Published in final edited form as:
J Comput Chem. 2005 December ; 26(16): 1752–1780. doi:10.1002/jcc.20292.

Integrated Modeling Program, Applied Chemical Theory (IMPACT)

JAY L. BANKS[1], **HEGE S. BEARD**[1], **YIXIANG CAO**[1], **ART E. CHO**[2], **WOLFGANG DAMM**[1], **RAMY FARID**[1], **ANTHONY K. FELTS**[3], **THOMAS A. HALGREN**[1], **DANIEL T. MAINZ**[1], **JON R. MAPLE**[1], **ROBERT MURPHY**[1], **DEAN M. PHILIPP**[1], **MATTHEW P. REPASKY**[1], **LINDA Y. ZHANG**[1], **BRUCE J. BERNE**[2], **RICHARD A. FRIESNER**[2], **EMILIO GALLICCHIO**[3], and **RONALD M. LEVY**[3]

[1] Schrödinger, Inc., New York, New York 10036

[2] Department of Chemistry, Columbia University, New York, New York 10027

[3] Department of Chemistry and Chemical Biology and BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, New Jersey 08854

Abstract

We provide an overview of the IMPACT molecular mechanics program with an emphasis on recent developments and a description of its current functionality. With respect to core molecular mechanics technologies we include a status report for the fixed charge and polarizable force fields that can be used with the program and illustrate how the force fields, when used together with new atom typing and parameter assignment modules, have greatly expanded the coverage of organic compounds and medicinally relevant ligands. As we discuss in this review, explicit solvent simulations have been used to guide our design of implicit solvent models based on the generalized Born framework and a novel nonpolar estimator that have recently been incorporated into the program. With IMPACT it is possible to use several different advanced conformational sampling algorithms based on combining features of molecular dynamics and Monte Carlo simulations. The program includes two specialized molecular mechanics modules: Glide, a high-throughput docking program, and QSite, a mixed quantum mechanics/molecular mechanics module. These modules employ the IMPACT infrastructure as a starting point for the construction of the protein model and assignment of molecular mechanics parameters, but have then been developed to meet specialized objectives with respect to sampling and the energy function.

Keywords

IMPACT; Monte Carlo simulation; QM/MM applications

Introduction

This article provides an overview of the IMPACT molecular mechanics program. The emphasis is on recent developments of the program with respect to both molecular mechanics core technologies and to specialized technologies developed in response to the needs of computational scientists working on current drug discovery projects in the pharmaceutical industry. We begin with a brief history of the development of IMPACT.

The first molecular dynamics simulation of a protein was reported in 1977.[1] Martin Karplus[2] has recently published a brief history of molecular dynamics simulations of biological

Correspondence to: R. M. Levy; e-mail: ronlevy@lutece.rutgers.edu.

macromolecules, and of the period in his laboratory, during which time one of us (R.M.L.) was a postdoctoral student in the group. It was difficult at the time to carry out molecular dynamics simulations of proteins using programs then available, and several members of the Karplus group at Harvard University discussed their ideas about how the situation might be improved. It is interesting to note that several of the articles in this special issue of the *Journal of Computational Chemistry* are coauthored by scientists who participated in the molecular mechanics and dynamics program development at Harvard during this period in the late 70s and early 80s, and that the genesis of the CHARMM, AMBER, GROMOS, and IMPACT simulation packages can be traced to this period. The development of IMPACT began in the Levy group at Rutgers University, in 1985, starting with a core set of molecular mechanics energy subroutines used to carry out molecular dynamics simulations of proteins. Early research using IMPACT focused on the relation between simulations and NMR experimental studies of protein structure and dynamics,[3,4] and on protein solvation.[5,6] In 1992, the Center for Theoretical Simulation of Biological Systems was established by Richard Friesner at Columbia University and three of the authors of this article (R.A.F., B.J.B., and R.M.L.) began to collaborate using IMPACT as a platform for methods development. In 1997, a strategic development partnership involving Columbia, Rutgers, Yale, and Schrödinger LLC was formed to provide a path for commercializing some of the new methods being developed for simulations of protein structural changes and interactions with ligands. As is apparent from the author list of this article and from the focus of those sections of our review concerned with specialized molecular mechanics technologies, the academic–commercial partnership has played an important role in the evolution of IMPACT. Basic research projects in our laboratories have benefited from this partnership, particularly by the development of tools to automate the preparation of ligand–protein complexes, by the expanded coverage of the force field, and by the increased coordination between various modeling packages.

Molecular simulations of protein structural changes and ligand binding are built upon two foundations: (1) the design of effective potentials that are matched to the requirements of accuracy and speed appropriate to particular modeling problems; and (2) the design of algorithms to sample the effective potentials in highly efficient ways so as to facilitate the convergence of the simulations in a thermodynamic sense and/or the coverage over large databases containing structures for which effective potential energy calculations are required. Developing algorithms to satisfy the competing goals of accuracy and speed is at the heart of the problem when considering computational models for use in structural biology, and strategies for achieving these twin goals in different molecular modeling contexts are emphasized throughout this review.

We have divided this article into sections that describe molecular mechanics core technologies, and ones that describe specialized technologies. In the former category we include a review of force field development and implicit solvation models, and also a description of the parallel and multicanonical molecular dynamics sampling algorithms available within IMPACT that take advantage of clusters of Linux processors that are now readily available. With IMPACT, it is possible to specify the use of a fixed charge force field OPLS_2003, which builds upon the OPLS_AA force field of Jorgensen and coworkers,[7,8] or the use of a polarizable force field that has been under development for several years and that we have previously described in a series of articles.[9–14] As discussed in this review, there has been a particular effort to extend the coverage of the molecular mechanics force field to include a very large number of different pharmaceutically relevant organic molecules and to facilitate atom typing and parameter assignment.

IMPACT includes two specialized molecular mechanics modules: Glide, a high-throughput docking program;[15–17] and QSite,[18,19] a mixed quantum mechanics/molecular mechanics module. These modules employ the IMPACT infrastructure as a starting point for the

construction of the protein model and assignment of molecular mechanics parameters, but have then been developed to meet specialized objectives with respect to sampling and the energy function. Glide has been designed to meet accuracy, speed, and coverage requirements for the identification of lead pharmaceutical compounds in high-throughput virtual screening tests. At the other end of the molecular mechanics spectrum, QSite has been designed to handle chemical reactions. The functionality of QSite has been achieved by tightly coupling IMPACT with the Jaguar (Schrödinger LLC, Portland, OR) suite of *ab initio* programs.

The following two sections describe the general organization of the IMPACT program and the Maestro graphical user interface, which can be used to control the process of setting up and running an IMPACT job. We then provide an overview of the molecular mechanics core and specialized technologies available with IMPACT. These sections include descriptions of some applications to research problems of current interest in our groups.

General Organization of the IMPACT Program

The IMPACT program is composed of a series of modules. The core module is responsible for loading system definitions from user files and setting up the structural and energetic data structures. Other modules provide energy functions and simulation protocols. Some of the modules are portable libraries that are utilized by several programs of the Schrödinger software suite (Macromodel, Jaguar, Prime, Maestro, etc.).

The user controls the behavior of the IMPACT program via a text input file (an example of which is given in Fig. 1), which contains instructions organized into tasks. The CREATE task contains instructions to define the chemical system, such as loading structure files and related system preparation tasks. The SET-MODEL task sets the parameters of the model, such as the type of solvation model, nonbonded cutoff definitions and QM/MM settings. The MINIMIZE task performs energy minimizations and the DYNAMICS task molecular dynamics simulations. Other tasks not shown in Figure 1 perform a variety of functions, some of which are described in this article. The syntax of the IMPACT input files also includes a pseudo programming language, called DICE (Dynamic IMPACT Control Environment), with which it is possible to define and manipulate scalar and multidimensional variables that interact with IMPACT's internal data structures. DICE also includes logical control constructs for branching, looping, and subroutine definition similar to more advanced structured programming languages. By using DICE it is possible to design complex computational tasks. DICE versatility is particularly useful in the analysis of trajectory files.

Maestro, described in the following section, is the graphical user interface for IMPACT developed at Schrödinger. The user prepares the system using Maestro. Maestro also collects simulation parameters from the user. This information is assembled, and an IMPACT input file is generated. The input file together with the relevant structure files are then used to drive the calculation using the IMPACT backend. At the end of the calculation information about the resulting molecular system is transmitted to the Maestro interface and the results of the calculation are displayed. This process occurs transparently from the user's point of view.

To make the interaction between Maestro and IMPACT possible a new protocol to load the molecular system into IMPACT has been developed. Prior to this innovation, the molecular system could only be defined using topology files for residues or individual molecules; a common protocol employed in molecular mechanics program packages. A database of topology files for standard protein and nucleic acid residues as well as frequently used solvent molecules existed; however, defining new chemical species involved compiling new topology files, a tedious and error-prone process. A new mechanism has been implemented whereby the molecular system definitions are loaded from one or more Maestro structure files, which

contain only connectivity information, bypassing the system construction procedure based on topology files. Energy parameters are assigned using an automatic scheme described in the Atom Typing and Parameter Assignment section. This allows for much greater flexibility in the kinds of molecular systems that can be handled by IMPACT; any chemical system that can be defined in Maestro can be simulated in IMPACT.

The new mechanism is invoked by assigning in the input file the “automatic” type to the molecular species being constructed; the previous mechanism is invoked by assigning the “protein,” “DNA,” etc., types depending on the kind of molecular system being studied. Simulation protocols that rely on internal coordinate definitions, such as internal coordinate Monte Carlo and Free Energy Perturbation (FEP) currently can only be performed using the system construction mechanism based on topology files, and consequently are not available through the Maestro graphical front end. The process of porting all the functionalities of IMPACT for access by Maestro is ongoing.

The Maestro Graphical User Interface

Maestro provides the graphical user interface (GUI) for IMPACT and several other computational chemistry programs. Maestro greatly simplifies the process of setting up, launching, and organizing IMPACT simulations. Maestro automatically generates an IMPACT input file based on the user input collected through Maestro’s graphical panels. For most applications the user does not need to interact directly with the IMPACT backend. Maestro also assists the user in preparing the molecular system. When a structure from the PDB database[20] or other type of structure is used as input for a specific job, structural inspection and modification is often necessary. Maestro’s build panel allows for mutating amino acid residues, changing bond orders, changing atom types, etc. Hydrogen atoms addition is performed automatically through a toolbar. The Maestro toolbar also allows for access to a number of common operations, such as deletion of water molecules, residues, and molecules.

IMPACT jobs, such as molecular dynamics, energy optimization, and hybrid Monte Carlo can be set up from Maestro panels. Within a panel, choices can be made with respect to force field, solvation treatment, use of periodic boundary conditions, type of implicit solvent, constraining/freezing parts of the system, etc. Selection of atoms that are to be kept fixed or frozen during a simulation is done using the Atom Specification Dialog (ASD). The ASD is highly versatile, allowing selection based on atoms (element, atom type, charge, etc.), residues (type, number, sequence, etc.), molecules, chains, and so on. Predefined groups of atoms may also be used. Atom selection can also be accomplished using commands for intersection and subtraction, as well as by defining spheres around a selected atom, residue, or molecule. The IMPACT job is usually started directly from Maestro, and is run in the background or on a remote computational host, while being constantly monitored by Maestro’s job control facility.

Maestro includes a project facility that helps organize the user’s work. Every structure read into Maestro becomes part of a project, and will be saved (unless the user chooses otherwise) in its last state when the project is saved or closed, or when the Maestro session is ended. If the starting structure belongs to a named project, upon job completion the resulting structures are incorporated into the project along with its associated job properties. The structures and properties of a project are stored in the project table, which behaves like a spreadsheet. Structures can be sorted, properties can be plotted, and data can be imported and exported to the spreadsheet table. In addition, visualization aids that have been applied to a structure or part of a structure will be saved with the project. The project table also features a tool (called e-player) used, for instance, to playback a molecular dynamics trajectory. Different operations can also be applied to each structure during playback, examples being rendering and coloring, displaying of hydrogen bonds and a variety of user-defined measurements.

The integration of IMPACT with Maestro, by streamlining system preparation procedures, job execution, and analysis, has both expanded the range of applicability of the program and shortened the time required to set up basic research projects using IMPACT. The ability to obtain visual clues at various stages of the project provides helpful insights into the behavior of the system. It is also helpful that IMPACT shares the same graphical user interface with a variety of other modeling programs.

Molecular Mechanics Core Technologies

Force Field Development

IMPACT contains facilities for modeling proteins, nucleic acids, and general organic molecules, utilizing both fixed-charge and polarizable molecular-mechanics force fields. We have made a major investment in force field development, with the goal of improving accuracy by fitting to high-level quantum-chemical data and by employing enhanced functional forms, with an emphasis on improving the treatment of nonbonded interactions. In what follows, we briefly describe our development methodology and summarize results, demonstrating accuracy via comparison with both gas-phase and condensed-phase data.

The development of improved fixed-charge and polarizable models has proceeded in parallel, sharing data sets and development tools when possible. In both cases, our philosophy has been to fit valence terms (particularly, torsions) to high-level quantum-chemical data for a training set of small molecules; this fitting is carried out after the nonbonded energy function has been specified. Although creating a database of valence parameters adequate to cover a significant fraction of medicinal chemistry space is a formidable technical challenge, it is straightforward conceptually. The greatest difficulties arise in generating an adequate set of torsional parameters; this is discussed below.

Our current fixed-charge force-field model, OPLS_2003, has its origins in the OPLS-AA force field of Jorgensen and coworkers.[7] We have attempted to retain the philosophy of that group, augmented by deploying substantially larger amounts of high-quality quantum-chemical data and by developing an automated atom-typing algorithm. We first discuss development of an improved version of the OPLS-AA force field for proteins, then present our efforts at expanding coverage of ligand functionalities for both fixed-charge and polarizable force fields.

In addition to providing a theoretically superior representation of the molecular charge distribution in an arbitrary environment, a polarizable force field has a second important advantage compared to a fixed-charge force field; it allows atom–atom pair interactions (van der Waals terms) to be optimized straightforwardly via fits to high-level quantum-chemical dimer interaction energies. This is not feasible with a fixed-charge force field because the gas-phase binding energies obtained from quantum-chemical calculations provide the wrong target; implicit inclusion of “average” polarization in the charge distributions, necessary to achieve reasonable results in the condensed phase, implies that it is not possible to simultaneously achieve quantitative agreement with gas-phase energetics. Scaled quantum-chemical binding energies and geometries can be used in this case,[21,22] but the scaling to be applied is always somewhat arbitrary.

A central objective has been to develop a nonbonded functional form, together with automated fitting protocols, that accurately reproduces a database of high-level binding energies. Results indicating our current precision and coverage are summarized below. We also reference results in which parameters developed via this type of fitting are used to carry out condensed-phase liquid-state simulations and to compare thermodynamic properties (heat of vaporization, density) with experiment. Finally, there are some subtleties in modeling polarization in the condensed phase; gas-phase polarizabilities appear not to be directly applicable, for reasons

discussed below. To address this difficulty, we use a heuristic approximation that performs reasonably well in initial testing. However, considerable further effort will be required to assess the overall accuracy obtainable using this, and other, approximations.

Fixed-Charge Protein Force Field

The set of 20 standard amino acids contains a relatively small number of chemical functional groups; furthermore, condensed-phase experimental data is available for many of these groups. The development of nonbonded parameters (charges, van der Waals parameters) for the OPLS-AA protein force field by Jorgensen and coworkers[7] exploited these observations. For many side-chain functionalities, as well as the backbone amide group, liquid-state simulations were performed and the charges and van der Waals radii were adjusted to reproduce experimental thermodynamic properties (heats of vaporization, densities) and were further evaluated by examining other properties such as solvation free energies in aqueous solution. We have retained the vast majority of the original OPLS-AA protein nonbonded parameters. The exception is the parameters for sulfur used in cysteine and methionine, where quantum-chemical calculations revealed an overbinding of small molecule dimers that was too large to be explained by polarization effects. A new set of charge parameters and van der Waals radii were developed that fit liquid-state simulation data equally well[8] and that yielded hydrogen-bonding energies much closer to the quantum-chemical results. These parameters also displayed substantially better performance in side-chain prediction tests (discussed further below).

For the valence force field, we retained the OPLS-AA stretches and bends and focused on refitting the torsional parameters to accurate quantum-chemical data.[8] In collaboration with the Jorgensen group, a set of rotamer states for model dipeptides were generated and relative energies of the various rotamer states for each amino acid were computed at the LMP2/cc-pVTZ (-f)//HF/6-31G** level, a level we have shown in previous work to be accurate to better than 0.5 kcal/mol.[23] New torsional parameters for the backbone and side chains were then fit to reproduce these relative energies. Results are summarized in Tables 1 and 2. The RMS error in side-chain conformational energies has been reduced by a factor of approximately two times compared to the original OPLS-AA parameterization and compared to MMFF94.[21,22,24–28] Backbone parameters were further tested by computing relative energies for a set of 10 conformations of the alanine tetrapeptide, for which quantum-chemical data has been computed at the same level as discussed above; again, a factor of approximately two times improvement in the RMSD is achieved by torsional refitting.

Condensed-phase performance of the protein force field has been investigated in the context of a continuum-solvation model, via conformational prediction of single side chains (keeping the rest of the protein constant) and loops. These tests have been carried out using the PLOP program, an offshoot of IMPACT that has now developed into a separate architecture focused on conformational-search (as opposed to molecular-dynamics) algorithms for modeling protein structure.[29,30] Our studies demonstrate that improved fitting of quantum-chemical rotamer energies leads to superior results in side chain prediction, when compared with structures in the PDB.[29] Similarly, the low RMSDs achieved for loop prediction for loops up to 12 residues in length validate both the backbone and side chain components of the potential functions.[30] These studies are complementary to molecular-dynamics simulations that examine the stability of the native protein structure over the trajectory of the simulation. Such simulations explore whether the native structure is a local minimum but not whether there are alternative local minima that may have lower free energies.

Force Field for Pharmaceutically Relevant Organic Molecules

Our objective in developing OPLS_2003 was to broaden the coverage of OPLS_2001, the previous standard implementation in IMPACT of the OPLS-AA force field,[7,31–35] and to improve the quality of the conformational energetics relative to MMFF94 and MMFF94s[21,22,24–28] as well as to earlier versions of OPLS-AA. OPLS_2001 includes parameters obtained from the parameter files for Boss4.0 and uses an automated atom-typing scheme that has been extended and improved for OPLS_2003. New force-field parameters were developed for OPLS_2003 for organic functional groups for which the OPLS_2001 force field does not provide specific parameters. Previously derived parameters for proteins,[8] discussed above, were implemented without modification in OPLS_2003.

Atom Typing and Parameter Assignment—All OPLS_2003 parameters for a given molecule are assigned automatically in an all-atom representation. The parameter assignment scheme employs atom types that are obtained by matching molecular fragments that describe the functional groups covered by the force field. These molecular fragments are stored as strings of characters using a notation similar to the SMILES/SMARTS language[36] used in cheminformatics. These strings of characters are called SMARTS patterns. For example, the SMARTS pattern that represents an alkene moiety is "C=C," and it is associated with the OPLS_2003 symbolic atom type "CM" for each of the carbon atoms; the SMARTS pattern for an amide group is "C(=O)N," and it is associated with the OPLS_2003 symbolic type "C" for the carbonyl carbon atom (the first atom in the pattern). For the OPLS_2003 force field the SMARTS pattern notation has been extended to increase the specificity of the pattern matching by introducing atom property labels. For example, the SMARTS pattern "[^sCM]-[^sCM]," where "^s" denotes the beginning of the symbolic atom type property label and "CM" is the atom type label, matches the two central carbon atoms of butadiene (C=C—C=C). Atom type properties can also include numerical values to, for example, match atomic partial charges. Atom property labels are used in pattern matches subsequent to atom typing in a manner similar to that employed in PATTY (Programmable ATom TYper).[37]

The SMARTS pattern-matching algorithm relies on the Lewis structure of the molecule. The atomic number and the formal charge of each atom as well as the bond orders of each covalent bond in the molecule define a Lewis structure. When the Lewis structure is not available or when an inconsistency of the Lewis structure is detected (by analyzing the formal charge and valence, computed as the sum of the bond orders, of every atom), a Lewis structure is derived from the atomic numbers and the interatomic connections. This is done in an iterative way. First, the formal charges are guessed by considering the number of connected atoms and their atomic numbers and then by finding an optimal set of bond orders for each connection. If this does not lead to a valid Lewis structure, the initial guess is varied and the process is repeated until a valid Lewis structure is obtained.

Each pattern is associated with a numerical atom type (used to assign bond charge increments, or BCIs for short),[21,24] a van der Waals type (used to assign Lennard–Jones parameters), and a symbolic type (used to assign stretching, bending, and torsional parameters). The valence parameters also depend on an index that, when set to a value other than the default of zero, allows special parameters to be assigned[24] without requiring that overly fine distinctions be built into the assignment of the symbolic atom types. For example, this approach allows the central C—C bond in butadiene to be recognized as being distinct from the terminal C=C bonds even though the four carbon atoms share the same symbolic type. This subtype index is assigned for bonds, angles, and torsions by matching the molecular connectivity and bond orders to a small list of patterns.

The partial atomic charges are assigned by first distributing any formal ionic charges over one or more atoms, using defined patterns, and by then adding contributions from the BCI

parameters associated with the chemical bonds. This approach ensures that the net charge on the molecule is maintained exactly and avoids the need for redistributing any "excess" molecular charge over the molecule, as is required in earlier implementations of the OPLS force field.

When no exact match of the numerical atom types can be found in the OPLS_2003 BCI database for a given bond, a charge increment based on the difference of the atomic electronegativities is assigned. When an exact match cannot be found for stretching, bending, or torsional interactions, parameters are assigned by similarity using a series of defined similarities between symbolic atom types. These similarities proceed from more specific to less specific symbolic types, which usually represent different atomic hybrids.

Parameterization—In the original development of the OPLS-AA force field, the partial charges and van der Waals parameters were adjusted to reproduce experimental heats of vaporization and densities for a series of pure liquids.[7,31–35] These parameters were further tested by comparison to experimental solvation energies, using explicit-solvent simulations. Additional comparisons were made in some cases to hydrogen-bond dimer interaction energies obtained from quantum-chemical calculations. These comparisons were used to detect large discrepancies that, when present, called for a reinvestigation of the nonbonded parameters. The OPLS-AA torsional parameters were fit to reproduce gas-phase conformational energies obtained from quantum-chemical calculations, and stretching and bending parameters were adapted from the CHARMM22 or AMBER force fields.

Our development of OPLS_2003 followed this general prescription. A central objective of this work was to significantly extend the range of chemical functionality covered by the force field. With this in mind, a training set of molecular structures was defined that consisted of the OPLS-AA training set, provided with the BOSS program, the training set used to develop MMFF94 and MMFF94s, a larger MMFF set that one of us (TAH) had prepared at Merck with the intention of extending the parameterization of MMFF, and additional compounds defined in this work. Authentic OPLS-AA nonbonded parameters and stretching and bending parameters were retained for the OPLS-AA core set of 112 compounds for which liquid-phase properties have been studied.[7,31–35] Initial estimates for the van der Waals parameters for new organic functional groups were assigned by analogy to OPLS-AA core parameters, and BCIs defined via additional combinations of numerical atom types were fit to reproduce molecular electrostatic potentials derived at the HF/6-31G*//B3LYP/6-31G* level of theory. Where necessary, these parameters were modified to improve the fit to scaled quantum-chemical geometries and interaction energies for the series of small-molecule dimers described below.

Stretching and bending interactions for which no parameters were available were adapted from MMFF94. Equilibrium values (ideal bond lengths and angles) were then adjusted to reduce the largest deviations observed in bond length and bond angles with respect to B3LYP/6-31G*-optimized geometries.

Much as was done for MMFF94s,[27] the out-of-plane bending of tri-coordinated nitrogen atoms that are conjugated to a pi-system (enamines, aromatic amines, etc.) was examined closely. These systems typically adopt a strongly pyramidal geometry at the nitrogen in the B3LYP/6-31G*-optimized gas-phase geometries, but most are regarded as being roughly planar in solution. To emulate condensed-phase behavior, the bending parameters of the nitrogen atom were adjusted to give nearly planar optimized OPLS_2003 geometries.

Finally, torsional parameters required for the expanded parameterization were fit to reproduce conformational energies obtained at the LMP2/cc-pVTZ(-f)//B3LYP/6-31G* level using a least-squares fitting protocol based on code originally used to develop MMFF94.[28]

Results—The datasets used to parameterize OPLS_2003 are considerably larger than those used for OPLS_2001, MMFF94, and other force fields with which we are familiar. For example, OPLS_2001 employs about 650 BCI parameters, while OPLS_2003 uses about 5700, of which 3200 are for 220 additional heterocyclic compounds. The quality of the charge distributions were assessed by comparing force-field and scaled quantum-chemical geometries and interaction energies for hydrogen-bond dimers, as has been done in previous studies.[21, 22] The MMFF dimer set, which numbered 65 for MMFF94 but subsequently was expanded to 195, was made available to us. We extended this set, most of which involve water, to a set of 550 structures by adding corresponding dimers involving methanol or *N*-methyl-acetamide, although we used mainly the water-dimer set in this work. Errors in dimer interaction energies larger than 1 kcal/mol were found for some of these complexes. For such cases, a conservative change to the non-bonded charge and van der Waals parameters was made, reducing the discrepancy between the force field and scaled quantum-chemical interaction energies to less than 1 kcal/mol in most cases.

A training set of 631 OPLS-AA, MMFF, and Schrödinger compounds was used for the torsional parameter fitting, and conformational energies and rotation profiles were obtained at the LMP2/cc-pvtz(-f) level using B3LYP/6-31G*-optimized geometries. We believe this to be the largest and highest quality set of conformational-energy data used in force-field development to date. The RMSD obtained for all conformational comparisons included in the fitting is 0.80 kcal/mol for OPLS_2003 compared to 2.97 kcal/mol for OPLS_2001 and 2.01 kcal/mol for MMFF94s. Further comparisons used equilibrium conformers for a set of 108 compounds taken from the published validation set for MMFF94. For this data set the RMSD is 0.48 kcal/mol for OPLS_2003, 2.43 kcal/mol for OPLS_2001, and 0.68 kcal/mol for MMFF94s.

To complete the development of OPLS_2003, surface-generalized Born (SGB) continuum-solvent parameters for the SGB/NP model were fit to reproduce experimentally derived free energies of hydration.[38] The dataset used in the previous work[38] was increased from 221 to 282 compounds. The RMSD for the difference in calculated and experimentally derived hydration free energies for the neutral and charged compounds used in the fitting is 0.38 kcal/mol for OPLS_2003 and 0.85 kcal/mol for OPLS_1999, representing a substantial improvement in this respect as well.

Polarizable Force Field for Pharmaceutically Relevant Organic Molecules

Over the past decade, a number of efforts have been made to explicitly incorporate polarization into molecular-mechanics force fields.[9–12,14,39–47] The present section provides a brief summary of our own work along these lines, discussing both the underlying theoretical approach and results that have been obtained to date. Key features of our approach include fitting to accurate quantum chemistry, coverage of a wide range of functionalities relevant to medicinal chemistry, and ability to carry out both explicit solvent and continuum modeling using a polarizable model.

We have described the basic philosophy, and algorithms, of our polarizable force-field development efforts in a series of previous articles.[9–14] The permanent electrostatic model consists of atom-centered point charges and dipoles, supplemented by lone-pair charges for oxygen atoms; the addition of quadrupoles produces marginally better agreement with quantum-chemical charge distributions, but not (in our view) at a level that exceeds other errors in the complete force field. These electrostatic parameters are fit to high-level quantum-chemical calculations; DFT methods produce reasonable charge distributions, but we have found that LMP2 calculations[23] yield somewhat better results, and we have employed such calculations in our most recent work.[13] At present, our methodology requires development of a new permanent electrostatic model for each new molecule from quantum-chemical

calculations. This is consistent with the idea that the polarizable force field will be deployed primarily for low-throughput applications (e.g., lead optimization) as opposed to, for example, screening of large libraries.

Our current model employs atom-centered dipole polarizabilities; a number of tests have established that the use of atom-centered fluctuating charges, without dipoles, is inadequate in some cases to represent the spatial character of the polarization. We have developed dipole polarizabilities for each of the polarizable atom types by fitting the polarization response for a series of small molecules with that from quantum-chemical calculations; details of our most recent results are reported in ref. [13]. Errors in three-body energies (determined by applying two point-charge or dipole probes to the test molecule) are typically less than 0.4 kcal/mol. Substantially, larger three-body energy errors can occur, but only for cases in which both probes are placed close to the same atom but on opposite sides of the atom. These latter errors can be shown to be due to an inherent limitation in the dipole polarizability model (i.e., atomic quadrupole polarizability terms would be needed to accurately model polarization energies in this particular case).

Various computational experiments have led to the conclusion that the polarizability in the condensed phase is smaller than that in the gas phase. If, for example, liquid-state simulations are carried out using the gas-phase polarizability, one observes substantial overbinding of the liquid state, due to overpolarization. We have hypothesized[11] that this is primarily due to the fact that diffuse functions contribute significantly to the polarization response in quantum-chemical calculations in the gas phase, and that these functions have substantially higher energy, due to overlap with the charge clouds of neighboring molecules, in the condensed phase. As a heuristic approach to this problem, we have investigated the effects of utilizing basis sets lacking diffuse functions in the computation of polarization parameters. This appears to work well in yielding agreement with liquid-state properties, as is discussed further below; a more extensive discussion of this issue is provided elsewhere[11] and similar discussions have also appeared.[48–50]

Once the electrostatic model is specified, the final task in completing the nonbonded polarizable energy model is determination of the nonelectrostatic atom–atom pair terms. This term represents two very different types of physical interaction. The long-range part is a true “van der Waals” interaction, modeling atom–atom dispersion interactions. In contrast, the short-range component incorporates contributions from exchange, Pauli repulsion, and other quantum-chemical forces that are very difficult to explicitly represent in a molecular-mechanics force field. The principal goal is to empirically adjust the heavy-atom distance in hydrogen-bonding (and other) short-range interactions, as well as the binding energy.

We first obtained long-range dispersion parameters, coefficients of the $1/r^6$ term in the pair function, by carrying out liquid-state simulations.[11] We have made an initial assumption that the dispersive term can be defined as dependent only on the atomic number of the atom, so that all oxygen atoms, for example, have the same dispersive coefficient. One could define a different co-efficient for each atom type, but the present more restrictive approximation appears to work well in the test cases we have investigated to date. A few small molecules (e.g., methane, ethane, methanol, and formamide for C, H, O, and N) were used to optimize the dispersive parameters.

The remainder of the pair function, which controls the short-range interactions (all of the remaining terms go rapidly to zero at long range), was optimized by fitting the structures and binding energies of small-molecule dimers. We have chosen to use a combination of the Lennard–Jones and exp-6 functional forms, the complete pair function being given by:

$$E_{nb}=\sum_{i<j} A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6}+C_{ij}\exp(r_{ij}/\alpha_{ij}). \quad (1)$$

Because the coefficient of the $1/r^6$ term is fixed for each atom, three parameters remain to be determined (as opposed to the single parameter that would be available if a Lennard–Jones 6–12 potential were employed). The additional functional flexibility enabled us simultaneously to fit both the hydrogen-bond distances and binding energies of the small-molecule dimer training set, and this in turn has been found to yield good results for both the heat of vaporization and density in liquid-state simulations; these simulations exhibited average errors in heats of vaporization of ~0.5 kcal/mol and less than 3% in the density.[11] The dimers were optimized at the LMP2/cc-pvtz(-f) level, followed by single-point extrapolated LMP2 calculations that are estimated to agree with the basis-set limit of MP2 to better than ~0.3 kcal/mol.[51] A wide range of organic functional groups was covered by fitting to a training set of 142 dimers (primarily various small molecules with water); the average deviation of the force field from the quantum-chemical results was 0.6 kcal/mol.[13] A test suite of 40 additional dimers yielded a similar level of error, suggesting that the parameters and combining rules derived in this effort are transferable and that overfitting has been avoided.

Once the nonbonded terms were defined, the valence component of the force field was fit in the same fashion as for the fixed-charge force field, using the same data sets. Tables 1 and 2 show that the residual errors for the protein force fields are similar for the fixed-charge and polarizable models; this suggests that these errors may now be dominated by stretching and bending terms, which are identical in both cases. Similarly, the average errors for the small-molecule training set used to obtain torsion parameters for the generalized polarizable force field are of the same order (1.0 kcal/mol) as for the fixed-charge force field.

IMPACT contains facilities for carrying out both explicit solvent and continuum-solvent modeling with the polarizable force field. The explicit-solvent simulation methodology has been derived from the SIM program of Stern et al.[52] which has been integrated with IMPACT's building, atomtyping, and other facilities. SIM contains technology for imposing Ewald boundary conditions using polarizable dipoles and/or fluctuating charges, as well as extended Lagrangian methods that render polarizable molecular-mechanics calculations considerably more efficient than approaches that rely on adiabatic optimization of the polarization at every time step; indeed, the computational effort for polarizable simulations in explicit solvent are only a factor of approximately two times larger than they are for fixed-charge simulations.[53] Continuum-solvation methods, including analytical gradients, are implemented via a self-consistent reaction field (SCRF) methodology using the Poisson–Boltzmann solver PBF (discussed further in the section on continuum-solvation models).[54] The SCRF formalism is isomorphic to that used in quantum-chemical continuum-solvation models. However, treatment of large systems such as proteins (as opposed to the small molecules typically modeled via quantum chemistry) is technically demanding, particularly for the gradient calculations. Successful minimization of 18 protein–ligand complexes using the PFF/PBF energy model have been performed.[13]

In principle, the polarizable force field described in this section should represent a significant improvement in accuracy and reliability compared to a fixed-charge force field. The performance in practice for problems of interest, such as the prediction of protein structure or protein–ligand binding, however, can only be determined by comparison with experimental structural and thermodynamic data for the relevant systems. Tests of this type are at present ongoing, but the data are as of yet insufficient to draw any firm conclusions.

Fast Multipole Method

There are two methods available in IMPACT to treat long-range electrostatic interactions that avoid the use of nonbonded cutoffs. In addition to the Ewald sum method, the Fast Multipole Method (FMM)[55,56] has been implemented in IMPACT. FMM allows simulations effectively without nonbonded cutoffs, which are known in some cases to introduce artifacts, without incurring the unfavorable O(N^2) scaling of the brute force calculation of the Coulomb interaction energy between every pair of atoms. The FMM algorithm can be used for periodic systems (with cubic periodic boundary conditions) as well as aperiodic systems (such as simulations in water droplets) for which standard Ewald and Particle Mesh Ewald (PME) methods cannot be applied. The FMM method starts with the observation that the electrostatic interaction energy between an atom and a set of distant atoms can be approximated by a multipolar expansion, up to a certain order, of the electric field of the distant atoms at the location of the atom. The multipolar expansion is expressed in terms of multipolar coefficients that depend only on the charge distribution of the distant atoms and not on the observation point. Therefore, once the multipoles of the distant atoms are computed, the cost of computing the long-range interaction energy of N atoms is proportional to N. Because in IMPACT's FMM implementation[57] the cost of calculating the multipolar expansions also grows linearly with system size, the overall cost of computing the long-range electrostatic energy of the system grows only linearly with system size. The O(N) scaling of FMM is superior to the O(NLog$N$) scaling of PME, the O($N^{3/2}$) scaling of the Ewald method and the O(N^2) scaling of the brute force algorithm. Therefore, for a large enough system, FMM is expected to be the most efficient algorithm to compute the long-range electrostatic energy. By indirect comparisons it was estimated that the FMM algorithm implemented in IMPACT becomes faster than PME for systems of about 20,000 atoms.[57] In the FMM implementation in IMPACT[57] the system is divided recursively into a tree of cubic clusters. Multipolar expansions are evaluated for the smallest clusters, which are then propagated to the parent clusters using efficient transformation relations. The implementation allows for arbitrary cluster tree depths and order of multipolar expansions. The FMM algorithm is coupled to a sophisticated MD RESPA scheme[57] that allows infrequent reevaluation of computationally expensive quantities.

Solvation Models

The accurate modeling of the effects of solvation on protein structural and thermodynamic properties has been a central concern in the evolution of modern molecular mechanics models for protein simulations. In the 1980s and the early 1990s much of the emphasis was on explicit molecular representations of the solvent. In principle, this is the most realistic approach to study solvation effects. Furthermore, the development of potential functions and advanced molecular dynamics and Monte Carlo algorithms for simulating biomolecules in solution and molecular liquids, have advanced in parallel and with considerable synergy. As described in the following section IMPACT has the ability to carry out explicit solvent simulations of proteins in solution using a variety of water models and methods for handling the periodic boundary conditions that arise in these kinds of simulations. However, computer simulations using explicit solvent models are computationally intensive, not just because of the very much larger number of atomic interactions that need to be calculated at every step, but also, and perhaps more importantly, because of the need in most problems of interest to average the fluctuating effects of the solvent reaction field on the biomolecular solute. For modeling the binding of ligands to proteins and other problems where the emphasis is on structure and thermodynamics, effective potential models that treat the solvent implicitly have much to offer. These models can be derived, at least conceptually, from the consideration of the solvent potential of mean force.[58] We have benefited from this perspective in our own work on the development of implicit solvent models. By carrying out "computer experiments" on solvation thermodynamics in explicit solvent and extracting potentials of mean force we have obtained

useful insights into the construction of functional forms for implicit solvent models and their parameterization.[38,59–65] The goal is to derive functional forms for the solvent-averaged potential of mean forces that have the "flavor" of a molecular mechanics potential functions; that is, that the solvation potential and its derivatives can be calculated as analytical functions of the atomic coordinates of the biomolecular solute. Much of our work in recent years on solvation models for IMPACT has been focused on this goal.

The starting point for our discussion of implicit solvation models is the expression for the total "effective energy" of the biomolecule in solution.

$$E_{\text{tot}}=E_0+\Delta G_{\text{solv}} \tag{2}$$

where E_0 represents the molecule's energy in the gas phase (E_{vac}) if the molecular mechanics model is a polarizable model, or, if more commonly E_0 is calculated using a fixed-charge molecular mechanics function, it represents the intramolecular effective energy of the solute including the electronic polarization effects of the solvent in an average way. ΔG_{solv} is the solvation free energy change for transferring the molecule from the gas phase to solution for the polarizable model; but for fixed charge molecular mechanics models ΔG_{solv} does not include the work required to polarize the solute's charges that are built into E_0.66 (We note that it is ΔE_{tot}, the change in the total effective energy of the solvated system with a change in the solute's coordinates, that is usually the quantity of interest. The solvent contribution, $\Delta\Delta G_{\text{solv}}$, is the solvent averaged potential of mean force for the conformational change of the solute in solution.)

To estimate the total solvation free energy ΔG_{solv} of the molecule, a standard charge decoupling procedure is followed by which ΔG_{solv} is decomposed into electrostatic and nonpolar contributions:

$$\Delta G_{\text{solv}}=\Delta G_{\text{elec}}+\Delta G_{\text{nonpolar}} \tag{3}$$

where ΔG_{elec} is the free energy change for removing all the charges from the molecular shaped cavity in the vacuum and adding them back to the corresponding cavity in solution. The nonpolar free energy change can be decomposed into contributions from cavity formation and the favorable van der Waals attraction between the solute and solvent molecules. With IMPACT a continuum dielectric estimate of ΔG_{elec} can be evaluated as described below using a Poisson–Boltzmann solver (PBF, see below). However, it is not in general practical to perform molecular dynamics simulations, or high-throughput ligand docking simulations using a Poisson–Boltzmann framework to evaluate ΔG_{elec}. Instead. the modeling community has developed a strong interest in a class of approximations known as Generalized Born models, [67,68] which can be derived from the Poisson equation. Two implementations of GB have been coded in IMPACT: the Analytical Generalized Born (AGB) model, and the Surface Generalized Born (SGB) model. These models are described below with an emphasis on the AGB model. Unlike most molecular mechanics packages for which $\Delta G_{\text{nonpolar}}$ is approximated by a surface area term, after extensive experimentation based on explicit solvent simulations, we have chosen to model the favorable van der Waals attraction energy between the solute and solvent using a different functional form than the estimator used to model the work of cavity formation; for this term we retain a standard surface area representation. The rationale for our choice of the nonpolar function and our initial parameterization of the model is explained in a following section.

Explicit Solvent

Explicit solvent simulations[69] using the SPC,[70] TIP3P, and TIP4P[71] water models are supported natively. Other solvents can be constructed from the topology file of a representative solvent molecule. A solute–solvent system using explicit water can be set up either from Maestro or by directly interacting with the IMPACT backend. The construction of solvated systems in solvents other than water necessitates the use of topology files and requires direct interaction with the backend. In Maestro, the procedure for constructing an explicit solvent system is called Soak. The starting point is a solute conformation. The user selects the water model and the dimensions of the solvent box. The IMPACT backend then inserts the solute in the solvent box using a prethermalized sample box, which is replicated and clipped to cover the final box size. Water molecules that overlap with the solute are removed. The resulting solvated system is usually energy minimized and thermalized using constant pressure molecular dynamics. The Ewald or the Fast Multipole Method (FMM) can be used to model long-range electrostatics with periodic boundary conditions. IMPACT uses a group-based (residue-based or molecule-based) Verlet neighbor list for short-range nonbonded interactions. The non-bonded interaction energy routine is multithreaded allowing for increased speed on symmetric multiprocessor computers.

Poisson–Boltzmann Solver

PBF is a program that uses finite element methods to numerically solve the linearized Poisson–Boltzmann (PB) equation:

$$\nabla \cdot (\varepsilon \nabla \varphi) = -\frac{4\pi\rho}{kT} + \varepsilon\kappa^2\varphi \tag{4}$$

where ε is the dielectric constant, φ is the electrostatic potential, ρ is the solute charge density, $\kappa^2 = (8\pi\varepsilon^2 I)/(\varepsilon kT)$ is the inverse Debye length squared, and I is the ionic strength. Unlike finite difference methods, the finite element approach enables the use of a nonuniform grid; in the present case, the grid points are concentrated at the dielectric boundary. However, to compute the multidimensional derivatives required to represent eq. (4), the grid must be assembled into connected volume elements. PBF employs tetrahedra as elements, each defined by the four grid points of the vertices. Tiling the grid with tetrahedra is a complex computational task; the approach taken in PBF is discussed in detail in ref. [72]. It is also essential that the dielectric surface can be constructed by putting together the appropriate tetrahedral faces; again, the methodology to accomplish this in PBF is discussed in ref. [72].

Once the tetrahedral tiling of the grid has been completed, gradients are defined using standard finite element formulas, and the results implemented as a matrix operation acting on the values of the electrostatic potential at the grid points. The matrix so generated is, for a given system and effective resolution, smaller than that used in a finite difference approach, as the finest resolution in the finite element grid is not required everywhere in the system. On the other hand, the matrix operator has more off diagonal elements in a finite element representation than it does in a finite difference representation, and the structure of the matrix is more irregular. These two factors probably roughly cancel each other in practice, suggesting that neither method has a fundamental performance advantage for iterative solution of the PB equation. At present, a significant amount of time is required to build the tetrahedral finite element mesh; efficiency could be greatly improved for geometry optimization, compared to what is shown below, by rebuilding the mesh when and where necessary, as opposed to a complete rebuild for every geometry step (which is what is currently implemented). Our focus at present is on obtaining increased accuracy, and the current level of performance is sufficient to carry out the relevant computational experiments.

The dielectric surface in PBF is defined using a sum of atom-based Gaussians; the parameters have been adjusted to qualitatively yield results similar to the classical solvent accessible surface (Connolly surface) that is used in DelPhi[73] and other finite difference PB solvers. However, the Gaussian surface is smooth and permits a straightforward implementation of analytical gradients, using methods discussed in ref. [54]. Both the Gaussian surface and the Connolly surface can be designed so as to properly define areas where water molecules cannot fit as low dielectric; however, there are still nontrivial differences in the details of the surface. Because neither model can be rigorously derived from an exact statistical mechanical treatment (all continuum solvation models in fact must be parameterized to provide a reasonable representation of first shell interactions with water in any case), it is unclear which surface will deliver better performance in practical calculations. Accuracy in various tasks (e.g., computing relative energies of different side chain or loop conformations) can only be ascertained by taking a particular total implementation (surface definition, numerical grid, partial atomic charges, etc.) and comparing with the relevant experimental data. Comparisons of PBF with DelPhi[73] in various benchmarks show good qualitative agreement for relative energies, [74] while comparisons with free energy perturbation calculations for solvation free energies of peptides and proteins have yielded remarkably good agreement.[62] Nevertheless, the quantitative accuracy of any PB methodology for structural or binding affinity predictions in complex systems has yet to be assessed in depth.

PBF has been designed to work with both fixed charge and polarizable force fields, as well as with a quantum chemical description of the solute. In both force fields cases, we employ the model discussed in the SGB/NP section to describe the nonpolar component of the free energy. For the fixed charge force field, we use the OPLS-AA charge model described in the Fixed-Charge Protein Force Field section of this article. For the polarizable force field and quantum chemical descriptions of the solute, a self-consistent reaction field (SCRF) formalism is employed, as is discussed in detail in refs. [75] and [76]. In this approach, an initial guess is made for the solute charge distribution, the solvent reaction field is evaluated using PBF, the reaction field (in the form of surface point charges) is determined, and these are then used to recalculate the solute charge distribution. This process is iterated until the solute charge distribution has converged. In this fashion, the polarization of the solute due to the electrostatic field of the solvent, and the solvent alignment due to the solute, are self-consistently adjusted. The technology to treat the polarizable force field and quantum chemical descriptions of the solute are isomorphic; algorithms for calculating the analytical gradient are complicated but straightforward, and are described in ref. [54].

For all three models, dielectric and nonpolar parameters must be optimized for each atom type if first shell interactions are to be described with a reasonable degree of accuracy. The principal means at present for such optimization is fitting the parameters to experimental solvation free energies of small molecules. Table 3 summarizes the results for the three models, presenting the number of molecules in the training set, and mean unsigned errors between the calculated and experimental values. Details of the fitting protocol, training set, and distribution of errors can be found in refs. [13] and [77].

To demonstrate the capability of our PBF implementation, we have carried out solution-phase optimization of a number of systems, including the cocrystallized ligands and native proteins of various sizes. Initial structures were taken from the PDB data bank. In carrying out such energy minimizations, we begin by minimizing the systems using a fixed charge force field and the SGB/NP implicit solvent model, and then we switch to PBF sequentially increasing model complexity from the fixed-charge model to the QM model. Using this staging scheme, whereby more complex models are used at later stages of the calculation, saves significant amount of CPU time. In this study, we have implemented some preliminary protocols to

increase the speed of the minimization procedure, such as calling the PBF solver infrequently. Work is in progress to further optimize both the PB solver and the minimization scheme.

The SCRF version of PBF has been designed to work with QSite, so that QM/MM as well as QM solutes can be investigated. Although numerous applications of PBF with Jaguar (pure QM model) have been carried out to date, the PBF/QSite capability is relatively new and has not been extensively investigated, or optimized, as of yet.

In summary, the PBF module provides the capability within IMPACT to numerically solve the full PB equation, for both single-point energy and gradient calculations, as opposed to the approximations inherent in the GB-based models discussed below. Such solutions are computationally more expensive, but may provide improved accuracy (although this is dependent upon the details of the parameterization, and remains to be demonstrated). The SCRF capabilities represent the principal approach to developing continuum solvation models when the solute is to be treated as polarizable (whether via a polarizable force field or via quantum mechanics). Some degrees of parameter optimization and testing have been performed for all of the models, but this is an area that will require a great deal of work in the future to ensure accuracy and robustness for a wide range of systems.

Generalized Born Models

Generalized Born (GB) models[67,78] estimate the electrostatic component ΔG_{elec} of the hydration free energy as

$$\Delta G_{\text{elec}} \approx \Delta G_{\text{GB}} = \sum_i \Delta G_{\text{self}}(i) - \left(\frac{1}{\varepsilon_{\text{in}}} - \frac{1}{\varepsilon_{\text{out}}}\right) \sum_{i<j} \frac{q_i q_j}{f_{ij}}, \tag{5}$$

$$f_{ij} = \sqrt{r_{ij}^2 + B_i B_j \exp(-r_{ij}^2/4B_i B_j)}, \tag{6}$$

where the summation runs over the atoms i and the atom pairs (ij, $i \leq j$) of the solute, $\varepsilon_{\text{in}} = 1$ and $\varepsilon_{\text{out}} = 80$ are the interior and exterior dielectric constants, q_i is the partial charge of atom i, and r_{ij} is the distance between atoms i and j. $\Delta G_{\text{self}}(i)$ is the self-energy of atom i, defined as the electrostatic solvation free energy of the solute when only the partial charge of atom i is turned on. The summation over atom pairs in eq. (5) is the sum of the GB pair energies. The Born radius B_i of atom i is defined as the radius that reproduces the self-energy, $\Delta G_{\text{self}}(i)$, of atom i according to the Born formula

$$\Delta G_{\text{self}}(i) = -\frac{1}{2}\left(\frac{1}{\varepsilon_{\text{in}}} - \frac{1}{\varepsilon_{\text{out}}}\right)\frac{q_i^2}{B_i}. \tag{7}$$

The self-energy is largest in absolute value for the atoms that are most exposed to the solvent because they are capable of inducing stronger polarization fields. This effect is captured by the GB model in that atoms exposed to the solvent have smaller Born radii, whereas buried atoms tend to have larger Born radii. The pair-energy term corresponds to the dampening of electrostatic interactions in a high dielectric medium due to the screening of the solute charges. The GB equation can be shown to be an exact representation of the electrostatic charging free

energy of the solute in a continuum dielectric in the two limiting cases of infinite atomic separation and complete atomic overlap.[67]

In the Coulomb Field approximation the Born radius is given by the integral over the solvent volume of the fourth power of the distance between the solute atom and the solvent[79]

$$\frac{1}{B_i}=\frac{1}{4\pi}\int_{\text{solvent}}\frac{1}{(\mathbf{r}-\mathbf{r}_i)^4}d^3\mathbf{r}. \tag{8}$$

The accuracy of the Coulomb field approximation has been analyzed using exact analytical models[78] and accurate numerical PB calculations.[80,81] It has been found to be generally acceptable with the exception of cases with very asymmetric solute geometries, where it tends to overestimate the values of the Born radii. Empirical corrections to Coulomb field approximation have been proposed.[79,81] It has been pointed out that approximations in the integration procedure to obtain the Born radii using eq. (8) may actually be of more significance than the Coulomb field approximation itself.[80]

AGBNP

AGBNP (Analytic Generalized Born plus Nonpolar) is a recently developed implicit solvent model.[61] Development versions of AGBNP were first incorporated in the 2.6 version of academic IMPACT. The AGBNP model will be available in the 2005 release from Schrödinger. The numerical implementation of AGBNP[61] includes several speed optimization features, and it makes use of multithreading techniques for increased speed on symmetric multiprocessor computers.

Several key objectives were pursued in the development of the AGBNP implicit solvent model: the applicability of the model over a wide range of molecular sizes, from small molecules to large proteins, and over the wide range of functional groups present in ligand databases; the ability to correctly model hydration free energy differences, including large-scale protein motions and the motion of only a few atoms; the ability to express the model in analytical form with analytical gradients, and finally, computational efficiency. These requirements were dictated by the range of applications, hydration free energy prediction, ligand affinity prediction, induced fit, loop modeling, protein folding, protein binding, and protein allostery that we wished to pursue with this model. Although several models exist with some of the above characteristics, [38,68,77–79,82–86] none of them, in our view, has the flexibility and computational efficiency we required.

AGBNP is based on the decomposition of the solvent potential of mean force into an electrostatic component and a nonpolar component. The electrostatic component is modeled using the Generalized Born model. AGBNP introduces two key innovations: a parameter-free analytical pairwise descreening scheme for the calculation of Born radii, and a nonpolar hydration free energy model that includes a solute–solvent van der Waals interaction energy estimator.

Parameter-Free GB Pairwise Descreening Scheme

The key quantities in Generalized Born models are the Born radii B_i of each atom. Generalized Born implementations in use in molecular simulation programs[87] differ mainly in the procedure used to calculate Born radii. In the Coulomb field approximation the Born radius of atom *i* is given by eq. (8). A more computationally convenient expression is obtained by adding and subtracting from eq. (8) the expression for the inverse of the Born radius of a solute composed only of atom *i*, yielding

$$\frac{1}{B_i}=\frac{1}{R_i}-\frac{1}{4\pi}\int_{\Omega_i} d^3\mathbf{r}\frac{1}{(\mathbf{r}-\mathbf{r}_i)^4} \tag{9}$$

where R_i is the van der Waals radius of atom i, $\mathbf{r}_i$ is the position of atom i, and Ω_i represents the solute volume outside atom i. The first term on the right-hand side of eq. (9) represents the Born radius of atom i in the absence of all other solute atoms; the integral expression takes into account the displacement of the solvent dielectric due to the other solute atoms. The goal is to evaluate this integral as accurately and efficiently as possible. AGBNP computes this term using an analytical expression inspired by dielectric descreening schemes.84 In the dielectric descreening method the integral on the right-hand side of eq. (9) is replaced by a pairwise sum over the solute atoms,

$$\frac{1}{B_i}=\frac{1}{R_i}-\frac{1}{4\pi}\sum_{j\neq i} Q_{ji} \tag{10}$$

where Q_{ji} is a quantity that corresponds to the contribution of the volume associated with atom j to the integral in eq. (9), given by

$$Q_{ji}=\int_{\Omega_{ji}} d^3\mathbf{r}\frac{1}{(\mathbf{r}-\mathbf{r}_i)^4} \tag{11}$$

where the integration domain is the region Ω_{ji}, corresponding to the volume of atom j that lies outside atom i. The integral in eq. (11) can be expressed in analytical form. However, due to the overcounting of regions that lie inside more than one atomic sphere, eq. (10) significantly overestimates the values of the Born radii. To correct for this the contribution from each atom j in eq. (10) is reduced by a scaling factor, s_{ji}, less than 1

$$\frac{1}{B_i}=\frac{1}{R_i}-\frac{1}{4\pi}\sum_{j\neq i} s_{ji}Q_{ji}. \tag{12}$$

The Generalized Born implementation in AGBNP is unique among pairwise descreening schemes in that the values of the factors to account for atomic overlaps are not predetermined by parameterization with respect to Poisson–Boltzmann results or experimental data, [68,84,86,88,89] but are instead computed on the fly from the geometry of each solute conformation. A detailed description of the algorithm to compute the scaling factors s_{ji} can be found in ref. [61]. Briefly, the scaling factor s_{ji} is defined as the fraction of volume of atom j assigned exclusively to that atom (called the self-volume)— excluding overlaps with atom i, which are already taken into account in the definition of Q_{ji}. The atomic self-volumes are defined by partitioning the solute volume into volumes occupied by one, two, three, etc., atoms. These volumes are assigned in equal fractions to the component atoms. For example, of the volume occupied both by atoms i and j, half is assigned to the self-volume of atom i and half is assigned to the self-volume of atom j. Due to their complex geometries, the volumes of these regions are not amenable to easy evaluation. Instead, each volume is decomposed into intersection volumes between pairs, triples, etc., of atoms and the following expression for the self-volume, V'_i, of atom i is obtained:

$$V'_i = V_i - \frac{1}{2}\sum_j V_{ij} + \frac{1}{3}\sum_{j<k} V_{ijk} - \cdots \tag{13}$$

where V_i is the van der Waals volume of atom i, V_{ij} is the intersection volume of atoms i and j, V_{ijk} is the intersection volume of atoms i, j, k, and so on. By summing the self-volumes of the atoms given by eq. (13), the volume of the molecule is obtained (a result known as the Poincarè formula for the volume of an object composed by the union of overlapping spheres[90]), confirming the interpretation of V'_i as the self-volume of atom i, that is, the volume of atom i that can be regarded to belong only to that atom. Two simplifications are then applied. The first, justified by the fact that the analytical expression for the intersection volume of multiple spheres is extremely complex, [91] consists of estimating the volume of intersection of multiple atoms as the overlap integral of Gaussian functions centered on each atom with parameters adjusted to reproduce the extent and the volume of each atom.[61,90] The second simplification is approximating the self-volume V'_{ji} of atom j in the absence of atom i, needed as mentioned above to compute the overlap scaling factor s_{ji}, as the self-volume of atom j plus one-half the intersection volume of atoms i and j. This approximation, valid when intersection volumes of higher order involving atoms i and j can be neglected, simplifies considerably the algorithm that would otherwise require the calculation of multiple self-volumes for each atom. We therefore finally set

$$s_{ji} = \frac{V'_j + V_{ij}/2}{V_j}. \tag{14}$$

The parameter-free approach employed by AGBNP to calculate pairwise descreening scaling coefficients is particularly useful when treating unusual functional groups often found when screening large numbers of ligand candidates from a database. Scaling coefficients derived from a training set[68,88,89,92] in which a particular functional group is not represented may be unsuitable for such a functional group. On the other hand, it is impractical to construct a training set in which all possible functional groups and combinations of functional groups are represented. The parameter-free analytical scheme used in the AGBNP model ensures that each atom in any molecule is assigned proper scaling coefficients. In the context of molecular modeling projects concerned with ligand binding, the parameter-free feature of AGBNP makes it possible to use a very diverse database of ligands.

Nonpolar Model—AGBNP includes a novel analytical nonpolar hydration free energy estimator. It is distinct from commonly used surface area-based nonpolar estimators in that it is based on the nonpolar hydration free energy decomposition into a cavity term, proportional to the solute surface area, and an attractive dispersion energy term, which approximates the solute–solvent van der Waals interaction energy assuming uniform solvent density outside the solute. We have shown that using independent models for the cavity and the dispersion energy components reproduces the nonelectrostatic solvation properties of small molecules and macromolecules qualitatively better than models that describe both of these effects through a model exclusively based on the solute surface area. In particular, it was observed[64] that, due to the medium-range nature of van der Waals interactions, the protein–ligand binding energy penalty incurred by the loss of ligand–water van der Waals interactions (often larger in magnitude than the corresponding binding free energy) is poorly correlated with the buried surface area upon binding, and that a surface area model parameterized on small molecules, peptides, and proteins consistently overestimated the loss of ligand–water van der Waals

interactions. Similar effects were shown to be responsible for the inability of surface area models to describe even in a qualitative sense the potential of mean force for dimerization of two alanine dipeptide molecules in water, which was instead reproduced by a model, described below, which accounts for solute–solvent attractive van der Waals interactions through an estimator based on the Born radii of the solute atoms.[65]

We obtained the first insights into this problem studying the hydration free energies of alkanes in explicit solvent. The well-known fact that cyclic alkanes are more soluble in water than linear and branched alkanes with similar surface area, although reproduced by explicit water simulations, lacked theoretical justification. It was found[60] that for these small alkanes the solute–solvent van der Waals energy per solute atom was approximately independent of the solute surface area. Therefore, the increased relative solubility of cyclic alkanes is due to the fact that cyclic alkanes interact as strongly with the solvent as linear and branched alkanes of similar chain length, suffering at the same time a smaller cavity formation free energy penalty due to their smaller surface areas. This observation helped us to also understand the very different values of the effective surface tension obtained when fitting a surface area model to either the hydration free energies of a series of alkanes or a series of different conformations of the same alkane molecule.[60,93] It was shown that conformational changes caused variations in the cavity component of the nonpolar hydration free energy, whereas solute–solvent dispersion interaction energies of these small molecules were much less affected by conformational changes. These observations suggested that an accurate nonpolar hydration free energy estimator applicable to a wide range of chemical functionalities and conformational variations should be composed of two independent components: one corresponding to the cavity hydration free energy, and the other corresponding to the solute–solvent van der Waals interaction energy. Free energy perturbation calculations in explicit solvent showed that for small molecules the cavity component was correlated with short-range first solvation shell estimators such as the solute surface area, whereas the solute–solvent dispersion energy depended also on second shell and longer range interactions less dependent on solute conformation. These ideas were applied with success to the prediction of experimental hydration free energies of a large set of small molecules[38] using SGB for the electrostatic component, a surface area model for the cavity component and an atom-type–dependent term independent of solute conformation. Later we showed some of the shortcomings of using a surface area model to describe solute–solvent van der Waals interaction energies of peptides, proteins, and protein–protein and protein–ligand complexes, obtained from explicit solvent simulations.[64] Based on this data we then developed a solute–solvent dispersion energy analytical expression based on the Born radius of each solute atom[61] that was simple enough to be used for molecular dynamics sampling. The nonpolar hydration free energy estimator finally used in the AGBNP model has the following expression:

$$\Delta G_{\mathrm{np}}=\sum_{i}(\gamma_i A_i+\alpha_i W_i) \tag{15}$$

where the summation runs over solute atoms, γ_i is the surface tension parameter assigned to atom i, α_i is the dimensionless solute–solvent van der Waals interaction energy parameter assigned to atom i, and A_i and W_i are geometrical estimators that depend on solute conformation. A_i is the surface area of atom i and W_i is defined as

$$W_i=\frac{a_i}{(B_i+R_w)^3} \tag{16}$$

where a_i depends on the Lennard–Jones parameters of atom i,61 B_i is the Born radius of atom i, and R_w = 1.4 Å is the radius of a water molecule. The nonpolar model is currently minimally parameterized;61 a single value of the surface tension parameter obtained by fitting to explicit solvent free energy perturbation calculation results of the cavity hydration free energies of alkanes is used for all atom types. The van der Waals parameter α_i have been minimally adjusted from their ideal value of 1 to better reproduce solute–solvent van der Waals interactions from explicit solvent simulations;61,64 their values range from 1 to 0.75.

SGB/NP

The Surface Generalized Born (SGB) [79] plus Nonpolar (NP) [38] implicit solvent model (SGB/NP) is based on the Generalized Born continuum dielectric model and a nonpolar hydration free energy estimator similar to the one employed by the AGBNP implicit solvent model described in the previous section. The SGB/NP Generalized Born implementation differs from other GB implementations in two main respects: (1) the computation of the Born radii is performed by integrating over the solute–solvent boundary rather than the solute volume, (2) Born radii are calculated in principle without introducing any approximation beyond the Coulomb field approximation (as opposed to pairwise descreening schemes, see above), and (3) it contains correction factors aimed at counterbalancing the effects of the Generalized Born and Coulomb field approximations relative to the exact solution of the Poisson–Boltzmann equation. The SGB/NP nonpolar hydration free energy estimator is based on the same decomposition as the AGBNP model (see previous section) into a cavity component, proportional to the solute surface area, and a solute–solvent van der Waals energy component based on the Born radius of each atom.[38]

The Born radius of atom i, B_i, is calculated in the Coulomb Field approximation by transforming, using Green's theorem, the integral of the $(\mathbf{r} - \mathbf{r}_i)^{-4}$ over the solute volume [see eq. (8)], into a surface integral

$$\frac{1}{B_i} = \frac{1}{4\pi} \int_s \frac{\mathbf{r} - \mathbf{r}_i}{(\mathbf{r} - \mathbf{r}_i)^4} \cdot \mathbf{n}(\mathbf{r}) d^2\mathbf{r} \tag{17}$$

where S is the solute–solvent boundary, $\mathbf{r}_i$ is the position of atom i, and $\mathbf{n}(\mathbf{r})$ is outward normal to the surface element at $\mathbf{r}$. The integral is performed over a surface grid constructed by placing a uniform grid of surface elements over each solute atom and deleting those surface elements that are found in the interior of solute atoms. When expressed on a surface grid, the integral in eq. (17) takes the form of a pairwise sum between the centers of atom i and of each surface element. To reduce the number of pair evaluations, a higher density of surface elements is used for nearby atoms, whereas fewer surface elements per atom are used to compute the contribution to the surface integral from distant atoms. Optionally a distance cutoff is also applied to omit the contribution of the surface of atoms beyond a certain distance from atom I.

Empirical corrections to the SGB model have been developed by examining systematic deviations between the primitive SGB results and the PB results for a large database of molecules.[79] Two sets of corrections for the Born radii were developed. These corrections assume the form of scaling factor $R_{sr,i}$ and $R_{lr,i}$ that multiply the original value of the Born radius of atom i given by eq. (17). The short-range corrections are based on the van der Waals radius of atom i, the number of neighbors of atom i, and on their van der Waals radii. The long-range corrections are based on an estimator function that measures the amount of solute surface area with outward normal pointing toward the position of each solute atom. The long-range corrections are designed to offset the tendency of eq. (17) to generally overestimate the Born

radii of atoms situated in invaginated portions of the surface of biomolecules. Empirical corrections for the GB pair interaction energies were also developed.[79] These take the form of quantities, R_{pr}, which depend exponentially on the distance between the two atoms, which are added to the original expression[67] of the pair GB interaction energy. The parameters of the expression of R_{pr} depend also on the number of van der Waals neighbors of each atom.[79]

The electrostatic portion of the SGB/NP model was shown to reproduce with good accuracy the charging free energies of various conformations of peptides and proteins obtained by solving the Poisson–Boltzmann equation.[79] We have also tested the predictions of the SGB model against explicit solvent free energy perturbation calculations of electrostatic solvation component of the free energy of binding of a peptide to the Major Histocompatibility Complex receptor (MHC I).[62] We found very good agreement between the SGB model estimates and the free energy perturbation results.

The expression of the nonpolar hydration free energy estimator of the SGB/NP model has the same general expression as the nonpolar model used in AGBNP, except that the van der Waals term, W_i, has a different but similar dependence on the Born radius.[38] The SGB/NP nonpolar model has been parameterized by fitting the adjustable parameters γ_i and α_i against the experimental hydration free energies of a large database of organic molecules.[38] The parameterization of the SGB/NP model has been recently extended to the OPLS_2003 force field (see section Force Field for Pharmaceutically Relevant Organic Molecules).

Conformational Sampling

The design of accurate and efficient effective potential energy models needs to be matched with algorithms to sample the effective potential energy surface over the relevant conformational space in efficient ways so as to allow for the estimation of thermodynamic quantities. In this section we review some of the sampling algorithms implemented in IMPACT. A description of the free energy perturbation method implementation is also included in this section.

Molecular Dynamics

The Molecular Dynamics (MD) facility in IMPACT employs the RESPA[94] multiple time step integrator. The RESPA implementation in IMPACT includes a recursive algorithm[57] that allows for a variety of RESPA integration levels to be employed. Normally two RESPA levels are used. The inner (fast) level is employed for bonding interactions (bonds, angles, and torsions) and the outer (slow) level for nonbonded and solvation components. Typically from four to eight inner integration steps per outer integration step are performed. This simple setup allows MD simulations without holonomic bond constraints (SHAKE/RATTLE) without significant performance loss. With the Fast Multipole Method (FMM) it is also possible to insert an extra RESPA level (medium) between the fast and slow levels for electrostatic and Lennard–Jones interactions of intermediate range. Microcanonical, canonical, and isobaric ensemble MD sampling schemes are implemented.

HMC, J-WALK, and S-WALK

A number of advanced sampling techniques based on the Hybrid Monte Carlo (HMC) sampling algorithm are available in IMPACT, including J-WALK and S-WALK. Conventional Monte Carlo (MC) methods are not easily applicable to biomolecular systems because of the difficulty of designing nonlocal collective moves without incurring prohibitively low acceptance ratios. HMC[95] is a general technique to generate collective sampling moves that are accepted with optimal probability. In HMC the MD integrator itself (RESPA in this case) is used to generate global MC moves. The idea is to use a large MD time step that does not lead to exact energy

conservation. Starting with random velocities, at the end of a series of MD steps the new total energy E_{new} is compared to the starting energy E_{old} and the new conformation is accepted or rejected by means of a standard Metropolis MC test based on $\exp[-E_{new} - E_{old})/kT]$. Because the RESPA integrator in IMPACT is time reversible and symplectic,[94] it can be shown[95] that HMC, as implemented in IMPACT, produces an equilibrium canonical probability distribution. HMC is an attractive canonical sampling scheme because it has the ability to perform multiparticle moves without incurring the instabilities and discretization errors inherent in MD methods. HMC does not require the atomic forces used in the MD updates to correspond exactly to the gradient of the potential energy being sampled (although better energy conservation and thus a higher HMC acceptance ratio is achieved when the forces correspond as closely as possible to the true gradient of the potential). Therefore, HMC is also very useful for sampling a potential energy surface whose exact gradient is either not available or too expensive to update at every step, such is the case for the numerical implementations of the SGB implicit solvent model. In these cases approximate gradients are used in the MD updates and HMC is used to perform exact canonical sampling.

The Jump-Walking (J-WALK) sampling method[96] is an advanced sampling protocol designed to reduce the sampling bottlenecks caused by the ruggedness of the potential energy surface of biomolecules. The idea is to enhance sampling by proposing new conformations drawn from a canonical ensemble at a higher temperature than the temperature of interest. An HMC run is conducted at a temperature T' greater than the temperature of interest and conformations are saved at regular intervals. Then an HMC run is conducted at the temperature of interest T attempting periodically to jump to one of the conformations saved from the high temperature run. The jump is accepted with probability

$$\min\left\{1, \exp\left[-\left(\frac{1}{kT} - \frac{1}{kT'}\right)(U' - U)\right]\right\},$$

where U is the current potential energy and U' is the potential energy of the selected high temperature conformation. This form of the acceptance probability ensures that J-WALK samples the equilibrium canonical ensemble. The J-WALK method as implemented in IMPACT is related to Replica Exchange sampling with two walkers (see below) except that J-WALK does not require the two walkers to be run in parallel.

The Smart-Walking (S-WALK) sampling method[97] is the same as J-WALK except that conformations generated by the high temperature walker are locally energy minimized before being saved. The saved conformations are then regarded as one of the possible trial moves at low temperature, which are visited according to the standard Metropolis MC acceptance probability function

$$\min\left\{1, \exp\left[-\frac{1}{kT}(U' - U)\right]\right\},$$

where U is the current potential energy and U' is the potential energy of the selected energy minimized conformation. The lower energies U' of the saved energy minimized conformations relative to the same conformations thermalized at the temperature of the high temperature walker make it easier for the low-temperature walker to jump to one of the saved conformations. Consequently, the S-WALK method enables the system to explore more conformational space and undergo more efficient barrier crossings due to the increase in the jump success ratio.

However, the minimized conformations are not exactly canonically distributed, and therefore, the distribution of conformations produced by S-WALK is only approximately canonical.

Replica Exchange Molecular Dynamics

An implementation of the Temperature Replica Exchange Molecular Dynamics sampling method (RXMD) [98] is available in the academic version of IMPACT. The RXMD algorithm is a powerful technique to efficiently sample the rough energy landscape of biomolecules, allowing rapid interconversion between conformations separated by high free-energy barriers, which are not normally accessible at room temperature. Several replicas of the system are run in parallel over a series of temperatures using constant temperature molecular dynamics. At regular intervals exchange of conformations are attempted between pairs of replicas, using a scheme designed to preserve canonical sampling at each temperature. Unlike conformational search techniques, RXMD can be used to calculate thermodynamic quantities, such as conformational free energies, at each simulated temperature. The RXMD method is also suited to take advantage of the large number of processors in modern computing clusters without requiring expensive low-latency networking hardware.

The RXMD implementation in IMPACT follows the scheme proposed by Sugita and Okamoto. [98] Replica simulations are distributed over a series of processors, and the communication between the replicas is implemented using MPI (Message Passing Interface) instructions. A replica exchange module is called at regular intervals within the MD loop. The master processor selects the pair of replicas scheduled for exchange and instructs the corresponding processors to coordinate the exchange. For maximum efficiency, instead of exchanging conformations, the replicas equivalently exchange temperatures. As a result, the temperature of each replica is not constant during the simulation. Efficient interconversion between low-energy conformations occurs whereby a replica assumes a temperature high enough to rapidly move to a new region of conformational space, and then assumes progressively lower temperatures allowing the formation of new stable conformations. The temperature of each replica is recorded at regular intervals in a trajectory file together with the system coordinates. Subsequent analysis of the trajectory files collates the trajectory file frames at a particular temperature and calculates ensemble averages at that temperature.

Free Energy Perturbation

The academic version of IMPACT includes an efficient Free Energy Perturbation (FEP) implementation. FEP is used to compute the free energy difference between two chemical systems by constructing a nonphysical thermodynamic path between the two states. A series of MD simulations are performed along the thermodynamic path, and averages are collected that then allow the computation of the free energy difference. The FEP implementation in IMPACT follows the GASP (Generalized Alteration of Structure and Parameters) scheme.[99] This scheme allows for mutations that involve simultaneous variation of internal coordinates and force field parameters. Topology files define the initial and final states. The files are parsed to detect variation of atom types and internal coordinates. Force field parameters are assigned to the initial and final states, and intermediate values of the parameters are obtained along the mutation path by linear interpolation. Dummy atom types are used to create or destroy atoms. Internal coordinates are varied in a similar way. In FEP applications aimed at computing conformational free energy profiles, the internal coordinates to be varied are often maintained fixed during MD sampling. The GASP method, however, does not impose this limitation. It has been shown[99] that, when the internal coordinate mutations are part of a more complex scheme involving simultaneous variation of both force field parameters and geometry designed to convert one molecule into another, it is actually beneficial to let the internal coordinates being mutated vary during the sampling trajectory. The GASP method does not include Jacobian correction factors for the conversion from internal to Cartesian coordinates.[99]

The FEP estimator $\langle \exp[-(U' - U)/kT] \rangle$,[69] where U is the current potential energy and U' the perturbed potential energy, is evaluated by recomputing the total potential energy of the system in the perturbed state, rather than updating only the energy components affected by the perturbation. This strategy has the advantage of being easily applicable to nonpairwise-decomposable potentials (i.e., Ewald) but each evaluation is computationally expensive. In most cases this does not significantly affect overall performance because the FEP estimator is sampled infrequently (frequent sampling does not necessarily improve convergence due to correlation between samples). FEP can be performed using constant temperature MD, HMC, or from precomputed trajectory files. The FEP facility in IMPACT has been used extensively to gather hydration free energy differences in explicit solvent used to validate and tune implicit solvent models.[62,63,65,100]

Applications

Simulation of Peptide Free Energy Surfaces and Folding

The simulation of polypeptide free energy surfaces demonstrates some of the features of advanced sampling techniques implemented in IMPACT as well as the accuracy achieved by the OPLS all-atom force field (OPLS-AA) [7] combined with the AGBNP implicit solvent model.[61] Because analytical gradients are provided by the AGBNP routine, the OPLS-AA/AGBNP potential is suitable for molecular dynamics (MD) simulations, which are made more efficient with the use of the RESPA algorithm in IMPACT.[57] Our study of polypeptide free energy surfaces is made possible by the use of the replica exchange molecular dynamics method (RXMD)[98] and the AGBNP implicit solvent model described previously.[61] In the RXMD calculations described here typically 20 RXMD replicas are employed at 20 different temperatures between 270 and 690 K. RXMD greatly enhances sampling efficiency compared to standard MD. These features of IMPACT are brought together in the challenge to obtain the correct thermodynamic populations of the conformers of several peptides.

We have simulated the free energy surface[59] of the C-terminal β-hairpin of the B1 domain of protein G (sequence: G41EWTYDDATKTFTVTE56),[101] referred to as the G-peptide.[102] Comparison of the results obtained with the AGBNP nonpolar estimator to those obtained with a standard surface area-dependent nonpolar estimator showed that the new nonpolar term helped to stabilize the association of the hydrophobic core (W43, Y45, F52, and V54 in the G-peptide sequence). The percentage of structures generated with RXMD sampling that had a collapsed hydrophobic core was only 13% when the surface area nonpolar model was used, while this percentage increased to 38% with the use of the AGBNP nonpolar model. This increased stabilization of the hydrophobic core using the new nonpolar term occurred even in the presence of disruptive salt bridges, in particular, between K50 and E56.[59]

From our previous work with protein decoys, [82] explicit solvent simulation results, [100] and the above simulations of the G-peptide, we recognized a problem with the overstabilization of salt bridges using generalized Born models such as SGB and AGB. To address this problem we implemented a modified GB pair interaction energy expression, which makes it possible to selectively apply additional dielectric screening to particular atomic pairs.[59] The PMF with respect to the radius of gyration (R_g) of the hydrophobic core residues and the number of β-hairpin hydrogen bonds of the G-peptide generated by RXMD sampling with the OPLS-AA/AGBNP potential is consistent with PMFs obtained by replica exchange MD with explicit solvent.[103,104]

Further principal component analysis using 42 C_α interatomic distances, showed two thermodynamically stable regions at room temperature: a β-hairpin region, and a less populated region composed of α-helical structures. The PMF, with respect to the second and third principal components, is shown in Figure 2. The occurrence of α-helical structures was

observed in some replica exchange explicit solvent studies,[103,105] but not in others.[104] The β-hairpin population predicted in our study is approximately 40%, based on the number of formed β-hairpin hydrogen bonds.[59] This agrees well with the experimental results of Blanco et al.,[101] who reported a 42% β-hairpin population. The degree of hydrophobic collapse (98%) also agrees reasonably well with the experimental results reported by Muñoz et al.[106] By employing the Temperature Weighted Histogram Analysis Method (T-WHAM) [107] to combine the data from the replica exchange trajectories at all temperatures, we were able to resolve the saddle point region of the PMF that connects the α-helical state to the β-hairpin state (see Fig. 2).[107]

We then turned our attention to assessing the sampling efficiency of RXMD and the accuracy of the OPLS-AA/AGBNP model in determining the correct thermodynamic populations of several other peptides. The peptides we investigated were either shown to be mostly helical or adopt no significant secondary structure when free in solution.[108,109] Along with the G-peptide, we summarize our results with the peptides we have simulated with RXMD sampling and the OPLS-AA/AGBNP model in Table 4. The agreement of the predicted thermodynamically stable populations and the experimental values is very good.

Although RXMD does not provide direct information about kinetics, we have recently employed the very diverse ensemble of conformations generated by RXMD sampling to infer the mechanism for folding and unfolding of the G-peptide.[110] To do so we constructed a kinetic network model, each node of the network representing a conformation generated by RXMD, with transitions between nodes allowed based on conformational similarity. Starting from an unfolded conformation associated with a high temperature replica, the system moves from one conformation to another using stochastic rules designed to asymptotically reproduce the thermodynamic properties of the system at a chosen temperature. We have applied this scheme to study the mechanism of folding of the G-peptide for which, as described above, RXMD with the OPLS-AA/AGBNP effective potential produces at room temperature a major β-hairpin population and a minor α-helical population. We observed that the great majority of the folding trajectories from unfolded conformations to α-hairpin conformations go through metastable α-helical conformations. Using graph theory techniques, we have also identified[110] high-probability pathways connecting the α-helical conformations to β-hairpin conformations, two of which are shown in Figure 3.

Allostery of the Ribose Binding Protein

The previous section focused on the problem of generating conformational free energy landscapes of peptides. The accuracy of the OPLS-AA/AGBNP effective potential and the conformational sampling capabilities of IMPACT with respect to the rather different problem of studying conformational changes of larger proteins are illustrated in this section, which presents a computational study of the allosteric equilibrium of the Ribose Binding protein. Allosteric transitions are essential for the function of many biological macromolecules; they mediate a variety of protein functions including transport, signaling and enzymatic activity. Ligand-induced allosteric conformational changes quite often involve domains of the protein moving with respect to one another as relatively rigid substructures.

The Ribose Binding Protein (RBP) is a 271 residue multidomain protein (see Fig. 4) member of a large class of bacterial periplasmic proteins involved in the sensing and transport of small molecule substrates. The mechanism of sugar binding and transport has been well characterized by crystallographic, spectroscopic, and biochemical studies.[111,112] RBP exists in an open conformation when not bound to the substrate (Fig. 4a); when the protein binds ribose the conversion to a closed, more compact, conformation occurs (Fig. 4b). This allows the protein to bind a membrane-bound permease complex that receives the ribose molecule and transports into the interior of the bacterial cell.

Computations with the IMPACT program have recently been completed[113] to study this allosteric system as a paradigm for more complex allosteric phenomena. These simulations, which require extensive computational resources to properly sample the conformational space of the protein, were made possible by the use of a large Linux cluster to perform MD simulations with many different biasing potentials in parallel, and by employing the AGBNP implicit solvent model.[61]

The closed and open conformations of RBP differ with respect to a rotation of the N-terminal domain, which is linked to the C-terminal domain by a three-stranded hinge domain[112] (Fig. 4).

The protein conformational space was sampled using umbrella sampling based on the order parameters θ and φ, which measure, respectively, the angle of opening between the two domains and the rotation of one domain with respect to the other (these order parameters are described in the caption of Fig. 5). The calculations were conducted using the OPLS-AA force field with the AGBNP implicit solvent model. The data from multiple simulations were analyzed using the Weighted Histogram Analysis Method (WHAM).[107] We obtained the conformational populations of the protein with respect to the order parameters spanning both closed and open conformations (Fig. 5). The shifts in conformational populations in going from the ribose-free to the ribose-bound forms of the protein are consistent with the available experimental data. Most of the predicted stable conformational states correspond to measured X-ray structures of RBP (indicated by crosses in Fig. 5); other conformational states were newly predicted. In particular, we were able to characterize the closed state of the protein in the absence of ribose, and a novel partially open ribose-bound conformation whose existence was previously postulated based on biochemical evidence. Further analysis revealed that the shift in the RBP population is driven by the favorable interactions between ribose and RBP in the closed conformation, whereas in the absence of these interactions the open conformation becomes predominant due to its larger conformational entropy.[113] Notably, the calculations showed that the allosteric transition upon binding of ribose involves a shift in the equilibrium between a set of conformations that exist in appreciable concentrations in solution in both the presence and the absence of the ligand (see Fig. 5).

Molecular Mechanics Specialized Technologies

Overview

In previous sections, we have discussed the features and performance of IMPACT in the context of applications which employ core molecular mechanics effective energy models and sampling techniques, that is, minimization, molecular dynamics, and Monte Carlo, as well as more advanced methods like RXMD, which combines features of MD and MC sampling. Modeling of this type is applicable to a wide range of important biological problems. However, there are also a number of situations in which alternative methods are required to meet accuracy and/or performance objectives, via the development of more specialized sampling algorithms and energy models. We have developed two modules of IMPACT in this category: Glide, a high-throughput docking program,[15–17] and QSite,[114,115] a mixed quantum mechanics/molecular mechanics methodology.

Current drug discovery projects in the pharmaceutical industry typically involve screening of hundreds of thousands, or millions, of ligands against a receptor to discover novel lead compounds with micromolar or nanomolar binding affinity. Virtual screening, in which computational methods are employed to predict binding affinity, can be employed if a high resolution crystal structure of the receptor is available, or if a suitable homology model can be constructed. Current molecular mechanics based simulation methods are unsuitable for large scale virtual screening for two reasons: (1) the computational effort required for carrying out

such simulations to achieve any sort of convergence of sampling is far too large to enable millions of compounds to be screened with an acceptable level of computational effort, and (2) even if such calculations could be carried out, prediction of absolute binding affinity (the requirement for lead discovery projects) in the absence of a training set is a highly demanding task, and it is unclear whether current force fields would provide the accuracy required to rank compounds accurately.

The Glide program employs the IMPACT infrastructure as a starting point; construction of the protein model, and assignment of molecular mechanics parameters to protein and ligand, is carried out as discussed in the sections The Maestro Graphical User Interface and Atom Typing and Parameter Assignment. However, the receptor is then modeled as a rigid object, and the molecular mechanics potential is mapped onto a grid to enable rapid evaluation of protein–ligand interaction energies and gradients using standard interpolation methods. Novel sampling methods, and empirical scoring functions, are then used in combination with the molecular mechanics energy to predict the binding mode, and binding affinity, of a ligand with the target receptor. Glide can be run at several different speeds; the three prepackaged modes are “Fast Screening” (3–5 CPU s/ligand); “Standard Precision” (~30 CPU s/ligand), and “Extra Precision” (~10 CPU min/ligand), with all timings referring to a single processor PC. We refer to these modes as FS, SP, and XP in what follows. Further description of the sampling methods and scoring function, as well as summaries of results, are presented below.

At the other end of the spectrum, current molecular mechanics force fields are not designed to handle chemical reactions. Mixed QM/MM methods, in which a region of the system is treated at the QM level (typically 50–200 atoms in our modeling of enzyme active sites) and the remainder of the system is treated at the MM level, provide the accuracy for *ab initio* quantum chemical approaches, yet enable treatment of large condensed phase systems in reasonable CPU times without truncation of the model, which can lead to erroneous neglect of environmental steric, electrostatic, and conformational effects. The QSite program, which has been developed via a tight coupling of IMPACT with the Jaguar suite of *ab initio* programs, contains an interface between the QM and MM regions that has been highly optimized for modeling protein active site chemistry. A description of the QM/MM methodology, as well as a summary of benchmark results calibrating accuracy and applications to a diverse set of proteins of biological and pharmaceutical interest, is presented in the section Mixed Quantum/Molecular Mechanics below. Finally, in the section Combining Glide and QSite to Obtain More Accurate Binding Mode Prediction, we present a recent, and novel, approach in which we used QSite to calculate polarized charge distributions for use in Glide docking. This methodology, which will be useful for lead docking of a relatively small number of compounds where accurate geometries are desired, for example, in a lead optimization context (as opposed to high-throughput screening), demonstrates significant improvements in docking accuracy compared to the use of force field based charge distributions.

Glide: High-Throughput Docking

High-throughput docking has become a standard tool in the pharmaceutical industry, both in lead discovery (virtual screening) and lead optimization applications. There are a number of such programs in current use, [116–119] all of which share the same basic objectives (docking of a flexible ligand in a rigid receptor to predict the binding mode and binding affinity), but differ considerably in the details of the sampling algorithms and scoring functions. Below, a brief description of the Glide module of IMPACT is presented, along with a representative set of results obtained using the current release of the program (v. 3.5).

Methodology

As was mentioned above, the first key assumption made by Glide is that the receptor conformation is rigid, whereas the ligand has fully flexible torsional degrees of freedom. This approximation is not always successful; there are numerous examples of pharmaceutically interesting targets in which large changes in the receptor structure are observed upon binding of a particular ligand, such that docking the ligand into the original conformation (either the apo structure, or the parent conformation of a different protein–ligand complex) is precluded by unfavorable steric interactions.[17] If the induced-fit effects are minor, scaling the van der Waals radii of the protein and ligand atoms can ameliorate this problem. This strategy is used by default in Glide, although care must be taken as too much scaling can result in an inability of the receptor to recognize the ligand. When the conformational changes are large, alternative methods taking induced-fit effects into account explicitly are required, necessitating substantially more computational effort. This subject, however, is beyond the scope of the current article.

It is common to find physically untenable steric clashes in crystallographically determined protein sites. Due to the rigid receptor approximation in Glide, such clashes can result in unfavorable van der Waals interaction energies for the native ligand and other known actives. To address this issue a protein preparation procedure designed to anneal away such steric clashes has been developed.[15] Beginning with a protein and cocrystallized ligand, the preferred preparation procedure results in a partially optimized protein–ligand complex to which hydrogen atoms have been added. Additionally, the protonation states of ionizable residues and the tautomeric form of histidines are adjusted and reorientable hydrogen atoms, such as in hydroxyl groups, are repositioned.

We begin with a brief overview of the sampling methodologies available in Glide.[15,16,120] The first step in using Glide to model protein–ligand binding is to generate Coulomb and van der Waals potentials on the receptor grid. A multigrid approach is used to minimize the required memory. In this approach higher resolution grids are applied at the protein–ligand surface, where high accuracy is required in modeling protein–ligand interactions, and lower resolution grids are used to treat more distant interactions. As mentioned above, the van der Waals radii of the protein and ligand atoms can be scaled (defaults are 1.0 for nonpolar protein atoms and 0.8 for nonpolar ligand atoms). A $2r$ distance-dependent dielectric constant is used to generate a screened Coulomb interaction. The energy and gradient attributable to a ligand atom in the field of the protein can then be rapidly computed via standard interpolation techniques. The grid is preprocessed prior to docking, and hence, does not contribute materially to the CPU time for a typical virtual-screening experiment. The general features of the grid-based docking methodology described here have been used by others,[118] although the multigrid aspect is a novel feature.

Once the grid is constructed, a suite of possible conformations of the core of the ligand is generated, and the ligand is docked using a series of hierarchical filters. Initial filters are primarily shape-based, and narrow the region of the receptor that can plausibly accommodate the ligand. Intermediate filters are based on fast approximate scoring functions (“rough scoring”) that do not reject poses due to small van der Waals overlaps and that detect potential hydrogen bonds and favorable hydrophobic interactions. Combinatorial search algorithms are used to efficiently position peripheral groups of the ligand, and promising poses are saved for further investigation. The algorithm is formally an exhaustive search, although there are limits to the resolution with regard to both the ligand conformations and positioning of the ligand in the receptor.

At the next stage of the calculation, the top-scoring subset of poses is selected and variable-metric minimization, using the grid-based molecular-mechanics energy function discussed

above, is performed, starting from the various poses that survive the rough scoring filters. Various techniques, such as annealing of the van der Waals radii, are used to enable the minimization to respond appropriately to large initial steric clashes. Finally, the poses are ranked by total energy (including an approximate representation of the ligand internal energy), and some number of top scoring poses is subjected to extensive torsional sampling as a method of optimizing the positions of ligand peripheral groups. The parameters used in these stages depend upon whether one is using fast-screening or standard-precision settings.

The above protocol provides a powerful and general method for predicting the binding mode of protein–ligand complexes, which may be examined in self-docking experiments where a ligand is docked into its parent receptor. A summary of Glide's performance in such experiments is given in Table 5. Additionally, data sets assembled by the developers of the GOLD and FlexX programs have been used to assess docking accuracy. Results over these common sets of systems, shown in Tables 6 and 7, enable a direct comparison of Glide with Gold and FlexX. The average RMSDs achieved by Glide represent a considerable improvement compared to the published alternatives cited. Note, however, that some residual errors do remain. Some—perhaps many—of these problematic cases appear to be primarily due to inaccurate modeling of the ligand charge distribution. Such errors can be corrected (at some computational cost) by the introduction of more accurate, polarized charges (in Section Combining Glide and QSite to Obtain More Accurate Binding Mode Prediction, this is carried out using QM/MM methods).

The FS and SP modes of Glide do not attempt to apply large desolvation penalties to the ligand when, for example, a protein or ligand charge is buried. XP Glide, on the other hand, does contain terms of this type, as is discussed further below. Although such terms can be extremely helpful in eliminating false positives, they also will penalize known active compounds unless the sampling is at a higher resolution than is obtained from the algorithms discussed above; small movements of the ligand may be necessary to evade burial penalties. Presumably, active compounds can execute these small displacements, whereas false positives may not be able to do so. A different type of sampling algorithm is therefore required to make XP Glide a useful approach.

We briefly outline the elements of this algorithm. First, normal Glide SP docking is performed, as discussed above. Poses that score well in SP are then decomposed into "anchor" fragments, typically rings, which are positioned in various locations in the receptor, as given by the ensemble of SP poses. A growing algorithm is used to regenerate the entire molecule, starting from the various attachment points on the anchor. Approximate filters select promising poses of various ligand "side chains," and torsional minimization on the Coulomb–van der Waals grids and rescoring is applied to a selected subset of fully assembled molecules to determine the final pose selection (in which penalized poses are rejected within a specified energy window). Finally, if a penalty term is detected, the side chain to which the penalty is applied is regrown at very high resolution in the ligand torsion angles, enabling the penalty to be evaded when feasible. The growing algorithm requires ~20× more CPU time than SP docking but provides the requisite fine grain sampling to eliminate penalties from properly docked active compounds in the great majority of test cases that we have examined. The XP sampling methodology will be discussed in greater detail elsewhere.[120] XP docking performance is comparable to that of SP for self-docking experiments, although significant differences do appear in cross docking and in penalty avoidance.

Once a docked pose is selected via the FS, SP, or XP modes, the binding affinity can be predicted. The molecular-mechanics potential function alone is insufficient to enable this prediction to be made with any degree of accuracy. Approximations that would lead to large errors in such a naïve approach include the neglect of solvation, entropy of binding, and protein

relaxation. However, specially designed empirical scoring functions can achieve much better results that are not clearly inferior to the current state of the art available from simulation methods. We have developed empirical scoring functions for Glide that improve on the performance of alternative methods described in the literature. Our scoring methodology is briefly described below. A more extended discussion can be found elsewhere.[15,16,120]

The Glide scoring function is a modified version of ChemScore.[121] The ChemScore scoring function is shown in eq. (18):

$$\Delta G_{\text{bind}} = C_0 + C_{\text{lipo}} \sum f(r_{\text{lr}}) + C_{\text{hbond}} \sum g(\Delta r) h(\Delta\alpha) + C_{\text{metal}} \sum f(r_{\text{lm}}) + C_{\text{rotb}} H_{\text{rotb}}. \quad (18)$$

The second term is an atom–atom pair term that assesses contacts between hydrophobic atoms of the ligand and protein. Such contacts involve the displacement of water molecules from positions adjacent to hydrophobic groups in the protein or ligand into bulk solution. Such contacts are presumably favorable as displaced water molecules can now make their full complement of hydrogen bonds without entropic penalty. A term of this sort is present in virtually every empirical scoring function.

The third term assigns favorable binding affinity for the formation of hydrogen bonds between the protein and ligand. The magnitude of the term is modulated by the quality of the hydrogen-bonding geometry, as evaluated based on distances and angles involved in the structure. Glide furthermore awards different free energy increments depending upon whether the hydrogen bond is neutral–neutral, neutral–charged, or charged–charged. The fourth term is a metal–ligand term, which in Glide considers only the strongest interaction with an anionic acceptor atom. The fifth term assigns a penalty for restricting ligand entropy, based on the number of rotatable bonds in the ligand.

Several new terms have been developed to improve the performance of the scoring function for both absolute binding affinity prediction and ranking of active compounds compared to those in a randomly chosen database (i.e., database enrichment studies). These terms both improve estimation of the attractive forces that drive binding affinity and enable the imposition of penalties that reflect the desolvation of polar and charged groups on the protein or ligand. A summary of these terms is as follows:

1. Desolvation penalties: 2.8-Å spheres representing explicit water molecules are added to a subset of high-scoring poses generated by the SP Glide docking protocol. This is achieved using a rapid, grid-based algorithm that fills a prescribed volume of the active site around the ligand. The number of waters surrounding each charged or polar atom can then be computed, and penalties are applied based on statistics developed from extensive studies of known active compounds for a wide variety of receptors. Our belief is that the use of descriptors based on a discrete representation of solvent has significant advantages over continuum based models for this particular type of calculation.
2. Molecular mechanics terms: the Coulomb and van der Waals ligand–protein interaction energies are incorporated into the overall scoring function.
3. Specialized hydrophobic and hydrogen bond terms: new non-linear functional forms have been developed that provide an improved description of hydrophobic interactions and hydrogen-bonding contributions to binding. These terms are discussed in more detail elsewhere.[120]

Selected terms of these types are used in both the XP and SP versions of Glide scoring. The XP scoring function, however, contains a much higher weighting for the desolvation and specialized molecular-recognition terms. As noted previously, obtaining accurate estimates of these terms is dependent upon the use of high-resolution sampling.

Results

Tables 5–7 present summaries of SP Glide root-mean-square deviations (RMSDs) for self-docking tests of docking accuracy, where the ligand is removed from its original cocrystallized complex, placed in the lowest energy conformation obtained from a short conformational search, and then redocked back into the receptor conformation observed in that complex. RMSDs are determined including only nonhydrogen atoms. A summary of the docking accuracy for 279 PDB complexes broken down by the number of rotatable bonds in each ligand is shown in Table 5. These results suggest the docking performance of Glide is very reasonable over a wide range of rotatable bonds. Similar results are found with Glide XP.

A head-to-head comparison of docking accuracy performance has been generated for Glide, FlexX, and GOLD using test sets of noncovalently bound ligands defined by the developers of these methods. The results presented in Tables 6 and 7 indicate that Glide represents a significant improvement in the accuracy of binding mode prediction compared to alternatives in the literature. If one considers only ligands with 10 or fewer rotatable bonds, the range of greatest interest for many library-screening exercises, Glide performs nearly twice as well as GOLD and more than twice as well as FlexX.

Although the Glide docking-accuracy results are generally quite good, a nontrivial fraction of the test cases (10–20%) result in incorrect binding modes, and in many cases we have found that these results arise from scoring rather than from sampling errors. In the section Combining Glide and QSite to Obtain More Accurate Binding Mode Prediction, we show that a primary source of the problem appears to be the use of fixed charges generated by the molecular-mechanics force field; employment of polarized charges obtained from mixed QM/MM calculations in the protein environment yields a qualitative reduction in the number of outliers. Implementation of an automated algorithm to carry out docking using this model, via a combination of Glide and QSite, is currently ongoing.

To evaluate the ability of Glide to rank known active ligands, enrichment has been evaluated by seeding known active compounds, with 10 micromolar or better experimental binding affinities, into a random database of 1000 ligands selected from a large database of purchasable compounds, using metrics for drug-like molecules (molecular weight, numbers of charged and neutral donors and acceptors, numbers of rotatable bonds and rings, percentage of phobic carbons, etc.) inferred from a selection of compounds from the Derwent World Drug Index (Derwent Information Limited, Alexandria, VA). The presumption is that such a database should contain few or no submicromolar inhibitors. The 13 screens in this article cover a wide range of receptor types and binding-site character.

Table 8 gives the percentage of ranked actives found by SP and XP Glide in the top 2, 5, and 10% of the 13 database screens. Because the present screens use only ~1000 ligands, 2% of the database (20 ranked ligands or so) is about the smallest percentage that can be examined, given that most screens contain about 10 active compounds. This will also lead to a cap on the maximum percent of active compounds found in the top 2% of the database in cases where the number of active compounds is greater than about 20 (1bl7, 1rt1 and 1e66).

It can be seen that the performance of SP Glide is quite respectable, as Glide is able to extract >50% of the known active compounds from the top 5% of the database for 4 of the 11 systems. XP Glide is more consistent across all of the receptors in the test suite, being able to extract at

least 29% of known active compounds from the top 2% of the databases for all included screens and at least 50% of known active compounds from the top 5% of the databases. Initial calculations show that the recently developed fast-screening mode gives results that are not quite as good those for SP Glide but nevertheless are very promising.

Mixed Quantum/Molecular Mechanics

Mixed quantum mechanics/molecular mechanics (QM/MM) programs have been under development in a number of laboratories now for the past several decades.[19,115,122–130] The basic idea is to use QM methods to treat reactive chemistry in a localized region of the system, while incorporating structural, steric, and electrostatic effects of the remainder of the system via molecular mechanics. The use of modern *ab initio* quantum chemical methods, particularly the hybrid density functional theory (DFT),[131] provides reasonable accuracy and robustness in first principles modeling of reactive chemistry for a wide range of systems (including those containing transition metals), while enabling QM regions of several hundred atoms to be treated on a routine basis. Although the QSite program allows the use of Hartree–Fock, local MP2 (LMP2), and DFT methods for the QM component of the calculations, the focus in what follows will be on the DFT implementation. We briefly summarize below the methodology employed in QSite and various biological applications; this work has recently been reviewed in more depth in ref. [132].

Methodology

The first key component of a robust and accurate mixed QM/MM methodology is the interface between the QM and MM regions. The greatest challenge is manifested when the interface occurs at a covalent bond, that is, an atom on one side of the bond is MM, and the atom on the other side is QM. A number of techniques have been developed to address this problem, including link atom methods,[125,126,130] and approaches based on frozen, localized molecular orbitals.[128,129] QSite employs the latter approach, as is described in detail in refs. [19, 115], and [129].

It is likely that good accuracy, compared to fully QM calculations, can be achieved using either link atom or frozen orbital methods.[133] However, errors are critically dependent upon details of the parameterization of the interface. We have developed parameters for QSite for all 20 amino acids by fitting to QM data for model dipeptides[115] enabling QM/MM partitioning in the backbone and between the α- and β-carbons of the various residues. The fitting database comprised 200 rotamer states of the dipeptide ensemble (identical, in fact, to the data used to develop torsional parameters for the OPLS2001 protein force field, discussed in the section Force-Field Development); parameters were optimized to reproduce the QM relative energies of these rotamer states. Deprotonation of the dipeptide side chains was also carried out, and the electrostatic balance of the model was properly adjusted (via a bond charge at the QM/MM boundary) by fitting the deprotonation energies to fully QM calculations.

Parameterization is also required at the noncovalent interface between the QM and MM regions; otherwise, the electrostatic interactions between these regions cannot be described accurately. We have developed a set of van der Waals parameters for the QM region by fitting the binding energy of a large number of hydrogen bonded small molecule dimers, representing the functionality of the backbone and various side chain groups, to fully QM results. We have also developed Pauli repulsion parameters for MM polar hydrogen atoms, which prevent charge density from leaking out of the QM region onto these hydrogen atoms. Details of this parameterization are provided in ref. [115].

Finally, we have tested the resulting energy model on a number of larger systems, including larger peptides and a model for dioxygen binding in hemerythrin, comparing QM/MM

energetics for various processes (e.g., deprotonation, dioxygen binding) with QM results. Over the entire test suite, errors average ~0.5 kcal/mol, which is significantly smaller than the intrinsic errors in the QM methods (principally hybrid DFT) that are suitable for modeling large QM regions. Based on these results, we would argue that the QM/MM interface in QSite is not at present the limiting factor in the accuracy of the methodology.

Once the energy model is defined, an implementation is required that enables ease of use and performance of large-scale calculations in reasonable CPU times. An analytical gradient of the energy model has been developed that allows geometry optimizations to be efficiently carried out.[19,115,129] Optimization of systems where the MM region is much larger than the QM region, such as a protein, is facilitated by carrying out adiabatic minimization of the MM region after each QM step; the MM minimizations take relatively little computational effort, rendering the required CPU time comparable to that needed to minimize the QM region alone. A parallel version of the methodology has also been developed, which provides respectable acceleration for two to eight processors, depending upon the details of the system being studied and the speed of the interconnect of the various processors that is available in the computational hardware.

The QSite energy model and sampling algorithms are controlled by a graphical user interface, which facilitates specification of the QM and MM regions (as well as other aspects of the calculation, such as the type of QM methodology to be used, and basis set). QM regions of between 50 and 200 atoms are typically used to model reactive chemistry in the protein active site, ordinarily including the ligand or cofactor as well as key amino acid side chains (and backbone atoms when relevant). Examples are presented in the following section.

QM/MM Applications

QSite has been applied to the study of a wide range of protein active site chemistry. The most extensive series of studies have been carried out for the enzyme methane monooxygenase (MMO). MMO is a nonheme iron-containing protein that catalyzes the conversion of methane to methanol, activating dioxygen, breaking the C—H bond in methane, and finally adding an —OH group to the methyl radical formed in the bond breaking step.[134] We have modeled the entire catalytic cycle of MMO using QM cluster methods with a large model system (~100 atoms), achieving good agreement with experiment for the activation energies of various steps in the cycle,[114] and have investigated hydroxylation of a number of small molecules other than methane. QM/MM calculations have thus far been applied principally to the hydroxylation step,[135] where the mode of binding of the ligand to the protein cavity is an important issue. QM/MM studies of the remaining steps in the cycle, in which dioxygen is activated, are currently ongoing.

Over the past several years, we have investigated a number of other systems using QM/MM methods, including dioxygen binding to hemerythrin,[76] dioxygen activation and substrate hydroxylation by cytochrome P450,[136,137] and the hydrolysis of the antibiotic cephalothin by a penicillin binding protein (PBP) and by a class C β-lactamase.[18] In all cases, agreement with experimental data for binding free energies and/or activation free energies, when available, has been encouraging (see below for a more extensive discussion of this point). For P450cam, we have modeled the entire catalytic cycle of the enzyme, achieving good qualitative agreement with available experimental data. A particularly interesting result that we have obtained is the substantial (~7–10 kcal/mol) lowering of the activation barrier for the initial step of the hydroxylation reaction in which a hydrogen atom is removed from the camphor substrate; this result explains the failure of experimental efforts to trap the catalytically competent intermediate, compound I,[138,139] which is presumed to therefore have a very short lifetime. In our calculations, the lowering of the activation barrier is achieved by the protein environment

electrostatically tuning a lone pair orbital of the peripheral carboxylate substituents of the porphyrin to be in near resonance with the cation radical orbital of the ferryl species constituting compound I; this resonance enables charge to migrate to the carboxylate as the hydrogen atom is transferred, thus stabilizing a salt bridge between the carboxylate and an arginine residue of the protein. Similarly, we provide a detailed explanation as to the difference in cephalothin hydrolysis barriers for the PBP and β-lactamase; the low barrier of the latter species is the source of antibiotic resistance to β-lactam–based antibiotics in bacteria.[140] This work demonstrates that QM/MM methods have reached the point where both qualitative and quantitative insight into biologically important enzymatic reactions can be achieved for a wide range of systems.

A final application of QSite, illustrating deployment of the methodology in a system where catalysis involves a substantial conformational change of the protein active site, is the modeling of the catalytic cycle of the enzyme triose phosphate isomerase (TIM).[141] TIM catalyzes the conversion of dihydroxyacetone phosphate (DHAP) to D-glyceraldehyde 3-phosphate (GAP); the rate-limiting step is a slow conformational change of the enzyme–substrate complex. The largest barrier from the point of view of chemical reactions (as opposed to conformational change) is removal of a proton from DHAP by a suitably positioned glutamate group of the protein; the free energy of this step, which is only slightly smaller than that associated with the conformational change, can be measured experimentally by looking at isotope effects and monitoring the reaction in the reverse direction (GAP to DHAP) where this proton transfer is rate limiting.[142,143] QSite yields good agreement with the experimental activation free energy for this step, as well as qualitatively matching other aspects of the experimental data. A crucial aspect of the TIM catalytic mechanism is the substantial motion of the catalytic loop region (loop 6, residues 166–177) of the TIM protein. In the absence of substrate, the loop in the crystallized version of the enzyme is predominantly in the "open" state, which enables substrate to enter the active site cavity and bind. Our methods for *ab initio* protein loop prediction (described elsewhere) are able to predict the structure of the open form, to within 0.43 Å backbone RMSD, starting from the apo structure of the enzyme. However, once the substrate is bound, the loop changes conformation to a "closed" form in which the substrate is now enclosed in the cavity, facilitating the catalytic mechanism. We have used QSite to compute the electronic charge distribution on the substrate in the active site cavity, and then repredicted the loop geometry in the presence of substrate, using the same loop prediction methodology. Agreement is obtained to better than 1.0 Å RMSD from the cocrystallized structure for the bound forms of both GAP and DHAP, and the energy gap between the open and closed forms is in qualitative agreement with estimated based on various experimental measurements of the activation energy required to open the loop in the presence of substrate. The more general use of QSite to compute ligand charge distributions in the protein environment, enabling improved modeling of the structures of protein–ligand complexes, is discussed further below in the following section.

An interesting, more general question arising from this work is the ability of DFT-based QM and QM/MM methods to compute activation barriers for enzymatic reactions. Assessments based on small molecule test cases have yielded rather contradictory results: Truhlar and coworkers have assembled a set of small molecule radical reactions for which hybrid DFT functionals (e.g., B3LYP) display quite large errors (~4–5 kcal/mol on average), whereas Houk and coworkers have obtained relatively small errors for a series of pericyclic reactions (~1–2 kcal/mol).[144] Arguably, most enzymatic reactions are closer in character to the latter as opposed to the former data set; however, explicit confrontation of theory and experiment is required to evaluate individual cases. Table 9 below presents a summary of results that we have obtained to date using Jaguar and QSite to compute activation barriers and free energy differences for enzymatic reactions (many of the systems were discussed above). The results are encouraging,

although a much larger number of test cases will have to be investigated before statistically valid conclusions can be drawn.

Combining Glide and QSite to Obtain More Accurate Binding Mode Prediction

The results shown in the section Glide: High-Throughput Docking for Glide self-docking experiments on protein–ligand complexes indicate a high degree of success in predicting the correct binding mode over a wide range of complexes, and an improvement compared to alternatives in the literature. However, there are still a nontrivial set of complexes for which Glide yields poor results, despite the fact that in self-docking, with proper protein preparation, steric clashes should not be a major problem. We discuss below our most recent efforts to address this problem,[145] which appear to potentially be a major step forward in both understanding the underlying difficulties and improving the results.

One possible source of error in Glide docking is the use of fixed charge force field charges for the ligands; the use of charges derived from quantum chemical calculations, in which polarization by the protein environment is incorporated, could yield a more accurate assessment of alternative hydrogen bond patterns available to the ligand. As an initial test of this idea, we selected 40 complexes from the standard Glide test suite, containing a distribution of RMSD errors in normal Glide docking, weighted towards cases with larger errors, as those are the population we would like to improve. QSite was then used to compute charges on the ligands starting from the cocrystallized complexes. The results of this experiment are summarized in Figure 6. A remarkable improvement in RMSD, across the board, is demonstrated; with the average error being reduced from 1.77 to 0.43 Å, and the maximum error being reduced from 6 to 2 Å, with most cases below 0.5 Å, or close to experimental error.

These calculations, of course, assume advance knowledge of the structure of the protein–ligand complex. To test the same idea using an unbiased algorithm with no presumptions concerning the initial structure, the following iterative protocol (“survival of the fittest,” or SOF) was followed. First, the ligands were docked using the normal Glide algorithm. Up to 10 poses were retained, QM/MM calculations of the ligand charge distributions for all retained poses were performed, and the ensemble of charge models was redocked, with the final single pose being selected on the basis of the Coulomb–van der Waals energy of the complex. The results of this approach are shown in Figure 7. Although there is some degradation from the results of Figure 6, the results still represent a qualitative improvement over the initial use of default force field charges, and suggest that polarization effects and accuracy of the ligand charge distribution is the dominant source error in standard force field-based docking algorithms. Improvements on the results of Figure 7 can likely be obtained by further refinement of the energy model.

Conclusions

As is clear from the title of this review, IMPACT is an acronym that stands for Integrated Modeling Program, Applied Chemical Theory. We believe the name is more apt now than ever. The first research with IMPACT in the 1980s focused on molecular dynamics simulations of proteins. As described in our review, IMPACT now consists of a set of sophisticated molecular modeling tools strongly grounded in chemical theory, which are being applied to help solve current problems in structural biology and in drug design. It has become possible to model biologically important conformational changes by constructing the corresponding complex free energy surfaces using the kinds of effective potential functions and advanced sampling techniques described in this review. The use of the specialized molecular mechanics program module Glide for drug lead identification and optimization that builds upon IMPACT, takes advantage of a hierarchical modeling approach using sampling algorithms and effective potentials that are highly tuned at each stage. As we look to the future, we can expect continued

algorithmic advances in sampling techniques to fuel the development of qualitatively more accurate effective potentials. This will occur as the interplay between theory and experiment is repeatedly tested and refined using much larger benchmark datasets than have typically been used in the past. The collaborations among our research groups, both academic and industrial, have contributed greatly to the expanded functionality of the program package. The basic and applied research projects in our laboratories have, in turn, benefited from this expanded functionality, particularly by the development of tools to automate the system set up, by the expanded coverage of the force field, and by the increased coordination between the various modeling modules described in this review.

Acknowledgments

Many people have contributed to IMPACT. We wish to acknowledge with gratitude key developers during the early IMPACT years in the 1980s and into the 1990s. These include Fumio Hirata, Douglas Kitchen, Francisco Figueirido, David Kofke, and John Westbrook. We also express our appreciation to Anders Wallqvist for his contributions, which greatly facilitated writing the Maestro–IMPACT link. We thank Ruhong Zhou for his work with IMPACT while at Columbia and at Schrödinger on sampling algorithms and on atom typing and parameter assignments, and Avi Ghosh for his work while at Columbia on the implementation of the SGB module in IMPACT. We thank Mike Andrec for providing the illustrations for the section on peptide thermodynamics and folding. We are grateful to the developers, quality-assurance analysts, and application developers at Schrödinger who have contributed computer code, documentation, and ideas to the development of IMPACT; we thank in particular Mike Beachy, Michael Bruce, Mike Campbell, Perry Francis, Shi-Yi Liu, Quentin McDonald, Jason Perry, Tom Pollard, Lynnette Sanders, Jeff Saunders, Mee Shelley, Peter Shenkin, Herc Silverstein, and Andy Spencer. The current fixed charge force field model implemented in IMPACT has its origins in the OPLS-AA force field of William Jorgensen and coworkers. We thank Bill for the close collaboration during the development of OPLS_2003.

Contract/grant sponsor: NIH; contract/grant numbers: GM30580 (to R.M.L.), GM52018 and GM40526 (to R.A.F.), and GM43340 (to B.J.B.)

Contract/grant sponsor: the NSF; contract/grant numer: CHE0316896 (to B.J.B.)

References

1. McCammon JA, Gelin BR, Karplus M. Nature 1977;267:585. [PubMed: 301613]
2. Karplus M. Biopolymers 2003;68:350. [PubMed: 12601794]
3. Levy RM, Bassolino D, Kitchen DB, Pardi A. Biochemistry 1989;28:9361. [PubMed: 2611235]
4. Fan P, Kominos D, Kitchen DB, Levy RM, Baum J. Chem Phys 1991;158:295.
5. Kitchen DB, Hirata F, Kofke DA, Westbrook JD, Yarmush M, Levy RM. J Comput Chem 1990;11:1169.
6. Kitchen DB, Reed LH, Levy RM. Biochemistry 1992;31:10083. [PubMed: 1382594]
7. Jorgensen WL, Maxwell DS, Tirado-Rives J. J Am Chem Soc 1996;118:11225.
8. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WJ. J Phys Chem B 2001;105:6474.
9. Banks JL, Kaminski GA, Zhou R, Mainz DT, Berne BJ, Friesner RA. J Chem Phys 1999;110:741.
10. Stern HA, Kaminski GA, Banks JL, Zhou R, Berne BJ, Friesner RA. J Phys Chem B 1999;103:4730.
11. Kaminski GA, Stern HA, Berne BJ, Friesner RA. J Phys Chem A 2004;108:621.
12. Kaminski GA, Stern HA, Berne BJ, Friesner RA, Cao YX, Murphy RB, Zhou R, Halgren TA. J Comput Chem 2002;23:1515. [PubMed: 12395421]
13. Maple JR, Cao YX, Damm W, Halgren TA, Kaminski GA, Zhang LY, Friesner RA. J Chem Theory Comput 2005;1:694.
14. Stern HA, Rittner F, Berne BJ, Friesner RA. J Chem Phys 2001;115:2237.
15. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. J Med Chem 2004;47:1739. [PubMed: 15027865]
16. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL. J Med Chem 2004;47:1750. [PubMed: 15027866]
17. Teague S. Nat Rev 2003;2:527.

18. Gherman BF, Goldberg SD, Cornish VW, Friesner RA. J Am Chem Soc 2004;126:7652. [PubMed: 15198613]
19. Murphy RB, Philipp DM, Friesner RA. Chem Phys Lett 2000;321:113.
20. Westbrook JD, Feng Z, Chen L, Yang H, Berman HM. Nucleic Acids Res 2003;31:489. [PubMed: 12520059]
21. Halgren TA. J Comput Chem 1996;17:520.
22. Halgren TA. J Comput Chem 1999;20:730.
23. Murphy RB, Beachy M, Ringnalda M, Friesner R. J Chem Phys 1995;103:1481.
24. Halgren TA. J Comput Chem 1996;17:490.
25. Halgren TA. J Comput Chem 1996;17:553.
26. Halgren TA. J Comput Chem 1996;17:616.
27. Halgren TA. J Comput Chem 1999;20:720.
28. Halgren TA, Nachbar RB. J Comput Chem 1996;17:587.
29. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. J Phys Chem B 2002;106:11673.
30. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. Proteins Struct Funct Bioinform 2004;55:351.
31. Damm W, Frontera A, Tirado-Rives J, Jorgensen WL. J Comput Chem 1997;18:1955.
32. Jorgensen WL, McDonald NA. Theochem 1998;424:145.
33. Jorgensen WL, McDonald NA. J Phys Chem B 1998;102:8049.
34. Rizzo RC, Jorgensen WL. J Am Chem Soc 1999;121:4827.
35. Watkins EK, Jorgensen WL. J Phys Chem A 2001;105:4118.
36. Weininger DJ. Chem Inf Comput Sci 1988;28:31.
37. Bush BL, Sheridan RP. J Chem Inf Comput Sci 1989;29:97.
38. Gallicchio E, Zhang LY, Levy RM. J Comput Chem 2002;23:517. [PubMed: 11948578]
39. Bernardo DN, Ding YB, Krogh-Jespersen K, Levy RM. J Phys Chem 1994;98:4180.
40. Caldwell JW, Kollman PA. J Phys Chem 1995;99:6208.
41. Dang LX. J Phys Chem B 1998;102:620.
42. Grossfield A, Ren PY, Ponder JW. J Am Chem Soc 2003;125:15671. [PubMed: 14664617]
43. Lamoureux G, MacKerell AD, Roux B. J Chem Phys 2003;119:5185.
44. Ren PY, Ponder JW. J Comput Chem 2002;23:1497. [PubMed: 12395419]
45. Ren PY, Ponder JW. J Phys Chem B 2004;108:13427.
46. Rick SW, Stuart SJ, Berne BJ. J Chem Phys 1994;101:6141.
47. Gao J, Pavelites J, Habibollazadeh D. J Phys Chem 1996;100:2689.
48. Giese TJ, York DM. J Chem Phys 2004;120:9903. [PubMed: 15268007]
49. Morita A. J Comput Chem 2002;23:1466. [PubMed: 12370948]
50. Panhuis MIH, Popelier PLA, Munn RW, Angyan JG. J Chem Phys 2001;114:7951.
51. Kaminski G, Maple J, Braden D, Murphy RB, Friesner RA. J Chem Theory Comput 2005;1:248.
52. Stern, H.; Rittner, F.; Pavese, M.; Harder, E.; Xu, H.; Kim, B. SIM: Molecular Dynamics Simulation Program. Columbia University; New York: 2001.
53. Harder E, Kim B, Friesner RA, Berne BJ. J Chem Theory Comput 2005;1:169.
54. Friedrichs M, Zhou R, Edinger SR, Friesner RA. J Phys Chem B 1999;103:3057.
55. Greengard L, Rokhlin V. J Comput Phys 1997;135:280.
56. Zhou R, Berne BJ. J Chem Phys 1995;103:9444.
57. Figuerido F, Zhou R, Levy RM, Berne BJ. J Chem Phys 1997;106:9835.
58. Roux B, Simonson T. Biophysical Chemistry 1999;78:1. [PubMed: 17030302]
59. Felts AK, Harano Y, Gallicchio E, Levy RM. Proteins Struct Funct Bioinform 2004;56:310.
60. Gallicchio E, Kubo MM, Levy RM. J Phys Chem B 2000;104:6271.
61. Gallicchio E, Levy RM. J Comput Chem 2004;25:479. [PubMed: 14735568]
62. Zhang L, Gallicchio E, Friesner R, Levy RM. J Comput Chem 2001;22:591.

63. Zhang, L.; Gallicchio, E.; Levy, RM. AIP Conference Proceedings (Simulation and Theory of Electrostatic Interactions in Solutions); 1999. p. 451
64. Levy RM, Zhang LY, Gallicchio E, Felts AK. J Am Chem Soc 2003;25:9523. [PubMed: 12889983]
65. Su Y, Gallicchio E. Biophys Chem 2004;109:251. [PubMed: 15110943]
66. Berendsen HJC, Grigera JR, Straatsma TP. J Phys Chem 1987;91:6269.
67. Still WC, Tempczyk A, Hawley RC, Hendrickson T. J Am Chem Soc 1990;112:6127.
68. Schaefer M, Karplus M. J Phys Chem 1996;100:1578.
69. Levy RM, Gallicchio E. Annu Rev Phys Chem 1998;49:531. [PubMed: 9933909]
70. Berendsen, HJC.; Postma, JPM.; von Gunsteren, WF.; Hermans, J., editors. Intermolecular Forces. Reidel; Dorfrecht, Holland: 1981.
71. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. J Chem Phys 1983;79:926.
72. Cortis CM, Friesner RA. J Comput Chem 1997;18:1570.
73. Nicholls A, Honig B. J Comput Chem 1991;12:435.
74. Cortis CM, Friesner RA. J Comput Chem 1997;18:1591.
75. Cortis CM, Langlois JM, Beachy MD, Friesner RA. J Chem Phys 1996;105:5472.
76. Tannor DJ, Marten B, Murphy R, Friesner RA, Sitkoff D, Nicholls A, Ringnalda M, Goddard WA, Honig B. J Am Chem Soc 1994;116:11875.
77. Marten B, Kim K, Cortis C, Friesner RA, Murphy RB, Ringnalda MN, Sitkoff D, Honig B. J Phys Chem 1996;100:11775.
78. Bashford D, Case DA. Annu Rev Phys Chem 2000;51:129. [PubMed: 11031278]
79. Ghosh A, Rapp CS, Friesner RA. J Phys Chem B 1998;102:10983.
80. Lee MS, Salsbury FF, Brooks CL III. J Chem Phys 2002;116:10606.
81. Onufriev A, Case DA, Bashford D. J Comput Chem 2002;23:1297. [PubMed: 12214312]
82. Felts AK, Gallicchio E, Wallqvist A, Levy RM. Proteins 2002;48:404. [PubMed: 12112706]
83. Froloff N, Windermuth A, Honig B. Protein Sci 1997;6:1293. [PubMed: 9194189]
84. Hawkins GD, Cramer CJ, Truhlar DG. J Phys Chem 1996;100:19824.
85. Lee MS, Feig M, Salsbury FR Jr, Brooks CL III. J Comput Chem 2003;24:1348. [PubMed: 12827676]
86. Qiu D, Shenkin PS, Hollinger FP, Still CW. J Phys Chem A 1997;101:3005.
87. Feig M, Onuvrief A, Lee MS, Im W, Case DA, Brooks CL III. J Comput Chem 2004;25:265. [PubMed: 14648625]
88. Dominy BN, Brooks CL III. J Phys Chem B 1999;103:3765.
89. Tsui V, Case DA. Biopolymers 2000;56:275. [PubMed: 11754341]
90. Grant JA, Pickup BT. J Phys Chem 1995;99:3503.
91. Petitjean M. J Comput Chem 1994;15:507.
92. Schaefer M, Bartels C, Leclerc F, Karplus M. J Comput Chem 2001;22:1857. [PubMed: 12116417]
93. Ashbaugh HS, Kaler EW, Paulaitis ME. J Am Chem Soc 1999;121:9243.
94. Tuckerman M, Berne BJ, Martyna GJ. J Chem Phys 1992;97:1990.
95. Duane S, Kennedy AD, Pendleton BJ, Roweth D. Phys Lett B 1987;195:216.
96. Frantz DD, Freeman DL, Doll JD. J Chem Phys 1990;93:2769.
97. Zhou R, Berne BJ. J Chem Phys 1997;107:9185.
98. Sugita Y, Okamoto Y. Chem Phys Lett 1999;314:141.
99. Severance DL, Essex JW, Jorgensen WL. J Comput Chem 1995;16:311.
100. Yu Z, Jacobson MP, Josovitz J, Rapp CS, Friesner RA. J Phys Chem B 2004;108:6643.
101. Blanco FJ, Rivas G, Serrano L. Nat Struct Biol 1994;1:584. [PubMed: 7634098]
102. Honda S, Kobayashi N, Munekata E. J Mol Biol 2000;295:269. [PubMed: 10623525]
103. Garcia AE, Sanbonmatsu KY. Proteins Struct Funct Genet 2001;42:345. [PubMed: 11151006]
104. Zhou R, Berne BJ, Germain R. Proc Natl Acad Sci USA 2001;98:14931. [PubMed: 11752441]
105. Zhou R. Proteins Struct Funct Genet 2003;53:148. [PubMed: 14517967]
106. Muñoz V, Thompson PA, Hofrichter J, Eaton WA. Nature 1997;390:196. [PubMed: 9367160]

107. Gallicchio E, Andrec M, Felts AK, Levy RM. J Phys Chem B 2005;109:6722. [PubMed: 16851756]
108. Brown JE, Klee WA. Biochemistry 1971;10:470. [PubMed: 5543977]
109. Muñoz V, Serrano L. J Mol Biol 1995;245:275. [PubMed: 7844817]
110. Andrec M, Felts AK, Gallicchio E, Levy RM. Proc Natl Acad Sci USA 2005;102:6801. [PubMed: 15800044]
111. Mowbray S, Cole L. J Mol Biol 1992;225:155. [PubMed: 1583688]
112. Björkman A, Mowbray S. J Mol Biol 1998;279:651. [PubMed: 9641984]
113. Ravindranathan KP, Gallicchio E, Levy RM. 2005in press
114. Gherman BF, Baik MH, Lippard SJ, Friesner RA. J Am Chem Soc 2004;126:2978. [PubMed: 14995216]
115. Murphy RB, Philipp DM, Friesner RA. J Comput Chem 2000;21:1442.
116. Ewing TJA, Kuntz ID. J Comput Chem 1997;18:1175.
117. Kramer B, Rarey M, Lengauer T. Proteins 1999;37:228. [PubMed: 10584068]
118. Meng EC, Shoichet BK, Kuntz ID. J Comput Chem 1992;13:505.
119. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. Proteins 2002;49:457. [PubMed: 12402356]
120. Murphy RB, Friesner RA, Halgren TA. 2005in preparation
121. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. J Comput Aided Mol Des 1997;11:425. [PubMed: 9385547]
122. Warshel A, Levitt M. J Mol Biol 1976;103:227. [PubMed: 985660]
123. Field MJ, Bash PA, Karplus M. J Comput Chem 1990;11:700.
124. Gao J, Xia XF. Science 1992;258:631. [PubMed: 1411573]
125. Crespo A, Scherlis DA, Marti MA, Ordejon P, Roitberg AE, Estrin DA. J Phys Chem B 2003;107:13728.
126. Das D, Eurenius KP, Billings EM, Sherwood P, Chatfield DC, Hodoscek M, Brooks BR. J Chem Phys 2002;117:10534.
127. Field MJ, Bash PA, Karplus M. J Comput Chem 1990;11:700.
128. Gao JL, Amara P, Alhambra C, Field MJ. J Phys Chem A 1998;102:4714.
129. Philipp DM, Friesner RA. J Comput Chem 1999;20:1468.
130. Svensson M, Humbel S, Froese RDJ, Matsubara T, Sieber S, Morokuma K. J Phys Chem 1996;100:19357.
131. Johnson BG, Gill PMW, Pople JA. J Chem Phys 1993;98:5612.
132. Friesner R, Guallar V. Annu Rev Phys Chem 2005;56:389. [PubMed: 15796706]
133. Reuter N, Dejaegere A, Maigret B, Karplus M. J Phys Chem A 2000;104:1720.
134. Liu KE, Valentine AM, Wang D, Huynh BH, Edmondson DE, Salifoglou A, Lippard SJ. J Am Chem Soc 1995;117:10174.
135. Gherman B, Lippard S, Friesner R. J Am Chem Soc 2005;127:1025. [PubMed: 15656641]
136. Guallar V, Baik MH, Lippard SJ, Friesner RA. Proc Natl Acad Sci USA 2003;100:6998. [PubMed: 12771375]
137. Guallar V, Friesner RA. J Am Chem Soc 2004;126:8501. [PubMed: 15238007]
138. Davydov R, Makris TM, Kofman V, Werst DE, Sligar SG, Hoffman BM. J Am Chem Soc 2001;123:1403. [PubMed: 11456714]
139. Schlichting I, Berendzen J, Chu K, Stock AM, Maves SA, Benson DE, Sweet RM, Ringe D, Petsko GA, Sligar SG. Science 2000;287:1615. [PubMed: 10698731]
140. Ghuysen JM. Annu Rev Microbiol 1991;45:37. [PubMed: 1741619]
141. Guallar V, Jacobson M, McDermott A, Friesner RA. J Mol Biol 2004;337:227. [PubMed: 15001364]
142. Albery WJ, Knowles JR. Biochemistry 1976;15:5627. [PubMed: 999838]
143. Knowles JR. Nature 1991;350:121. [PubMed: 2005961]
144. Guner V, Khuong KS, Leach AG, Lee PS, Bartberger MD, Houk KN. J Phys Chem A 2003;107:11445.

145. Cho AE, Guallar V, Berne BJ, Friesner RA. J Comput Chem 2005;26:915. [PubMed: 15841474]
146. Bierzynski A, Kim PS, Baldwin RL. Proc Natl Acad Sci USA 1982;79:2470. [PubMed: 6283528]
147. Kraulis PJ. J Appl Crystallogr 1991;24:946.

```
write file "ts3.out" -
      title "TS3 MIN + MD" *

SET FFIELD OPLS2003

CREATE
  build primary name species1 type auto read maestro file
     "ts3.mae"
  build types name species1
QUIT

SETMODEL
  setpotential
    mmechanics
  quit
  read parm file "paramstd.dat" noprint
  energy parm dielectric 1 nodist -
   listupdate 10 cutoff 12
  energy rescutoff byname all
  zonecons auto
QUIT

MINIMIZE
  conjugate dx0 0.05 dxm 1.0
  input cntl mxcyc 400 rmscut 0.01 deltae 0.1
  run
  write maestro file "ts3_min.mae"
QUIT

DYNAMICS
 input cntl -
   nstep 1000 delt 0.001 stop rotations -
   constant totalenergy nprnt 50 tol 1.e-7
 run rrespa fast 4
 write trajectory coordinates and velocities -
   file "ts3.trj"
QUIT

END
```

Figure 1.
A sample IMPACT input file.

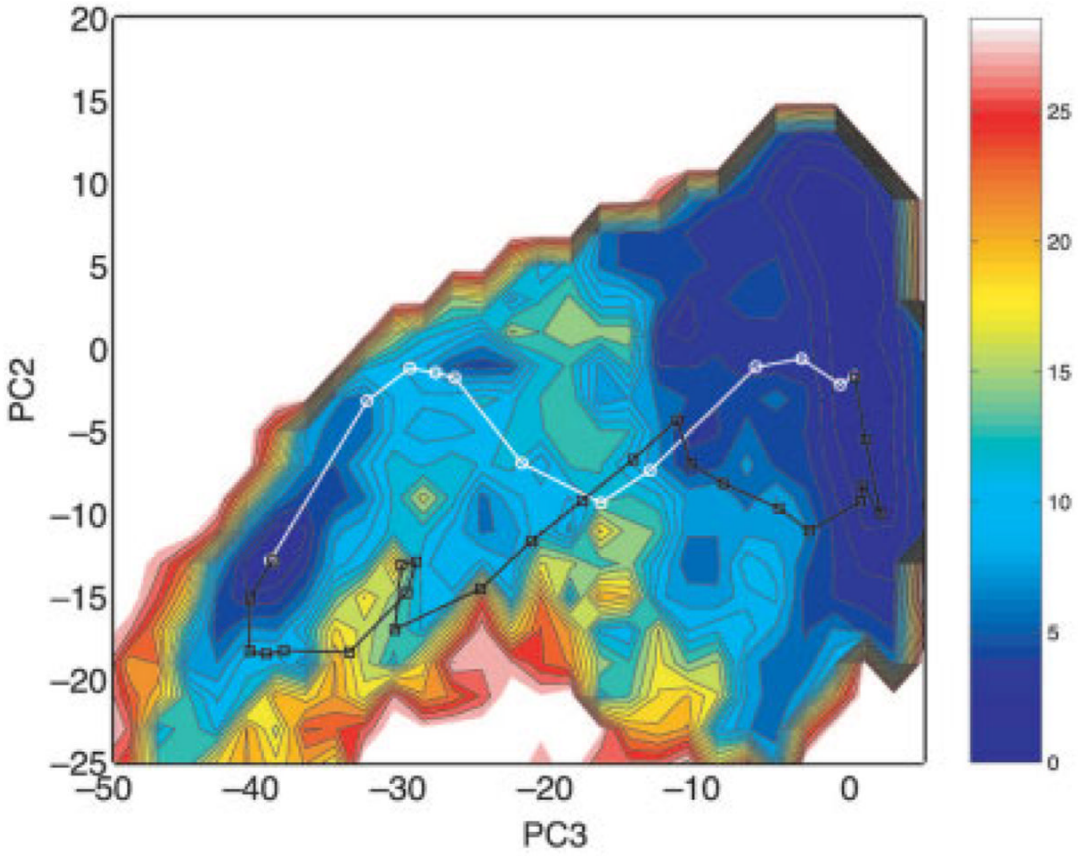

Figure 2.

The potential of mean force at 298 K of the capped C-terminal peptide from protein G with respect to the second (PC2) and third (PC3) principal components[107] (PC3 corresponds approximately to the end-to-end distance), using the OPLS-AA/AGBNP potential with additional dielectric screening of charged side chains. The PMF is calculated from T-WHAM analysis[107] of the ensemble of structures generated by all RXMD replicas with temperatures from 270 to 700 K. The energy is in units of kcal/mol. The low free energy region to the left correspond to α-helical conformations. The wider free energy basin to the right correspond to β-hairpin conformations. The black and white paths correspond to the upper and lower α-helical to β-hairpin conversion mechanisms shown in Figure 3, respectively. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

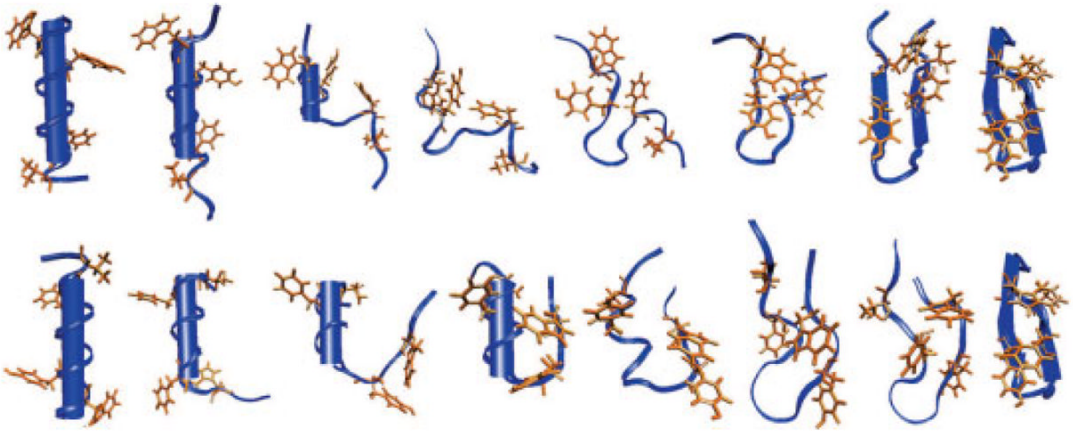

Figure 3.

Two possible pathways for the interconversion of an α-helix into a β-hairpin. Backbone trace is shown in blue, and the hydrophobic core residues (W43, Y45, F52, and V54) side chains are shown in gold. The upper path corresponds to unraveling of the helix at both ends and formation of a β-turn from a residual turn of α-helix. The lower path corresponds to unraveling of one end of the helix, which loops back.[110]

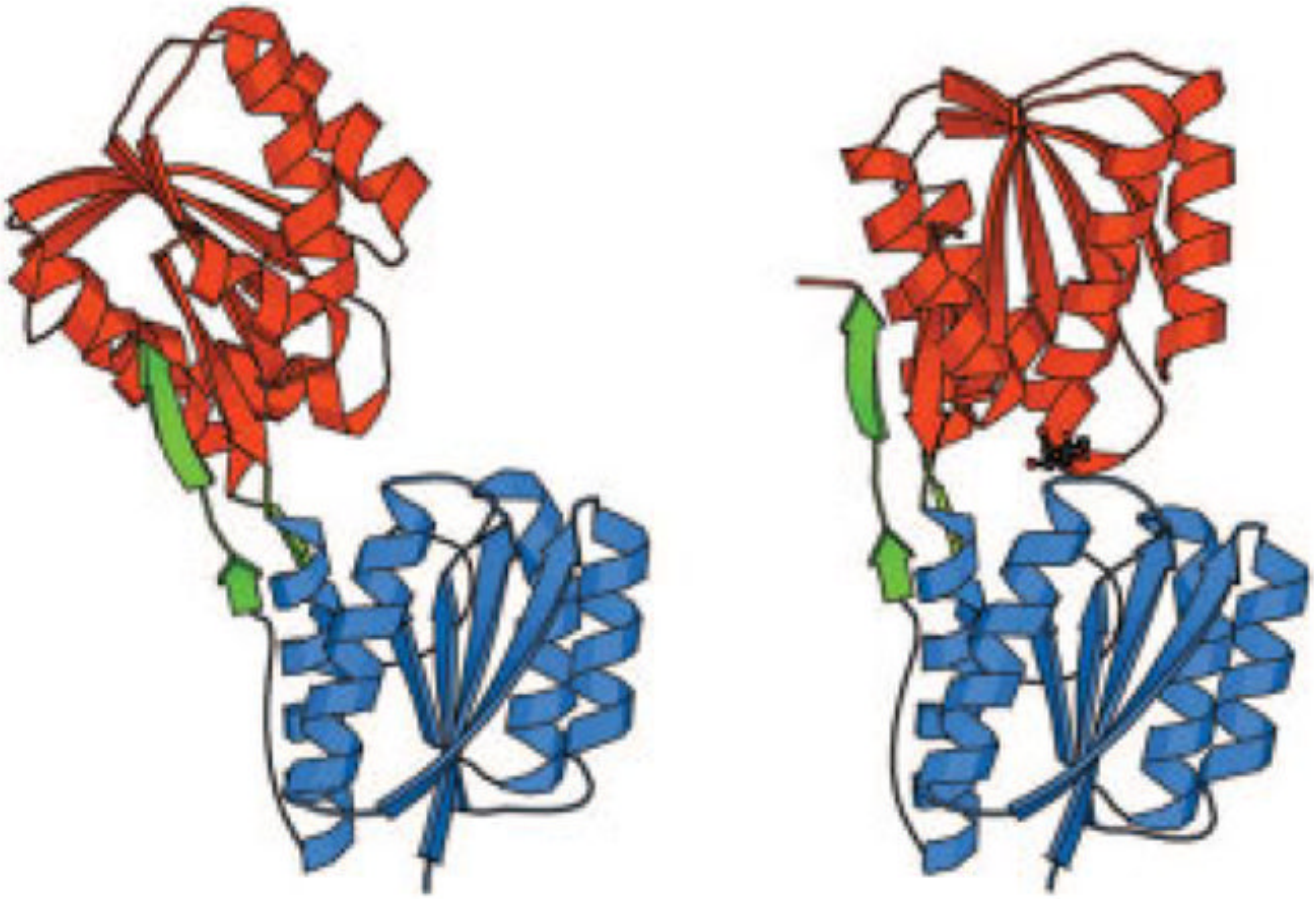

Figure 4.

The open ribose-free crystal structure of RBP (PDB id 1urp, left) and the closed ribose-bound crystal structure of RBP (PDB id 2dri, right). Ribose (D-ribopyranose form) is shown in ball-and-stick representation. The N-terminal domain of RBP is shown in blue and the C-terminal domain is shown in red. Three strands connecting the N-terminal domain to the C-terminal domain form the hinge region shown in green. The conformational change from the open (a) to the closed (b) conformation consists of a rotation of the C-terminal domain (red) toward the N-terminal domain (blue) around the axis perpendicular to the page centered on the hinge region (bending), followed by a rotation towards the viewer of the C-terminal domain around the axis longitudinal to the hinge region and parallel to the page (twisting). This figure has been generated using the program MOLSCRIPT.[147]

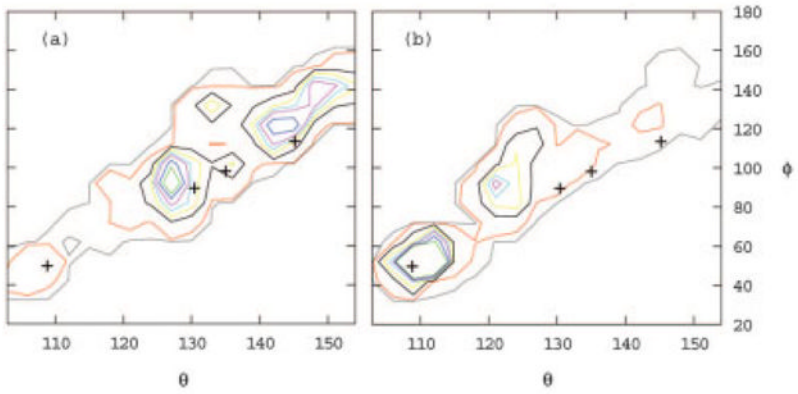

Figure 5.

Calculated population distribution of (a) ribose-free and (b) ribose-bound RBP as a function of the interdomain bending angle θ and twisting angle φ (see Fig. 4). Contours are drawn at 0.08 (green), 0.06 (blue), 0.04 (magenta), 0.03 (sky blue), 0.02 (yellow), 0.01 (brown), 0.001 (red), and 0.0001(gray) relative populations. The crosses indicate the (θ, φ) coordinates of the crystal structures of, with increasing angle θ, the closed ribose-bound (2dri) and open ribose-free (1urp) conformations of RBP, and open ribose-free conformations of a single-point mutant of RBP (1ba2). The angle θ is defined as the angle formed between the centers of mass of the C- and N-terminal domains and the center of mass of the hinge region. The angle φ is defined as the dihedral angle formed by the center of mass of the N-terminal domain, the center of mass of the residues on the N-terminal domain side of each of the three hinge strands, the center of mass of the corresponding residues on the C-terminal domain side of the three hinge strands, and the center of mass of the C-terminal domain. Angles are expressed in degrees. Highly populated conformations generally correspond to experimental crystal conformations. In agreement with experiments, the predicted population of the closed conformation ($\theta \approx 110°$, $\varphi \approx 50°$) in the presence of ribose (b) is larger than in the absence of ribose (a). Similarly, the population of open conformations ($\theta > 115°$) is larger in the absence of ribose. The population peak at ($\theta \approx 122°$, $\varphi \approx 95°$) in the presence of ribose (b) corresponds to a partially open conformation not yet observed experimentally postulated to play a role in the mechanism of ribose transport.

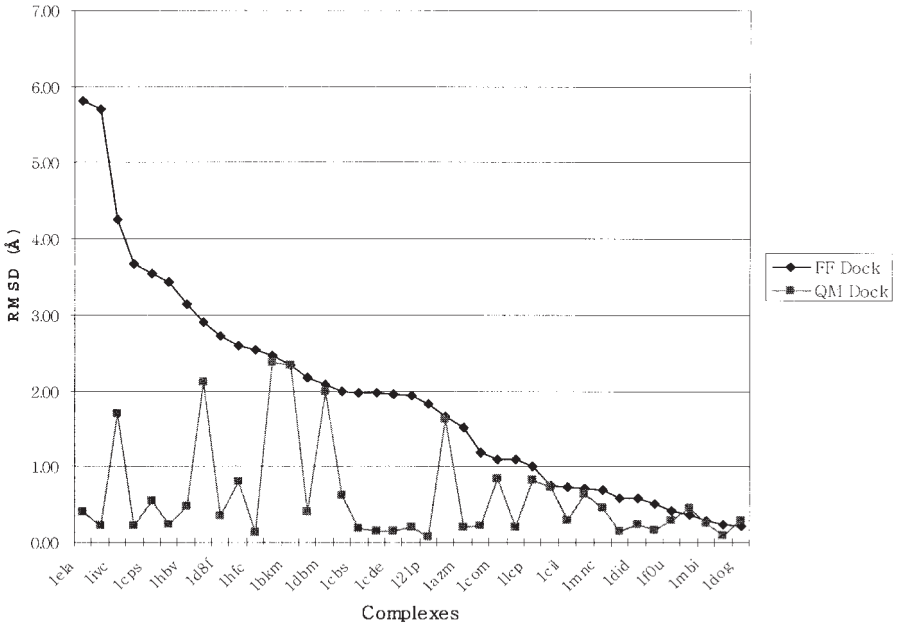

Figure 6.
RMSD from the native of the docked ligands using FF Dock and QM Dock.

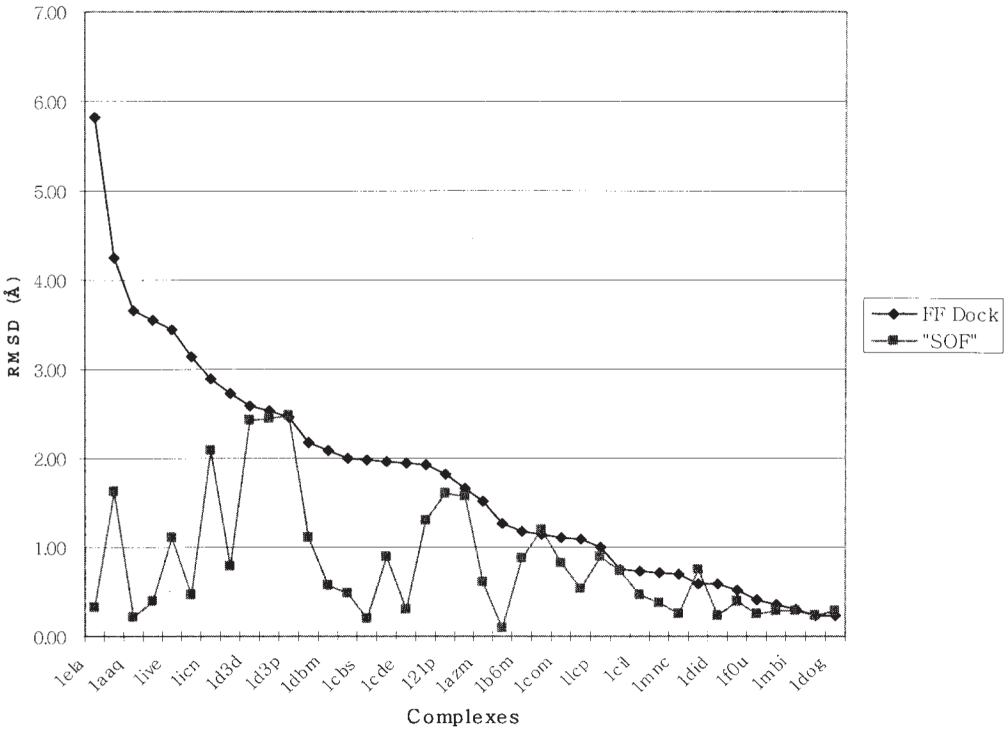

Figure 7.
RMSD from the native of the docked ligands using FF Dock and QM Dock with the “SOF” algorithm.

Table 1

RMS Energy Deviations (kcal/mol) from LMP2/cc-pVTZ(-f)//HF/6-31G** for Peptides.

Peptide	Original OPLS-AA[a]	PFF	OPLS-AA/L[a]	MMFF94[a]
Tetrapeptide				
Alanine	1.47	0.81	0.56	
Dipeptides				
Alanine	0.43	0.20	0.27	
Serine	0.47	0.16	0.44/0.34	0.97
Phenylalanine	0.35	0.05	0.15	0.21
Cysteine	1.91	0.31	0.35	1.21
Asparagine	1.30	0.19	0.16	2.25
Glutamine	0.98	0.69	0.96	1.00
Histidine	0.79	0.90	0.96/0.72	1.60
Leucine	0.37	0.57	0.34/0.38	1.27
Isoleucine	0.88	1.04	0.38	0.66
Valine	0.39	0.14	0.08/0.16	1.01
Methionine	1.00	0.59	0.59	1.05
Proline	2.25	0.76	1.54	
Tryptophan	0.56	0.63	0.50	0.83
Threonine	0.77	0.61	0.87	1.15
Tyrosine	0.35	0.25	0.39	0.28
Average[b]	0.81	0.55	0.47	1.04

[a] From ref. 8.

[b] Proline not included.

Table 2

RMS Energy Deviations (kcal/mol) from LMP2/cc-pVTZ(-f)//HF/6-31G** for Charged Dipeptides.

Peptide	Original OPLS-AA[a]	PFF	OPLS-AA/L
Aspartic acid	4.15	0.41	0.16/1.95
Glutamic acid	2.24	1.41	1.53
Lysine	1.09	0.32	0.88
Protonated His	2.05	0.57	0.97
Arginine	1.50	0.72	1.15
Average	2.20	0.69	0.94/1.29

[a]From ref. 8.

Table 3

The Number of Molecules in the Training Set (#Mol) and the Average Unsigned Errors (AVE in kcal/mol) Between the Predicted and the Experimental Hydration Free Energies for PBF with the OPLS-AA (OPLS_2003) Fixed Charge Force Field, the Polarizable Force Field (PFF) and Quantum Chemical Description (QM) of the Solute.

	PBF/OPLS-AA		PBF/PFF[a]		PBF/QM[b]	
	#Mol	AVE	#Mol	AVE	#Mol	AVE
Neutral	130	0.24	126	0.26	120	0.36
Ionic	21	0.65	21	0.83		
Total	151	0.30	147	0.34	120	0.36

[a] From ref. 13.

[b] From ref. 77.

Table 4

Comparison of Experimental and Predicted α and β Secondary Structure Content of a Series of Peptides.

		Sec. struct. content	
Name	**Sequence**	**Experimental**	**Predicted**
G-peptide	GEWTYDDATKTFTVTE	42% β[a]	40% β
C-peptide	KETAAAKFERQHM	29% α[b]	30% α
CheY2	EDGVDALNKLQAGGY	2% α[c]	2% α
CheY2-mu	EDAVEALRKLQAGGY	39% α[c]	45% α
SH3Lo	DYQEKSPREVAMKKG	2% α[c]	6% α

[a] Ref. 101.

[b] Ref. 146.

[c] Ref. 109.

Table 5

Average RMS Deviations for Flexible Ligand Docking on 279 PDB Complexes.

Number of rot. bonds	Number of cases	Average RMSD top ranked pose (Å)	Average CPU time (min.)
0–3	51	1.23	0.1
4–6	92	1.39	0.4
7–10	48	1.47	0.9
0–8	164	1.34	0.3
0–10	191	1.37	0.4
0–20	263	1.78	1.1

Times are CPU-minutes on a 1.6-GHz AMD Opteron 242 processor.

Table 6

Comparison of RMS Deviations (Å) for Flexible Docking by Glide SP and GOLD.

	≤10 rotbonds (72 cases)		≤20 rotbonds (86 cases)		All ligands (93 cases)	
	avg. RMSD	max. RMSD	avg. RMSD	max. RMSD	avg. RMSD	max. RMSD
Glide	1.41	6.1	1.77	11.8	2.13	11.9
GOLD	2.56	14.0	2.92	14.0	3.06	14.0

Table 7

Comparison of RMS Deviations (Å) for Flexible Docking by Glide SP and FlexX.

	≤10 rotbonds (133 cases)		≤20 rotbonds (174 cases)		All ligands (187 cases)	
	avg. RMSD	max. RMSD	avg. RMSD	max. RMSD	avg. RMSD	max. RMSD
Glide	1.31	6.1	1.69	11.8	1.96	11.9
FlexX	2.99	12.6	3.47	13.4	3.73	15.5

Table 8

Fraction of Known Actives Ranked by SP and XP Glide in the Top 2, 5, and 10% of a 1000 Ligand Database Seeded with the Active Compounds.

			% Actives in top 2% of database		% Actives in top 5% of database		% Actives in top 10% of database	
Screen	**Site**	**Number actives**	**SP**	**XP**	**SP**	**XP**	**SP**	**XP**
Thymidine kinase	1kim	7	42.9	85.7	71.4	85.7	85.7	85.7
CDK-2 kinase	1aq1	10	20.0	60.0	30.0	60.0	50.0	80.0
p38 MAP kinase	1bl7	50	2.0	30.0	14.0	56.0	30.0	68.0
p38 MAP Kinase	1kv2	10	30.0	100.0	40.0	100.0	60.0	100.0
Estrogen receptor	3ert	10	60.0	70.0	60.0	70.0	70.0	80.0
Thrombin	1ett	16	56.2	100.0	87.5	100.0	100.0	100.0
HIV protease	1hpx	16	53.3	80.0	73.3	86.7	86.7	86.7
Cox-2	1cx2	20	52.6	42.1	63.2	78.9	68.4	84.2
HIV rev. transcriptase	1rt1	33	15.2	36.4	30.3	60.6	54.5	69.7
Acetylcholinesterase	1e66	27	7.4	29.6	11.1	51.9	11.1	70.4
Ferredoxin	1fxa	20	35.0	65.0	60.0	80.0	80.0	80.0

Table 9

Comparison of Calculated and Experimental Activation Free Energies for Protein Active Site Chemistry.

Protein	Substrate	Reaction	Calculated	Experimental
MMO	dioxygen	H_{red}–P	22.1	18.2–20.0
MMO	dioxygen	P–Q	17.9	15.7–16.6
MMO	Methane	H atom abstraction	18.6	15.4
MMO	Acetonitrile	H atom abstraction	13.5	13.9
MMO	Nitromethane	H atom abstraction	18.1	16.2
P450cam	Camphor	H atom abstraction	8.2	fast
Hr	Dioxygen	$Hr_{red} - Hr_{ox}$	−5.2	−7.3
TIM	DHAP	H atom abstraction	14.1	13.0
P99-betalactamase	Cephalothin	Hydrolysis; formation of tetrahedral intermediate	14.3	14.3

All calculated and experimental quantities are activation free energies required to reach the transition state for the reaction, with the exception of Hr, where the free energy of dioxygen binding is reported. Calculated results include zero-point energies, but incorporate quantum tunneling corrections only for the hydrogen atom abstraction reaction of methane catalyzed by MMO.