experimental  $pK_i$  data for HIV protease inhibitors. All data points, correlation coefficient squared = 0.09. Removal of (–) outliers,  $R^2 = 0.60$ .

removed the  $R^2$  jumps to  $>0.98$ . The upper series exhibits an  $R^2$  of 0.93 with no outliers. Can we explain these outliers and the divergence in series from a modeling perspective? Compound **5** is another example of a nitro aromatic in a highly polarizable region. Again, it is likely that a QM calculation of atom partial charges would better account for the BE of this compound. However, with a difference of 20 kcal/mol additional sources of error are apparent. Compound **10** is a trifluoromethoxy substituted aromatic and is modeled in the plane of the ring, which is contrary to both *ab initio* and experimental observations.<sup>45,46</sup> The lower series ring systems more closely overlay and perhaps most noting are the more planar  $R_4$  configurations, which contrast to those in the upper series where the  $R_4$  substituents more closely resemble the out of plane configuration found in the **1FJS** ligand. The nonplanarity of the  $R_4$  substituent resolves steric clashes in the binding site in cases of N,N-dimethyl substitution. Thus, these two tracks may represent divergent congeneric series within this set of compounds, as interpreted by the MM-GBSW rescoring function. These systematic differences may reflect ligand strain associated with a fixed protein binding site along with an underestimation of  $\pi$ -stacking interactions as noted above. The Amgen results indicated an  $R^2 = 0.71$  for this data set, and they did not observe divergent series.

**HIV Protease.** This is a well studied  $C_2$ -symmetric aspartyl protease essential to replication of the HIV virus. We evaluated the congeneric series based on a pyrrolidine scaffold described by Diederich et al.<sup>29</sup> The *E-*Novo** results are shown in Supporting Information Table 6. The correlation of BE and experimental  $pK_i$  is shown in Figure 13. The results were slightly better using the neutral form of the pyrrolidine nitrogen. The drug binding pocket is flexible and subject to greater induced fit between protein and inhibitor. Given the rigid protein treatment by *E-*Novo** this provides an interesting test case. Unlike the more traditional inhibitors based on statines, the basic pyrrolidine moiety forms a salt bridge to the catalytic dyad Asp25A/Asp25B. There were 2 significant outliers in this series, where all data points gave a correlation coefficient squared of 0.09. Removing compounds **6ab** and **6dd** increased the  $R^2$  to 0.60. Compound **6dd** is yet another example of a nitro aromatic where potential polarization effects are being ignored. Outlier **6ab**

Correlation of Calculated Binding Energy with Experimental Thrombin Inhibition Data



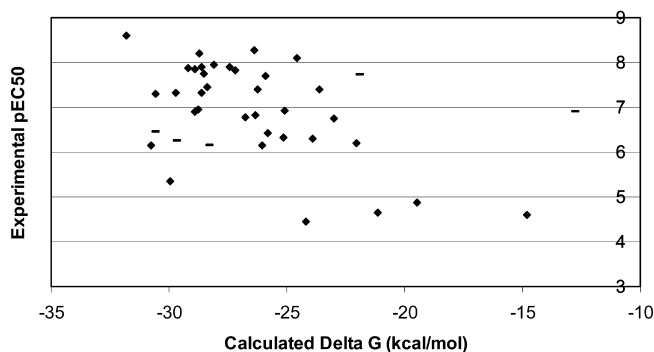
**Figure 14.** Correlation between MM-GBSW calculated binding energy and experimental  $pK_i$  data for thrombin inhibitors. All data points, correlation coefficient squared = 0.68.

is very hard to explain from a modeling perspective. These N-substituents fit into the hydrophobic  $S_1/S_1'$  binding pocket. Why the related N-allyl and N-prenyl derivatives differ significantly is unknown.

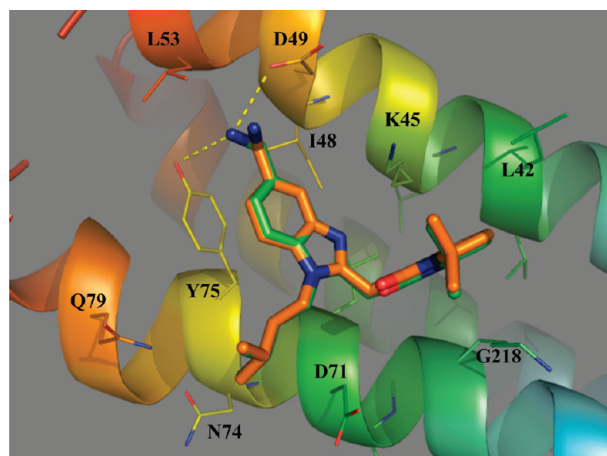
**Thrombin.** Another key enzyme in the blood coagulation cascade is the serine protease thrombin. This also represents the final series for comparison with the Amgen study.<sup>8,32</sup> The *E-Novo* results are shown in Supporting Information Table 7. The correlation of BE and experimental  $pK_i$  is shown in Figure 14. In this case the results were slightly better using the protonated form of the basic amidine. There were no significant outliers in this series, where all data points gave a correlation coefficient squared of 0.68. This compares favorably with the Amgen results where they reported an  $R^2 = 0.71$ .

**Respiratory Syncytial Virus (RSV).** RSV is an underestimated pathogen inflicting considerable mortality and morbidity worldwide. A significant medical need exists to find effective antiviral agents against RSV.<sup>47</sup> We recently described a series of RSV inhibitors that target the  $F_1$  subunit, which interfere with viral–host cell membrane fusion.<sup>33</sup> Based on photoaffinity labeling studies a protein–ligand pocket binding model was proposed using the published crystal structure of Zhao.<sup>34</sup> Our model, which is essentially that shown in Figure 3, has scaffold substituents  $R_1$  and  $R_2$  interacting directly with  $F_1$  protein residues. The  $R_3$  substituent projects into putative binding domains with the C-terminus heptad repeat strand which is not shown. Since direct protein binding interactions are necessary for this study, we chose to limit  $R_3$  to an isopropyl moiety which was widely used in the course of SAR development.<sup>48–53</sup> The ligands for this study were assembled from a corporate database containing all the RSV inhibitors (>1600 compounds) and were selected, without filtering, by using the  $R_3$  isopropyl containing template. In addition, to limit the number of compounds we only looked at 5-substituted benzimidazoles, which included the ligand from which the model was built. The RSV data were selected as a more challenging test of *E-Novo*, given the caveats of cell-based data and potential errors inherent in hand-built models. The *E-Novo* results are shown in Supporting Information Table 8. The correlation of BE and experimental  $pEC_{50}$  data for all 40 ligands was not inspiring having an  $R^2 = 0.19$  (Figure 15). Again, use of neutral ligands resulted in better correla-

Correlation of Calculated Binding Energy with Experimental Cell-Based RSV Inhibition Data



**Figure 15.** Correlation between MM-GBMV calculated binding energy and experimental  $pEC_{50}$  data for RSV inhibitors. All data points, correlation coefficient squared = 0.19. Removal of (–) outliers,  $R^2 = 0.39$ .



**Figure 16.** Top scoring RSV MM-GBMV CDOCKER pose for compound **RSV-17** (green carbons) along with the protonated form of **RSV-17** from the MD model (orange carbons). Benzylic ammonium salt bridging to Asp49 and H-bonding of the benzylic amino group to Tyr75 and Asp49.

tions, but in this case the GBMV solvation model worked better than GBSW. A more careful examination of the data shows five significant outliers that have binding pose issues. Legitimate removal of these five compounds increases the correlation coefficient squared to 0.39. There were four additional outliers not explained from the model. Removal of all 9 compounds yields a more respectable correlation coefficient of 0.62, but in practice this could not be done. Increasing the number of generated conformations at the MD stage from 20 (default) to 30 increased the  $R^2$  to 0.68 from 0.62. Increasing the number of MD simulations generally results in a slight tightening of data at the expense of modestly increased run times.<sup>54</sup> **RSV-17** is the most potent compound of the series and is correctly predicted as the most potent compound based on its MM-GBMV calculated  $\Delta G$  in the neutral form. In the model, the isopentyl side chain packs up against Asn74, Asp71, and Tyr75, and the benzyl amine is perfectly situated for hydrogen bonding to Asp49 and Tyr75. *E-Novo* correctly predicts this binding pose as shown in Figure 16 for both the neutral and charged forms. In the charged form, shown in Figure 16 as the MD model, the protonated benzylic amine forms a salt bridge to Asp49 and exhibits a dihedral rotation to accommodate this. In other

cases such as the R<sub>1</sub> amine-containing side chains there are differences between the neutral and charged forms in the 2–7 methylene linked amines. The binding poses with charged amines **RSV-1**, **RSV-14**, **RSV-18**, **RSV-20**, and **RSV-21** all show salt bridging to Asp71 at the expense of shifting the core and in cases of the longer linkages, pushing the side chains into the solvent exposed surface. With the neutral amines, hydrogen bonding dominates where possible and core variation is minimal. This may lead to an underestimation of the hydrophobic contributions of the side chains. The phosphonate (**RSV-7**) shows high binding affinity and commensurate RSV inhibition with the hydrophobic diethyl ester. The hydrophilic bis-acid phosphonate **RSV-8** is predicted to be less active along the lines of its experimentally observed bioactivity. However, the mono ester phosphonate **RSV-9** is a significant outlier having a bioactivity significantly below that predicted in both neutral and charged forms. Both **RSV-8** and **RSV-9** exhibit a similar binding motif having H-bonds to Gln79 in the charged form and H-bonds to Asp71 in the neutral form. **RSV-10** exhibits a significant shift in the core placement to accommodate the sterically demanding and hydrophobic tert-butyl ester and probably contributes to its lower than expected binding affinity. Along similar lines, the diphenyl methyl imine **RSV-24** also underestimates binding affinity. In this case the large diphenyl methyl moiety packs up against Leu42 but in order to do so pushes the core out as much as 4 Å. Without the ability of the protein to induce ligand fit errors such as this become more pronounced. The 5-ethyl derivative **RSV-27** is predicted to strongly bind to the protein with good hydrophobic interactions packing the ethyl group into Ile48 and Leu53. However, this compound is significantly less active than predicted. The mono substituted benzyl amine derivatives **RSV-32** and **RSV-33** and the geminal dimethyl derivative **RSV-40** exhibit vastly lower bioactivities than their parent **RSV-17** but exhibit similar binding poses having the Asp49 H-bond. Why these compounds are less active is not known but may relate to negative interactions with the putative binding site of the C-terminus heptad repeat strand. In all three cases an alkyl moiety is pushed into the protein-water interface. Given the decoupling between cell-based biology and structural chemistry the RSV results are understandable, but, again, a subset selection of library compounds for synthesis would have been largely successful in spite of some false positives.

## CONCLUSION

This workflow was generated for fast lead optimization by performing “all-in-one” *E-*Novo** design, including enumeration from scaffold/fragment, ligand conformer generation, core-constrained docking, and physics-based rescoring by CHARMM. The *E-*Novo** protocol can be run from Pipeline Pilot or within Discovery Studio. All that is required is a set of ligand compounds, a ligand template with docked 3D coordinates, and a protein of interest. The user can select the default parameters out of the box or experiment with using neutral versus charged ligands, enumerated ligand tautomers, conformational sampling, and continuum solvation models. Additional exposed or hidden parameters can be modified by drilling into the various subprotocols. Protocol validation was performed using diverse selections from

published examples, having appropriate crystallographic and biological assay data. In the examples investigated, the MM-GBSW method outperformed MM-GBMV with the exception of RSV. Outliers were few (RSV exception), with most of the data points exhibiting tight correlations between ligand experimental affinities and calculated binding energies. Whenever possible, systems should be validated with experimental data to select the most appropriate solvation rescoring model. In the absence of such data, using neutral ligands with the MM-GBSW rescoring method would be a reasonable first choice.

Where possible, comparisons to the Amgen results were made. *E-*Novo** provided similar results, with generally more outliers as might be expected given the differing methods. Ensemble averaging, flexible protein minimizations, and solute entropy contributions improve the results and help to minimize outliers as demonstrated by the Amgen studies. No consistent correlation trends were observed between experimental binding data and docking scores or other molecular descriptors such as nonpolar surface area and molecular weight, confirming the Amgen results. For example, nonpolar surface area will only correlate when these binding site interactions dominate to a significant extent (e.g., Src kinase). As with any molecular modeling results, the computational chemist needs to view the outcome in context with the observed binding poses. A number of the outliers we observed had binding pose issues or could be explained by inherent limitations of the molecular mechanics force fields. Ligand-protein polarization effects are not captured explicitly by the force field assigned atom partial charges and were noted in certain situations as contributing to BE errors. In the protocol's current form, highly polarizable ligand systems including nitro aromatics should be viewed with caution. Fundamental errors associated with force field parametrization may exist but were not evaluated using side-by-side comparisons in this study. Other potential sources of error include the use of a rigid protein throughout this process, where induced fit of the ligand is not taken into account. In our experience, performing bound state minimizations of crystal structures with subsequent extraction of the scaffold leads to similar or worse results, the likely outcome of protein overfitting. However, if one could run flexible protein MD or SD/PRCG minimizations on the top CDOCKER poses, prior to MM-GBSA rescoring, the results will likely improve. Effectively this is what the Amgen researchers did with their Embrace minimizations prior to rescoring. However, this does add to the overall experimental time frame. These solvation models are sensitive protonation states, whose uncertainty can introduce significant errors. Using ligand ensembles has the effect of smoothing the unbound ligand energies, and, as the Amgen results indicate, outliers are sometimes removed. However, it may be more prudent to evaluate the results in context of the observed binding poses and other data to decide reliability on an individual basis. Other sources of error include dismissal of the solute entropy terms although for relative BE comparisons this may not be a significant issue.<sup>9</sup> With modeled protein systems one always runs into caveats of structural waters, activation loops, and undefined protein domains.

Our results show that the *E-*Novo** docking and BE rescoring tool is an effective SBDD method in the seven test cases examined. The results were good for well-defined



ligand-protein active sites and expectedly less accurate with the RSV example. Given the weight of evidence suggesting compounds within a congeneric series bind similarly in the active site, *E-Novo*'s unique feature of core-constrained docking is beneficial. For most situations, *E-Novo* does a respectable job in guiding structure-based lead optimization efforts for the drug discovery scientist. Future efforts include the incorporation of flexible protein minimizations of docked poses and the recalculation of atom partial charges to better capture ligand polarization. In addition, prospective studies will be reported in due course.

#### ACKNOWLEDGMENT

Thanks to members of the BMS CADD group for helpful discussions, Ram Rajamani, Deborah Loughney, Arthur Doweiko, and Andrew Tebben. Special thanks to Roy Kimura and Brett Beno for manuscript review.

**Supporting Information Available:** SAR Tables 2–8 define core-constraints and contain *E-Novo* results for the examples in this study. This material is available free of charge via the Internet at <http://pubs.acs.org>.

#### REFERENCES AND NOTES

- Graves, A. P.; Shivakumar, D. M.; Boyce, S. E.; Jacobson, M. P.; Case, D. A.; Shoichet, B. K. Rescoring Docking Hit Lists for Model Cavity Sites: Predictions and Experimental Testing. *J. Mol. Biol.* **2008**, *377* (3), 914–934.
- Huang, N.; Jacobson, M. P. Physics-based methods for studying protein-ligand interactions. *Curr. Opin. Drug Discovery Dev.* **2007**, *10* (3), 325–331.
- Ji, H.; Zhang, W.; Zhang, M.; Kudo, M.; Aoyama, Y.; Yoshida, Y.; Sheng, C.; Song, Y.; Yang, S.; Zhou, Y.; Lue, J.; Zhu, J. Structure-Based de Novo Design, Synthesis, and Biological Evaluation of Non-Azole Inhibitors Specific for Lanosterol 14 $\alpha$ -Demethylase of Fungi. *J. Med. Chem.* **2003**, *46* (4), 474–485.
- Obst, U.; Banner, D. W.; Weber, L.; Diederich, F. Molecular recognition at the thrombin active site: Structure-based design and synthesis of potent and selective thrombin inhibitors and the X-ray crystal structures of two thrombin-inhibitor complexes. *Chem. Biol.* **1997**, *4* (4), 287–295.
- Günther, J.; Bergner, A.; Hendlich, M.; Klebe, G. Utilising structural knowledge in drug design strategies: Applications using relibase. *J. Mol. Biol.* **2003**, *326* (2), 621–636.
- Huang, H.; Kulkarni, A. Automated De Novo Design Workflow with Physics-Based Scoring Function for Fast Lead Identification and Optimization; Accelrys: San Diego, CA, 2007. <http://www.accelrys.com> (accessed May 2009).
- Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217.
- Guimarães, C. R. W.; Cardozo, M. MM-GB/SA rescoring of docking poses in structure-based lead optimization. *J. Chem. Inf. Model.* **2008**, *48* (5), 958–970.
- Foloppe, N.; Hubbard, R.; Foloppe, N. Towards predictive ligand design with free-energy based computational methods. *Curr. Med. Chem.* **2006**, *13* (29), 3583–3608.
- Brooks, C. III Assessing, improving and using grid-based docking algorithms in CHARMM. *Abstracts of Papers, Proceedings of the 233rd National Meeting and Exposition of the American Chemical Society*, Chicago, IL, Mar 25–29, 2007; Abstract COMP-250.
- (a) Wu, G.; Robertson, D. H.; Brooks, C. L., III; Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMM-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24* (13), 1549–1562. (b) Vieth, M.; Hirst, J. D.; Dominy, B. N.; Daigler, H.; Brooks III, C. L. Assessing Search Strategies For Flexible Docking. *J. Comput. Chem.* **1998**, *19* (14), 1623–1631.
- Wu, G.; Vieth, M. SDOCKER: A Method Utilizing Existing X-ray Structures To Improve Docking Accuracy. *J. Med. Chem.* **2004**, *47* (12), 3142–3148.
- Lyne, P. D.; Lamb, M. L.; Saeh, J. C. Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J. Med. Chem.* **2006**, *49* (16), 4805–4808.
- Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153*, (Suppl. 1).
- The ligand is edited in Discovery Studio (DS) to remove all appended groups [including hydrogens], leaving a basic core framework and adjustment of bond orders [aromatic bonds are designated as alternating double]. Save the file as a temporary.mol file. Import the core framework mol file into DS and use the add hydrogens feature from the chemistry menu and manually remove all hydrogens except those at the designated R-group positions. Save as the core-H.mol file. In order to get the correct core RG file format it is recommended that the RG file be processed using ISIS Draw although this can be done manually in a text editing program. The core mol file having the H for R-group placement is imported into ISIS Draw, and, without moving the molecule, the hydrogens are changed to the designated R-groups using the edit atom query tool. The 3D structure is then selected and exported as a ligand\_RG.mol file. This will create a scaffold with consecutive atom numbering in the core having the correct RG file format. *E-Novo* uses the atom numbering as defined in this fixed core. Results are sensitive to having the correct fixed atoms and their R-group vector alignment with the pdb ligand. Verify the atom numbering in a viewing program such as ISIS Draw, DS Visualizer, or DS ViewerPro. The order of the R-groups is unimportant.
- (a) *Pipeline Pilot, version 6.02*; Accelrys Inc.: San Diego, CA, 2007. (b) *Discovery Studio, version 1.7*; Accelrys Inc.: San Diego, CA, 2007. (c) Accelrys Inc. <http://forums.accelrys.org/> (accessed May 2009).
- Symyx. <http://www.mdli.com> (accessed May 2009).
- Daylight Chemical Information Systems Inc. [http://www.daylight.com/dayhtml\\_tutorials/languages/smirks/index.html](http://www.daylight.com/dayhtml_tutorials/languages/smirks/index.html) (accessed May 2009).
- Hahn, M. Receptor surface models. 1. Definition and construction. *J. Med. Chem.* **1995**, *38* (12), 2080–2090.
- Momany, F. A.; Rone, R. Validation of the general purpose QUANTA 3.2/CHARMM force field. *J. Comput. Chem.* **1992**, *13* (7), 888–900.
- Jayaram, B.; Sprou, D.; Beveridge, D. L. Solvation Free Energy of Biomacromolecules: Parameters for a Modified Generalized Born Model Consistent with the AMBER Force Field. *J. Phys. Chem. B* **1998**, *102* (47), 9571–9576.
- Massova, I.; Kollman, P. A. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA). to predict ligand binding. *Perspect. Drug Discovery* **2000**, *18* (Hydrophobicity and Solvation in Drug Design, Pt. II), 113–135.
- Srinivasan, J.; Cheatham, T. E., III; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. *J. Am. Chem. Soc.* **1998**, *120* (37), 9401–9409.
- Dalgarno, D.; Stehle, T.; Narula, S.; Schelling, P.; van Schravendijk, M. R.; Adams, S.; Andrade, L.; Keats, J.; Ram, M.; Jin, L.; Grossman, T.; MacNeil, I.; Metcalf, C., III.; Shakespeare, W.; Wang, Y.; Keenan, T.; Sundaramoorthi, R.; Bohacek, R.; Weigle, M.; Sawyer, T. Structural basis of Src tyrosine kinase inhibition with a new class of potent and selective trisubstituted purine-based compounds. *Chem. Biol. Drug Des.* **2006**, *67* (1), 46–57.
- RCSB Protein Data Bank. <http://www.rcsb.org/pdb> (accessed May 2009).
- Bramson, H. N.; Corona, J.; Davis, S. T.; Dickerson, S. H.; Edelstein, M.; Frye, S. V.; Gampe, R. T., Jr.; Harris, P. A.; Hassell, A.; Holmes, W. D.; Hunter, R. N.; Lackey, K. E.; Lovejoy, B.; Luzzio, M. J.; Montana, V.; Rocque, W. J.; Rusnak, D.; Shewchuk, L.; Veal, J. M.; Walker, D. H.; Kuyper, L. F. Oxindole-based inhibitors of cyclin-dependent kinase 2 (CDK2): Design, synthesis, enzymatic activities, and X-ray crystallographic analysis. *J. Med. Chem.* **2001**, *44* (25), 4339–4358.
- Ghosh, A. K.; Kumaragurubaran, N.; Hong, L.; Kulkarni, S. S.; Xu, X.; Chang, W.; Weerasena, V.; Turner, R.; Koelsch, G.; Bilcer, G.; Tang, J. Design, Synthesis, and X-ray Structure of Potent Memapsin 2 ( $\beta$ -Secretase). Inhibitors with Isophthalamide Derivatives as the P2-P3-Ligands. *J. Med. Chem.* **2007**, *50* (10), 2399–2407.
- Phillips, G.; Guilford, W. J.; Buckman, B. O.; Davey, D. D.; Eagen, K. A.; Koovakkat, S.; Liang, A.; McCarrick, M.; Mohan, R.; Ng, H. P.; Pinkerton, M.; Subramanyam, B.; Ho, E.; Trinh, L.; Whitlow, M.; Wu, S.; Xu, W.; Morrissey, M. M. Design, synthesis, and activity of a novel series of factor Xa inhibitors: Optimization of arylamidine groups. *J. Med. Chem.* **2002**, *45* (12), 2484–2493.
- Blum, A.; Böttcher, J.; Heine, A.; Klebe, G.; Diederich, W. E. Structure-guided design of C2-symmetric HIV-1 protease inhibitors based on a pyrrolidine scaffold. *J. Med. Chem.* **2008**, *51* (7), 2078–2087.
- Kim, S.; Hwang, S. Y.; Kim, Y. K.; Yun, M.; Oh, Y. S. Rational design of selective thrombin inhibitors. *Bioorg. Med. Chem. Lett.* **1997**, *7* (7), 769–774.

- (31) Oh, Y. S.; Yun, M.; Hwang, S. Y.; Hong, S.; Shin, Y.; Lee, K.; Yoon, K. H.; Yoo, Y. J.; Kim, D. S.; Lee, S. H.; Lee, Y. H.; Park, H. D.; Lee, C. H.; Lee, S. K.; Kim, S. Discovery of LB30057, a benzamidine-based selective oral thrombin inhibitor. *Bioorg. Med. Chem. Lett.* **1998**, *8* (6), 631–634.
- (32) Lee, K.; Jung, W. H.; Park, C. W.; Hong, C. Y.; Kim, I. C.; Kim, S.; Oh, Y. S.; Kwon, O. H.; Lee, S. H.; Park, H. D.; Kim, S. W.; Lee, Y. H.; Yoo, Y. J. Benzylamine-based selective and orally bioavailable inhibitors of thrombin. *Bioorg. Med. Chem. Lett.* **1998**, *8* (18), 2563–2568.
- (33) Cianci, C.; Langley, D. R.; Dischino, D. D.; Sun, Y.; Yu, K.-L.; Stanley, A.; Roach, J.; Li, Z.; Dalterio, R.; Colonna, R.; Meanwell, N. A.; Krystal, M. Targeting a binding pocket within the trimer-of-hairpins: Small-molecule inhibition of viral fusion. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (42), 15046–15051.
- (34) Zhao, X.; Singh, M.; Malashkevich, V. N.; Kim, P. S. Structural characterization of the human respiratory syncytial virus fusion protein core. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (26), 14172–7.
- (35) 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.* **1983**, *79* (2), 926–35.
- (36) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.
- (37) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–97.
- (38) 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.* **2004**, *25* (9), 1157–1174.
- (39) 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.* **2002**, *23* (16), 1623–1641.
- (40) Feig, M.; Onufriev, A.; Lee Michael, S.; Im, W.; Case David, A.; Brooks Charles, L. 3rd Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **2004**, *25* (2), 265–84.
- (41) Tounge, B. A.; Rajamani, R.; Baxter, E. W.; Reitz, A. B.; Reynolds, C. H. Linear interaction energy models for  $\beta$ -secretase (BACE) inhibitors: Role of van der Waals, electrostatic, and continuum-solvation terms. *J. Mol. Graphics Modell.* **2006**, *24* (6), 475–484.
- (42) Zhou, T.; Huang, D.; Cafilisch, A. Is quantum mechanics necessary for predicting binding free energy. *J. Med. Chem.* **2008**, *51* (14), 4280–4288.
- (43) Rajamani, R.; Good, A. C. Ranking poses in structure-based lead discovery and optimization: Current trends in scoring function development. *Curr. Opin. Drug Discovery Dev.* **2007**, *10* (3), 308–315.
- (44) Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. Importance of accurate charges in molecular docking: Quantum Mechanical/Molecular Mechanical (QM/MM) approach. *J. Comput. Chem.* **2005**, *26* (9), 915–931.
- (45) Kapustin, E. G.; Bzhezovsky, V. M.; Yagupolskii, L. M. Torsion potentials and electronic structure of trifluoromethoxy- and trifluoromethylthiobenzene: An ab initio study. *J. Fluorine Chem.* **2002**, *113* (2), 227–237.
- (46) Cambridge Structural Database, examples: AFABOY, CAVQEV, FUSTIW. <http://www.ccdc.cam.ac.uk/products/csd/> (accessed May 2009).
- (47) Sidwell, R. W.; Barnard, D. L. Respiratory syncytial virus infections: Recent prospects for control. *Antiviral Res.* **2006**, *71* (2–3), 379–390.
- (48) Yu, K.-L.; Zhang, Y.; Civiello, R. L.; Kadow, K. F.; Cianci, C.; Krystal, M.; Meanwell, N. A. Fundamental structure-activity relationships associated with a new structural class of respiratory syncytial virus inhibitor. *Bioorg. Med. Chem. Lett.* **2003**, *13* (13), 2141–2144.
- (49) Yu, K.-L.; Zhang, Y.; Civiello, R. L.; Trehan, A. K.; Pearce, B. C.; Yin, Z.; Combrink, K. D.; Gulgeze, H. B.; Wang, X. A.; Kadow, K. F.; Cianci, C. W.; Krystal, M.; Meanwell, N. A. Respiratory syncytial virus inhibitors. Part 2: Benzimidazol-2-one derivatives. *Bioorg. Med. Chem. Lett.* **2004**, *14* (5), 1133–1137.
- (50) Yu, K.-L.; Wang, X. A.; Civiello, R. L.; Trehan, A. K.; Pearce, B. C.; Yin, Z.; Combrink, K. D.; Gulgeze, H. B.; Zhang, Y.; Kadow, K. F.; Cianci, C. W.; Clarke, J.; Genovesi, E. V.; Medina, I.; Lamb, L.; Wyde, P. R.; Krystal, M.; Meanwell, N. A. Respiratory syncytial virus fusion inhibitors. Part 3: Water-soluble benzimidazol-2-one derivatives with antiviral activity in vivo. *Bioorg. Med. Chem. Lett.* **2006**, *16* (5), 1115–1122.
- (51) Yu, K.-L.; Sin, N.; Civiello, R. L.; Wang, X. A.; Combrink, K. D.; Gulgeze, H. B.; Venables, B. L.; Wright, J. J. K.; Dalterio, R. A.; Zadjura, L.; Marino, A.; Dando, S.; D'Arienzo, C.; Kadow, K. F.; Cianci, C. W.; Li, Z.; Clarke, J.; Genovesi, E. V.; Medina, I.; Lamb, L.; Colonna, R. J.; Yang, Z.; Krystal, M.; Meanwell, N. A. Respiratory syncytial virus fusion inhibitors. Part 4: Optimization for oral bioavailability. *Bioorg. Med. Chem. Lett.* **2007**, *17* (4), 895–901.
- (52) Wang, X. A.; Cianci, C. W.; Yu, K.-L.; Combrink, K. D.; Thuring, J. W.; Zhang, Y.; Civiello, R. L.; Kadow, K. F.; Roach, J.; Li, Z.; Langley, D. R.; Krystal, M.; Meanwell, N. A. Respiratory syncytial virus fusion inhibitors. Part 5: Optimization of benzimidazole substitution patterns towards derivatives with improved activity. *Bioorg. Med. Chem. Lett.* **2007**, *17* (16), 4592–4598.
- (53) Combrink, K. D.; Gulgeze, H. B.; Thuring, J. W.; Yu, K.-L.; Civiello, R. L.; Zhang, Y.; Pearce, B. C.; Yin, Z.; Langley, D. R.; Kadow, K. F.; Cianci, C. W.; Li, Z.; Clarke, J.; Genovesi, E. V.; Medina, I.; Lamb, L.; Yang, Z.; Zadjura, L.; Krystal, M.; Meanwell, N. A. Respiratory syncytial virus fusion inhibitors. Part 6: An examination of the effect of structural variation of the benzimidazol-2-one heterocycle moiety. *Bioorg. Med. Chem. Lett.* **2007**, *17* (17), 4784–4790.
- (54) For the RSV data [40 compounds], E-NOVO run time was 5 h at 20 conformations and 5 h 30 min at 30 conformations: 3 GHz, P-IV, 2Gb RAM.

CI900073K