אור ותמ
ים ים

LUX ET VERITAS

# Implementation of an Efficient Method of Calculating Generalized Born Energies in Monte Carlo Simulations using the GB/SA Hydration Model

A Prospectus
Presented to the Faculty of the Department of Chemistry at
Yale University
in Candidacy for the Degree of
Master of Science

by
John Philip Terhorst

Thesis Advisor: William L. Jorgensen

May, 2008

# Contents

## List of Figures

## List of Tables

# 1 Specific Aims

In computational simulations of biomolecules, description of the solvent with explicit solvent molecules yields slow convergence rates and long simulation times as a result of the many degrees of freedom in accounting for all the solvent molecules in the system. Explicitly treated water molecules that are equilibrated in the binding site of the protein can also complicate the system setup and affect the outcome of the simulation. To circumvent these problems, the generalized Born (GB) theory of implicit solvation was formulated. The generalized Born / surface area (GB/SA) model has emerged as a fast and accurate alternative, and is now widely used to model the solvation of such complex systems. Although efficient in molecular dynamics simulations, the GB/SA solvation model does not integrate well into Monte Carlo simulations, because of the pairwise description of the Born radii in the equations for the generalized Born energy. Whereas molecular dynamics simulations in GB/SA exhibit a 4–5-fold increase in computation time relative to gas phase, we find that Monte Carlo simulations in GB/SA exhibit a 15–20-fold increase. Thus, a more efficient implementation of the GB/SA solvation model within the Monte Carlo program *MCPRO* would be of great interest and impact. The following aims are therefore proposed:

1. *Evaluation of the current GB/SA algorithm as implemented in* MCPRO *via simulations of drug-like molecules.* Throughout the course of this research, we will use as a test system a compound from the development of an NNRTI series for HIV-RT. We will supplement previous calculations of free energies of binding for this system with our own gas phase and GB/SA simulations to obtain relative free energies of hydration in both TIP4P (explicit) and GB/SA (implicit) water. In doing so, we will be able to evaluate the precision and accuracy of the current implementation of GB/SA in the context of a real-world application in lead optimization.

2. *Modification of* MCPRO *to include the GB/SA algorithm within the Monte Carlo free-energy perturbation (MC/FEP) manifold.* In order to predict protein–ligand binding affinities using implicit solvation, our GB/SA model must first be integrated into the MC/FEP manifold. Then, we will launch fully rigorous MC/FEP simulations of our test system with GB/SA solvation in order to evaluate the free energies of binding in GB/SA and compare them to those from simulations in TIP4P (explicit) water.

3. *Enhancement of the current GB/SA algorithm as implemented in* MCPRO *such that calculations are made more efficient with little or no loss of accuracy.* This includes an approximated generalized Born potential in which the generalized Born energy between each pair of atoms is re-calculated only after moves in which the Born radius of either atom changes by more than a specified amount. The completion of Aims 2 and 3 should allow us to predict protein–ligand binding affinities via MC/FEP in GB/SA more rapidly. Previous work in our lab has studied extensively the protein–ligand binding affinities for systems such as HIV-RT using MC/FEP in TIP4P (explicit) water, and it would be of great interest to be able to obtain results of similar precision and accuracy with implicit solvent.

# 2 Background and Significance

## 2.1 The GB/SA Continuum Model of Solvation

The generalized Born / surface area (GB/SA) solvation model of Still and co-workers was developed in the early 1990s as a fast, semi-analytical treatment of solvation for molecular mechanics and molecular dynamics simulations.[1] Motivated by the observation that modeling large systems with explicitly treated solvent yielded slow convergence rates and required long simulation times, the GB/SA solvent model was designed to simulate solvation of a solute by abstraction of the discrete solvent molecules into a statistical dielectric continuum.[2] In doing so, computational cost could be significantly reduced while maintaining a high degree of accuracy and precision.

In the GB/SA model, the free energy of solvation, $G_{\text{sol}}$, is given as the sum of three terms: $G_{\text{cav}}$, $G_{\text{vdW}}$, and $G_{\text{pol}}$, eq 1.

$$G_{\text{sol}} = G_{\text{cav}} + G_{\text{vdW}} + G_{\text{pol}} \tag{1}$$

$G_{\text{cav}}$ is a solvent–solvent cavity term describing the free energy required to form a cavity within the solvent to accommodate the shape and volume of the solute. $G_{\text{vdW}}$ describes the solute–solvent van der Waals interactions, and $G_{\text{pol}}$ describes solute–solvent electrostatic polarization. The $G_{\text{cav}}$ and $G_{\text{vdW}}$ terms can be grouped into a nonpolar term, $G_{\text{np}}$, which is linearly related to the solvent-accessible surface area[3] for atom type $i$, $SA_i$, eq 2. The $\sigma_i$ in eq 2 is an empirically determined solvation parameter for cavity formation.

$$G_{\text{cav}} + G_{\text{vdW}} = G_{\text{np}} = \sum_i \sigma_i SA_i \tag{2}$$

In a calculation that includes GB/SA, the free energy of solvation is added to the potential energy $U$ for the protein–ligand complex *in vacuo*. The majority of the work that has gone into the development of continuum solvation models has focused on the the description of the final term, i.e., the electrostatic contribution.[2,4,5,6] In the original formulation of GB/SA,[1] $G_{\mathrm{pol}}$ was described as shown in eq 3, which is an elaboration of the original Born equation, eq 4. This elaboration allows for the electrostatic polarization term to be calculated for non-spherical solutes, relating $G_{\mathrm{pol}}$ to a number of physical parameters. These parameters describe the dielectric constant, $\epsilon$, of the solvent being simulated, the charge associated with each atom, $q_i$, and the Born radius associated with each atom, $\alpha_i$. The Born radius $\alpha_i$ is given as the spherically averaged distance from the center of atom $i$ to its dielectric boundary. The parameter $r_{ij}$ describes the distance between two atoms $i$ and $j$, and $\alpha_{ij} = (\alpha_i \alpha_j)^{1/2}$.

$$G_{\mathrm{pol}} = -166.0 \left(1 - \frac{1}{\epsilon}\right) \sum_i \sum_j \frac{q_i q_j}{r_{ij}^2 + \alpha_{ij}^2 \exp(-r_{ij}^2/2\alpha_{ij}^2)^{1/2}} \tag{3}$$

$$G_{\mathrm{pol}} = -166.0 \left(1 - \frac{1}{\epsilon}\right) \frac{q^2}{\alpha} \tag{4}$$

For solutes that are anything but spherical, e.g., for irregularly shaped molecular solutes, the calculation of $G_{\mathrm{pol}}$ in eq 3 is not trivial, due to the pairwise nature of $\alpha_{ij}^2$. In this description, the Born radius of any given solute atom $i$ is dependent on the position of all the other atoms $j$ in the solute. Thus, for solutes of high molecular weight, e.g., a protein, calculation of $G_{\mathrm{pol}}$ requires a numerical finite-difference method, which, while yielding well-defined and accurate Born radii, becomes prohibitively time-consuming as the systems under investigation become more complex.

In 1997, Qiu and Still[7] revised eq 3, making the calculation of Born radii fully analytical, circumventing the need for a numerical approach to the electrostatic polarization term. In their revision, they showed that the contribution of each atom to $G_{\mathrm{pol}}$ could be approximated without the need to know its respective $\alpha_{ij}$. This was accomplished by the definition of a new term, $G'_{\text{pol},i}$, which includes empirical scaling parameters $P_1$–$P_5$, as shown in eq 5. The dielectric offset, $\phi$, specifies a gap between the edge of the van der Waals radius and the beginning of the dielectric continuum, making the continuum solvent behave more realistically, i.e., like that of an explicit solvent. $V_j$ is the volume occupied by atom $j$, and CCF is a close-contact function for nonbonded (i.e., $1, \geq 4$) interactions of nearby atoms $j$ with atom $i$.

$$G'_{\text{pol},i} = \frac{-166.0}{R_{\text{vdW},i} + \phi + P_1} + \sum^{\text{Stretch}} \frac{P_2 V_j}{r_{ij}^4} + \sum^{\text{Bend}} \frac{P_3 V_j}{r_{ij}^4} + \sum^{\text{NB}} \frac{P_4 V_j \text{CCF}}{r_{ij}^4} \tag{5}$$

$$\text{CCF} = 1.0 \text{ if } \left( \frac{r_{ij}}{R_{\text{vdW},i} + R_{\text{vdW},j}} \right)^2 > \frac{1}{P_5}\text{; otherwise,}$$

$$\text{CCF} = \left\{ 0.5 \left[ 1 - \cos \left\{ \left( \frac{r_{ij}}{R_{\text{vdW},i} + R_{\text{vdW},j}} \right)^2 P_5 \pi \right\} \right] \right\}^2$$

It should be noted that $G'_{\text{pol},i}$, while being independent of $\alpha_{ij}$, is also independent of both the dielectric $\epsilon$ and the charge distribution. Furthermore, in the simplified Born equation, eq 6, $\alpha_{ij}$ is dependent solely on $G'_{\text{pol},i}$, making it also independent of $\epsilon$ and the charge distribution.

$$\alpha_{ij} = \frac{-166.0}{G'_{\text{pol},i}} \tag{6}$$

The analytical approach to calculating the Born radii made the simulation of large biomolecular systems much more feasible. Yet, until 2004, the implementation of the GB/SA solvation model in our lab relied upon the OPLS united-atom force field.[8] In 2004, the GB/SA solvation model as described by Qiu and Still[7] was migrated[9] to the newer OPLS all-atom force field,[10] allowing for broader application and improved accuracy.

## 2.2 GB/SA Solvation in Protein Simulations

There is great appeal in employing implicit solvents in simulations of large biomolecular systems[11] because of the enormous speed increases they afford over explicit solvent models. In the last 10 years, GB/SA has become more widely used in the context of molecular dynamics (MD) simulations, including studies of RNA hairpin unfolding,[12] nucleic acid conformational dynamics,[13] and determination of free energy surfaces of $\beta$-hairpin and $\alpha$-helical peptides by replica-exchange molecular dynamics.[14] There is also precedent for the use of GB/SA in Metropolis[15] Monte Carlo (MC) simulations of biomolecular systems. For instance, flexible docking[16] and concerted rotation with angles (CRA)[17,18] MC algorithms take advantage of GB/SA solvation. Unfortunately, the number of reported studies using GB/SA in MC simulations of proteins pales in comparison to the number of reported studies using GB/SA in MD simulations of proteins, and few have exploited GB/SA in combination with rigorous calculations of binding affinities.[19,20,21,22]

Even with the enormous speed increases obtained by switching from an explicit to an implicit solvent model, simulations of large systems in implicit solvent can become similarly as computationally demanding as those of smaller systems in explicit solvent. This is because of the sheer number of solute atoms alone in a biomolecular simulation; whereas the computational demand and slow rate of convergence of modeling a smaller system in explicit solvent arises from the large number of solvent molecules, high computational demand of modeling a larger system in GB/SA arises from the large number of solute atoms and from eqs 3 or 5, upon inspection of which one finds that the energies between each solute atom and every other solute atom in the system must be recalculated after every move, even if the position(s) of one or both of the atoms in the pair did not change. For a system of greater than 1000 solute atoms, this is not a trivial task. Furthermore, a MC simulation requires significantly more moves than a MD simulation requires time steps; it is therefore not surprising that molecular dynamics simulations in GB/SA exhibit only a 4–5-fold increase in computation time relative to the gas phase,[23] whereas we find that Monte Carlo simulations in GB/SA exhibit a 15–20-fold increase. In this light, MC may quickly lose its appeal to some as a method for simulating large biomolecular systems. Nevertheless, the many powerful algorithms only available within a MC manifold make it an indispensable tool for studying these systems. Thus, a number of other modifications to the GB/SA model have been proposed.[24,25,26,27]

Rather than focusing on the form of $G_{\mathrm{pol}}$, these more sophisticated approaches have manipulated the procedure that is used to determine which energies need to be recalculated after ever move and which ones do not. For instance, the "frozen atom" approximation of Still and co-workers[24] that was published in 2002 approximates the effects of atoms distant from the site of interest by freezing their coordinates throughout the duration of the simulation. The strictly pairwise terms involving these "frozen" atoms therefore need to be calculated only once, and the derivatives thereof need not be calculated at all. In addition, this approximation defines a "buffer region" containing frozen atoms that are close enough to the site of interest to experience changes to their Born radii as a result of the moving atoms. Thus, different series of calculations are performed for each atom pair based upon whether the pair is defined as frozen–frozen, moving–moving, buffer–frozen, or buffer–moving. This method was benchmarked using trial systems of camphor bound to cytochrome P450 and bezamidine bound to $\beta$-trypsin. Depending upon the system and the cutoff distance that was chosen, speed increases of approximately 1.5–10.6-fold were achieved.

A different approach was taken by Michel and co-workers[25] in 2006, wherein they used the pairwise descreening approximation (PDA)[28] for their description of the Born radii, and structured their GB/SA implementation in such a way that the energy of any given atom pair is only recalculated if the Born radius of eithÄer atom changes by more than a specified threshold after a MC move. The optimized threshold was determined to be 0.005 Å, giving results within less than 0.1% error of a fully rigorous calculation over the course of 5000 moves. With this modification alone, a 3-fold speed increase was attained, and even with a threshold as low as 0.001 Å, a 2.5-fold speed increase was attained. In addition, the "simplified sampling potential" methodology proposed by Gelb[26] in 2003 was implemented alongside, and the combination of the two methods afforded a 7–8-fold speed increase over a fully rigorous MC GB/SA calculation.

As implemented in *MCPRO*,[29,30] the GB/SA model lacks any of the advanced sampling or cutoff procedures mentioned above. And, with more and more interest in biomolecular systems, the ability to simulate complex systems in a more computationally efficient manner becomes essential. Thus, employing a number of modifications that would lead to a more efficient implementation of GB/SA in *MCPRO* would be of great interest and impact.

## 2.3 Monte Carlo Free-Energy Perturbation (MC/FEP)

Free-energy perturbation (FEP) theory is a powerful method for calculating free-energy differences ($\Delta G$) of two different chemical states, A and B. Based on statistical mechanics and the Zwanzig equation,[31] eq 7, it is used in both the molecular dynamics and Monte Carlo manifolds.

$$\Delta G(\mathrm{A} \rightarrow \mathrm{B}) = G_{\mathrm{B}} - G_{\mathrm{A}} = -k_B T \ln \left\langle \exp\left(-\frac{E_{\mathrm{B}} - E_{\mathrm{A}}}{k_B T}\right)\right\rangle_{\mathrm{A}} \tag{7}$$

In eq 7, the difference in free energy between states A and B is calculated as an ensemble average of a simulation for state A (denoted by the brackets). Thus, a full Monte Carlo simulation is run for state A to determine $E_{\mathrm{A}}$, and each time a new configuration is accepted, the energy $E_{\mathrm{B}}$ is calculated for state B as well. For states that vary by a large amount, a scaling parameter $\lambda$ is used to divide the simulation into a series of small "windows," allowing the system to be perturbed smoothly over a number of simulations from $\lambda = 0$ at state A to $\lambda = 1$ at state B. Using small increments of $\Delta\lambda$ ensures smooth convergence, and simulating multiple windows in parallel shortens overall computer time.

Typically, a "double-wide" sampling method is used, where a single accepted configuration is perturbed to both $-\Delta\lambda$ and $+\Delta\lambda$, requiring fewer simulations to be run overall. Other sampling methods have been proposed, including overlap and double-ended variations.[32]

States A and B can differ in a number of different ways, making FEP such a versatile computational tool. For instance, bond lengths or non-bonded interatomic distances can be perturbed from state A to B, yielding a free-energy map along one or more sets of reaction coordinates, or atom types can be changed, simulating an "alchemical" mutation from one molecule to another *in silico*. The latter method is particularly useful in determining relative free energies of hydration or relative free energies of binding. Typically, when doing so, the appropriate thermodynamic cycle is set up, and two FEP calculations are performed independently. Examples are shown in Figure 1.

$$\begin{tikzcd}
\mathrm{A}_{\mathrm{gas}} \arrow[r, "\Delta G_{\mathrm{gas}}"] \arrow[d, "\Delta G_{\mathrm{hyd}}(\mathrm{A})"] & \mathrm{B}_{\mathrm{gas}} \arrow[d, "\Delta G_{\mathrm{hyd}}(\mathrm{B})"] \\
\mathrm{A}_{\mathrm{aq}} \arrow[r, "\Delta G_{\mathrm{aq}}"] & \mathrm{B}_{\mathrm{aq}}
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathrm{A} + \mathrm{P} \arrow[r, "\Delta G_{\mathrm{bind}}(\mathrm{A})"] \arrow[d, "\Delta G_{\mathrm{unbound}}"] & \mathrm{AP} \arrow[d, "\Delta G_{\mathrm{bound}}"] \\
\mathrm{B} + \mathrm{P} \arrow[r, "\Delta G_{\mathrm{bind}}(\mathrm{B})"] & \mathrm{BP}
\end{tikzcd}$$

Figure 1: Thermodynamic cycles for FEP-based determination of free energies of hydration (left) and binding (right).

The thermodynamic cycle on the left depicts a FEP simulation for calculating relative free energies of hydration, $\Delta\Delta G_{\mathrm{hyd}}$, between two molecules A and B. This quantity would be determined by perturbing molecule A into molecule B in both the gas phase and in aqueous solution. The thermodynamic cycle on the right depicts a FEP simulation for calculating relative free energies of binding, $\Delta\Delta G_{\mathrm{bind}}$, between two ligands A and B to a protein P. This quantity would be determined by perturbing ligand A into ligand B unbound in solution and bound to protein P in solution. Thus, in a FEP simulation with double-wide sampling and GB/SA solvation, three full GB energy calculations must be performed after each accepted move—one for the reference state and one for each of its two perturbed states at $-\Delta\lambda$ and $+\Delta\lambda$—making FEP with GB/SA incredibly time consuming. Therefore, the same limitations to general MC simulations also apply in MC/FEP with GB/SA.

## 3 Preliminary Results

### 3.1 Early Benchmarking Studies

In the search for non-nucleoside inhibitors of HIV-1 reverse transcriptase, lead optimization of structures that were generated from similarity searches have lead to active compounds consisting of a 1,3,4-oxadiazole core linking a phenyl and an anilinyl ring.[33,34] We chose a test system used in the development of an NNRTI series for HIV-RT, consisting of a parent ligand 5-benzyl-*N*-phenyl-1,3,4-oxadiazol-2-amine and its monochloro analogs bound to HIV-1 RT. The parent ligand is shown in Figure 2.

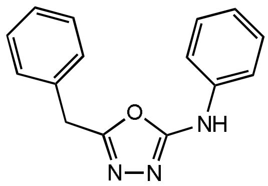

Figure 2: The parent ligand, 5-benzyl-*N*-phenyl-1,3,4-oxadiazol-2-amine, in our test system.

The protein scoop contained 2,728 atoms in 178 residues, and the parent ligand contained 30 atoms. Both the protein and the ligand have been set up previously with the OPLS-AA (all-atom) force field. Full standard Monte Carlo simulations were performed in the gas phase and in GB/SA as currently implemented in *MCPRO* version 2.1, each with 1

M configurations of equilibration followed by 1 M configurations of averaging. Cutoffs for solvent–solvent, solute–solvent, and intrasolute non-bonded interactions were set at 9.0 Å. Throughout the course of the simulation, the interactions of residues of similar charges were monitored. In our experience, when using GB/SA in simulations of large proteins, the Coulombic attractive and repulsive forces between charged residues are often miscalculated if the GB/SA parameterization for the Born radii is not correct. For example, two lysine residues might be brought together or even fused if the exaggeration is large enough. This is a problem that would need to be addressed before proceeding with any further enhancements of the GB/SA model. Our inspections revealed no such phenomena. Early benchmark times are given in Table 1, which shows that the simulation of the bound ligand in GB/SA takes nearly 17 times longer than in the gas phase. This leaves ample room for improvement.

Table 1: Early Monte Carlo benchmarking results for our system.

| Conditions | Time (h) | $\Delta t_{\mathrm{unbound} \to \mathrm{bound}}$ | $\Delta t_{\mathrm{gas} \to \mathrm{GB/SA}}$ |
|---|---|---|---|
| Unbound, gas phase | 0.43 | – | – |
| Bound, gas phase | 4.51 | $10.5x$ | – |
| Unbound, GB/SA | 0.62 | – | $1.44x$ |
| Bound, GB/SA | 76.0 | $123x$ | $16.8x$ |

We then turned our attention to the steric and electrostatic interactions between the ligand and the residues in the binding pocket. As a starting point, we examined the effects of applying perturbations to the ligand's heterocyclic core, given that cell-based assays have shown that modifying the heterocyclic core can lead to either decreased or enhanced activity, presumably because the modifications alter the spatial orientation of the core and/or its substituents in the binding pocket.

## 3.2 Fine-Tuning Intersubstituent Distances in Heterocycles

Current work in our lab has utilized free-energy perturbation (FEP) to guide the development of new analogues of structures like the one used in our test system (Figure 2). Results have indicated that by altering the $CH_2$ and NH linker groups, significant structural variation can be attained. Enhanced binding affinities *in silico* are observed in some of these analogs, ostensibly due to more favorable interactions between the ligand and the residues of the binding pocket. For instance, perturbation of the $CH_2$ (methylene) linker to an S (thio) linker widens the span of the two rings, while imparting a sharper angle to the thiophenyl (previously benzyl) ring and giving a more favorable $\Delta\Delta G_{\text{binding}}$. Thus, the intersubstituent distances within these ligands can be fine-tuned to allow for applications to more specific target structures.

We investigated the intersubstituent distances of 36 dimethyl-substituted heterocyclic cores, i.e., variants of the oxadiazole core in Figure 2. Structures were optimized using each of B3LYP/6-31G(d) density functional theory[35,36] in Gaussian03,[37] PDDG/PM3 semi-emperical molecular orbital calculations,[38,39] and the OPLS molecular mechanics force field[10] using CM1A*1.14 charges (OPLS/CM1A). The OPLS/CM1A and PDDG/PM3 optimizations were performed using *BOSS*[29] version 4.8. The the methyl–methyl distances were measured. The B3LYP/6-31G(d) intersubstituent distances for all 36 cores are given in Table 2. These calculations illustrate that intersubstituent distances can be manipulated to a fine degree by making simple changes to the core. Heterocycles with a "1,3"-dimethyl arrangement exhibited a much wider range of intersubstituent distances than those with a "1,2"-dimethyl arrangement, making the "1,3" motif more accessible to a wider range of conformational space. The OPLS/CM1A optimizations gave similar results, within 0.048 ± 0.037 Å of the DFT distances, Figure 3. The PDDG/PM3 distances were within 0.060 ± 0.030 Å of DFT, Figure 4. For both OPLS/CM1A and PDDG/PM3, better correlation with DFT was attained for the "1,3"-dimethyl heterocycles. Results are summarized in Table 3.

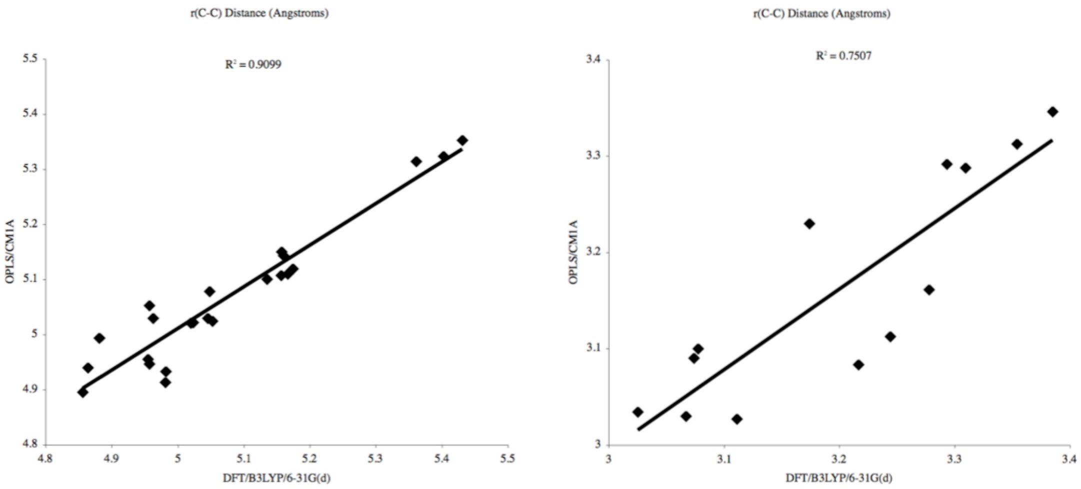

Figure 3: Comparison of B3LYP/6-31G(d) and OPLS/CM1A intersubstituent distances. Separate graphs are shown for the "1,3"-disubstituted analogs (left) and for the "1,2"-disubstitued analogs (right).

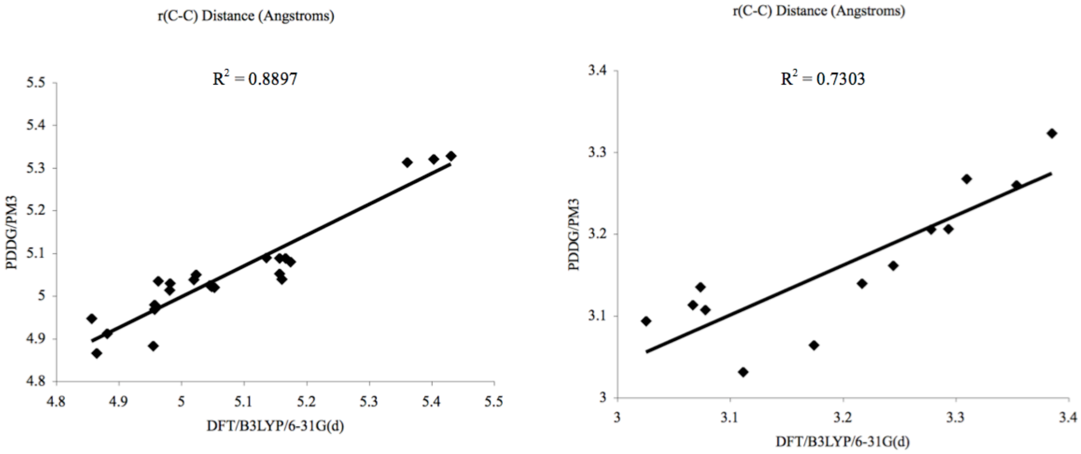

Figure 4: Comparison of B3LYP/6-31G(d) and PDDG/PM3 intersubstituent distances. Separate graphs are shown for the "1,3"-disubstituted analogs (left) and for the "1,2"-disubstitued analogs (right).

Table 2: B3LYP/6-31G(d) intersubstituent distances in heterocyclic cores, in Å.

| Core | Distance | Core | Distance |
|---|---|---|---|
| 2,5-dimethylthiophene | 5.431 | 2,4-dimethylthiazole | 4.956 |
| 2,5-dimethylthiazole | 5.403 | 2,5-dimethyloxazole | 4.954 |
| 2,5-dimethyl-1,3,4-thiadiazole | 5.361 | 2,5-dimethyl-1,3,4-oxadiazole | 4.880 |
| 2,4-dimethylpyrrole | 5.174 | 2,5-dimethylfuran | 4.864 |
| 2,4-dimethylfuran | 5.166 | 1,3-dimethylpyrazole | 4.856 |
| 2,4-dimethylthiophene | 5.160 | 4,5-dimethyloxazole | 3.385 |
| 3,5-dimethylisothiazole | 5.157 | 4,5-dimethylisoxazole | 3.353 |
| 3,5-dimethylpyrazole | 5.156 | 2,3-dimethylfuran | 3.310 |
| 3,5-dimethylisoxazole | 5.135 | 2,3-dimethylpyrrole | 3.293 |
| 1,4-dimethylpyrazole | 5.052 | 3,4-dimethylisoxazole | 3.278 |
| 1,4-dimethylimidazole | 5.048 | 3,4-dimethylfuran | 3.244 |
| 1,4-dimethyl-1,2,3-triazole | 5.045 | 3,4-dimethylpyrrole | 3.217 |
| 2,5-dimethylimidazole | 5.023 | 2,3-dimethylthiophene | 3.174 |
| 2,5-dimethylpyrrole | 5.019 | 3,4-dimethylthiophene | 3.111 |
| 2,4-dimethyloxazole | 4.982 | 1,5-dimethylimidazole | 3.077 |
| 2,4-dimethylimidazole | 4.981 | 1,2-dimethylimidazole | 3.074 |
| 3,5-dimethyl-1,2,4-oxadiazole | 4.963 | 1,5-dimethylpyrazole | 3.067 |
| 3,3-dimethyl-1,2,4-thiadiazole | 4.957 | 1,2-dimethylpyrrole | 3.025 |

Table 3: Intersubstituent Distances Deviation from DFT.

| Analog Subset | PDDG/PM3 | OPLS/CM1A |
|---|---|---|
| "1,3"-dimethyl | 0.054 ± 0.033 | 0.044 ± 0.032 |
| Correlation ($R^2$) | 0.8897 | 0.9909 |
| "1,2"-dimethyl | 0.070 ± 0.022 | 0.055 ± 0.046 |
| Correlation ($R^2$) | 0.7303 | 0.7507 |
| Entire Set | 0.060 ± 0.030 | 0.048 ± 0.037 |

In the case of HIV-RT, molecules containing the cores with the longest intersubstituent distances, e.g., 2,5-dimethylthiophene, thiazole, and thiadiazole (whose methyl–methyl distances are greater than 5.3 Å), have tested inactive in cell-based assays. The 1,3,4-oxadiazole core in the active derivatives affords a much smaller methyl–methyl distance of 0.4880 Å. As mentioned above, the identity of the linkers also affects the spatial orientation of the rings. Thus, for comparison, a we determined that a 2-amino-5-methyl-1,3,4-oxadiazole core exhibits a methyl–methyl distance of 4.782 Å by DFT, and that a 5-amino-2-mercapto-1,3,4-oxadiazole core exhibits a methyl–methyl distance of 5.003 Å by DFT.

We have also explored the 36 heterocyclic cores in Table 2 in much greater detail, in an auxiliary project aimed at refining the stretch–bend parameters of the OPLS force field to give better agreement with DFT in the overall geometries. Out of the nearly 900 angles examined, approximately 4% of them deviated from DFT by 2.433–4.000 degrees, and approximately 2% of them deviated by more than 4 degrees, with some deviations as large as 6.410 Å. Our goal in this respect is therefore to correct these discrepancies to give deviations of less 2 degrees.

## 4 Research Design and Methods

The methods for this project have been designed in such a way that useful results can be obtained before restructuring significantly the MC and GB/SA code in *MCPRO*.

### 4.1 Aim 1: Evaluation of the current GB/SA algorithm as implemented in *MCPRO* via simulations of drug-like molecules.

*Rationale:* Several preliminary simulations will be carried out in order to validate our current GB/SA method in the calculation of free energies of hydration for drug-like molecules. A MC/FEP chlorine "scan" has previously been performed on our test system, i.e., the parent compound in Figure 2 and each of its monochlorinated analogs, in TIP4P (explicit) water both unbound and bound to HIV-RT (Figure 1, right) in order to determine which sites optimize free energies of binding upon chlorination. This scan was performed by perturbing each aromatic hydrogen to a chlorine in both the unbound and bound structures, and then calculating the relative free energies of binding for each analog. These data can be used in part of the determination of free energies of hydration; in the general thermodynamic cycle for FEP-based determination of free energies of hydration for an unbound ligand (Figure 1, left), $\Delta G_{\text{aq}}$ have thus already been determined. Our work in Aim 1 will both take from and add to this previous work.

*Design and Methods for Aim 1:* The chlorine scan for our test system has been setup using the OPLS-AA (all-atom) force field and will be run for the unbound ligand in the gas phase. Simulations will be run for 1 M configurations of equilibration and 1 M configurations of averaging, using blocks of 100,000 configurations each, perturbing each of the chlorinated analogs (A) to the parent ligand (B). Thus, in conjunction with the chlorine scans run previously for the unbound parent ligand and its monochlorinated analogs in TIP4P (explicit) water, we will obtain free energies of hydration in TIP4P for each analog relative to the parent compound, as given in eqs 8 and 9.

$$\Delta G_{\text{gas}} + \Delta G_{\text{hyd}}^{\text{TIP4P}}(\text{B}) = \Delta G_{\text{hyd}}^{\text{TIP4P}}(\text{A}) + \Delta G_{\text{aq}} \tag{8}$$

$$\Delta G_{\text{aq}} - \Delta G_{\text{gas}} = \Delta G_{\text{hyd}}^{\text{TIP4P}}(\text{B}) - \Delta G_{\text{hyd}}^{\text{TIP4P}}(\text{A}) = \Delta\Delta G_{\text{hyd}}^{\text{TIP4P}} \tag{9}$$

Once these values are obtained, then simulations of the parent ligand and each of the chlorinated analogs will be run (unbound) using GB/SA solvation as currently implemented in *MCPRO*. Simulations will be run for 1 M configurations of equilibration and 1 M configurations of averaging, using blocks of 100,000 configurations each, to get free energies of hydration in GB/SA and relative free energies of hydration in GB/SA, according to eq 10.

$$\Delta G_{\text{hyd}}^{\text{GB/SA}}(\text{B}) - \Delta G_{\text{hyd}}^{\text{GB/SA}}(\text{A}) = \Delta\Delta G_{\text{hyd}}^{\text{GB/SA}} \tag{10}$$

At this point, relative free energies of hydration in GB/SA will be compared to relative free energies of hydration in TIP4P.

*Expectations for Aim 1:* We expect to see good correlation between relative free energies of hydration in GB/SA and relative free energies of hydration in TIP4P. Simulations of a single solute consisting of a few dozen atoms, as is the case for the present ligand, are computationally inexpensive in the gas phase and in GB/SA, and we expect the energies to converge rapidly.

## 4.2 Aim 2: Modification of *MCPRO* to include the GB/SA algorithm within the Monte Carlo free-energy perturbation (MC/FEP) manifold.

*Rationale:* In addition to relative free energies of hydration, relative free energies of binding are also valuable thermodynamic quantities to be able to calculate, especially in the context of drug design and lead optimization. In order to determine free energies of binding, two FEP simulations must be run, one perturbing the unbound protein or ligand in solution and one perturbing the protein–ligand complex in solution (Figure 1, right). However, at the present time, the GB/SA solvent model is not implemented in the MC/FEP protocol, and so free energies of binding in solution cannot currently be determined using implicit solvation. Therefore, we will revise *MCPRO* to allow for GB/SA to be used in MC/FEP simulations.

*Design and Methods for Aim 2:* A number of concerns are present. First, as mentioned in Section 2.3, fully rigorous MC/FEP calculations with the current GB/SA implementation will be incredibly time consuming. This, however, will provide us with a reference against which we will measure computer time once a faster GB/SA method is implemented. Second, and perhaps more importantly, there exists a very physical difference between FEP in explicit and implicit solvent. In a MC/FEP simulation with explicit solvent, energies $E_{\mathrm{A}}$ and $E_{\mathrm{B}}$ are computed and applied to eq 7 to obtain the change in free energy between states A and B. However, in a MC/FEP simulation with GB/SA, free energies $G_{\mathrm{A}}$ and $G_{\mathrm{B}}$ will be applied to eq 7. This will, ostensibly, lead to a very different description of the change in free energy with the two applications. After implementation, we will launch MC/FEP simulations for the chlorine scan for our test system (Figure 2) using the fully rigorous GB/SA algorithm. Ligand moves will be attempted 10% of the time, and the remaining moves will sample the protein side-chains.

*Expectations for Aim 2:* We expect implementation to be relatively straightforward, but we expect full MC/FEP simulations with GB/SA to take anywhere from several weeks to months to complete. Based on previous successful use of the GB/SA model in the context of MC/FEP simulations,[25] we expect the binding affinities obtained with GB/SA to compare relatively well with those obtained with TIP4P (explicit) water, despite the fact that they are being derived from quantities of physically different origins.

## 4.3 Aim 3: Enhancement of the current GB/SA algorithm as implemented in *MCPRO* such that calculations are made more efficient with little or no loss of accuracy.

*Rationale:* Monte Carlo simulations of large proteins in GB/SA solvation can be prohibitively time consuming as a result of the pairwise formulation in the GB/SA algorithm for calculating Born radii. However, the movement of an atom should have little impact on the Born radius of a distant atom. Therefore, we propose an implementation in which the generalized Born energy between each pair of atoms is recalculated only after moves in which the Born radius of either atom changes by more than a pre-specified threshold. This will reduce computer time significantly.

*Design and Methods for Aim 3.1:* Our implementation will be structured to mimic in spirit that of Michel and co-workers.[25] The technical details of our implementation are likely to be quite different than those of Michel and co-workers, since their implementation was based upon the pairwise descreening approximation (PDA)[28] for the Born radii, whereas the approximation for Born radii in *MCPRO* is based on that of Qiu and co-workers[7] and eqs 3 and 5. The threshold parameter will be optimized to best minimize both computer time and error in chlorine scans for our test system, relative to the fully rigorous Monte Carlo and MC/FEP simulations launched after completion of Aim 2, with the goal of assessing the impact of our approximations. Threshold values ranging from 0.001 Å to 0.1 Å will be examined. The threshold parameter will be implemented in the software as a variable that can be defined by the end-user for any custom system, with the optimized value set as the default.

*Expectations for Aim 3.1:* We expect that as long as the threshold parameter ensures that the impact of the approximation is small, the ensemble of states that is generated using the approximated potential will mimic closely the ensemble generated in a fully rigorous simulation.

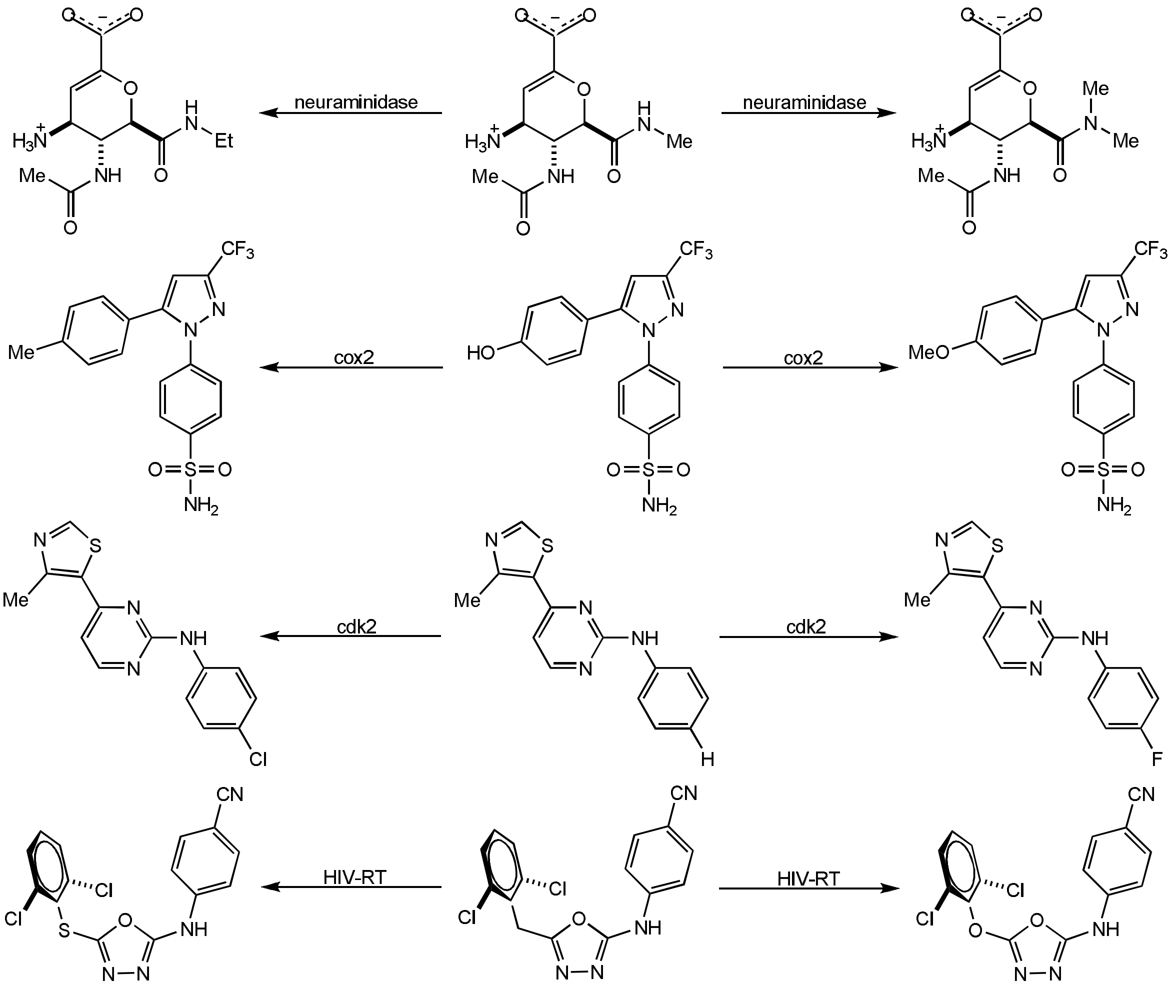

Figure 5: Systems that will be examined by standard MC and MC/FEP with the newly implemented GB/SA algorithm. Representative perturbations are shown.

*Design and Methods for Aim 3.2:* Systems for neuraminidase, cyclooxygenase 2 (cox2), cyclin-dependent kinase 2 (cdk2) and HIV-1 reverse transcriptase (HIV-RT) will be examined using MC and MC/FEP simulations with the newly implemented GB/SA model. Protein–ligand binding affinities for these systems have been studied using both explicit water[23,40] and GB/SA implicit solvation.[23,25] We hope to achieve similar results and to further assess the impact of our approximations. Representative perturbations for these systems are shown in Figure 5.

These systems have been chosen from a methodological standpoint because of the variety of physical properties they exhibit. For instance, the proposed perturbations cover a range of apolar–apolar, polar–polar, and polar–apolar, and are reminiscent of those performed in a typical MC/FEP simulation. Furthermore, the characteristics of the binding sites are quite different. For instance, neuraminidase has a polar, solvent-exposed binding site, whereas cox2 has a non-polar, buried binding site. This variety is essential for proper assessment of the new approximations of the new GB/SA algorithm. In practice, perturbations will be performed from the structure with the greater number of atoms to the structure with the fewer number of atoms. The systems will be setup with the OPLS-AA (all-atom) force field. Ligand moves will be attempted 10% of the time, and the remaining moves will sample the protein side-chains. We will use the optimized threshold as determined in Aim 3.1.

*Expectations for Aim 3.2:* The successful completion of both Aims 3.1 and 3.2 should allow us to predict protein–ligand binding affinities via MC/FEP in GB/SA more rapidly. We hope to obtain a 3–5-fold decrease in computer time by implementing the approach to calculating approximated Born radii. Energy deviations of less than 1.0 kcal/mol from simulations in explicit solvent are desired.

## 5 Summary

This proposal seeks to implement into *MCPRO* a method for calculating approximate Born radii in such a way that computer time for simulations of large biomolecular systems in GB/SA implicit solvent is reduced significantly. The ability to simulate biomolecular systems with efficient yet accurate approximations to solvent effects will enable thorough investigations of important biochemical processes to be more computationally feasible. Pairwise interactions involved in the calculation of Born radii, one of the more demanding aspects of a GB/SA calculation, will only be recalculated when one or both of the atoms move by more than a specified amount. Systems for neuraminidase, cox2, cdk2, and HIV-RT will be investigated from a diagnostic standpoint. Similar implementations have led to approximately 4-fold decreases in computer time, and we seek to achieve similar results.

## Notes

[1] Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

[2] Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187.

[3] Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754.

[4] Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616.

[5] Grant, J. A.; Sykes, M. J.; Kitchen, C. A.; Nicholls, A. *Phys. Chem. Chem. Phys.* **2007**, *9*, 4913.

[6] Jayaram, B.; Liu, Y.; Beveridge, L. *J. Chem. Phys.* **1998**, *109*, 1465.

[7] Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005.

[8] Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.

[9] Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. *J. Phys. Chem. B* **2004**, *108*, 16264.

[10] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

[11] Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.

[12] Sorin, E. J.; Engelhard, M. A.; Herschlag, D.; Pande, V. S. *J. Mol. Biol.* **2002**, *317*, 493.

[13] Sorin, E. J.; Rhee, Y, M.; Nakatani, B. J.; Pande, V. S. *Biophys. J.* **2003**, *85*, 790.

[14] Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Prot. Struct., Funct., Bioinformat.* **2004**, *56*, 310.

[15] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. *J. Chem. Phys.* **1953**, *21*, 1987.

[16] Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. *J. Comput. Chem.* **2004**, *24*, 1637.

[17] Ulmschneider, J. P.; Jorgensen, W. L. *J. Chem. Phys.* **2003**, *118*, 4261.

[18] Ulmschneider, J. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2004**, *126*, 1949.

[19]Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167.

[20]Henchman, R. H.; Kilburn, J. A.; Turner, D. L.; Essex, J. W. *J. Phys. Chem. B* **2004**, *108*, 17571.

[21]Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517.

[22]Zhnang, L. Y.; Gallicchio, E.; Friesner, R. A.; Levy, R. M. *J. Comput. Chem.* **2001**, *22*, 591.

[23]Michel, J.; Verdonk, M. L.; Essex, J. W. *J. Med. Chem.* **2006**, *49*, 7427.

[24]Guvench, O.; Weiser, J.; Shenkin, P.; Kolossváry, I.; Still, W. C. *J. Comput. Chem.* **2002**, *23*, 214.

[25]Michel, J.; Taylor, R. D.; Essex, J. W. *J. Chem. Theory Comput.* **2006**, *2*, 732.

[26]Gelb, L. D. *J. Chem. Phys.* **2003**, *118*, 7747.

[27]Felts, A. K.; Gallicchio, D.; Paris, K. A.; Friesner, R. A.; Levy, R. M. *J. Chem. Theory. Comput.* **2008**, *ASAP*

[28]Hawkins, C. J.; Cramer, D. *Chem. Phys. Lett.* **1995**, *246*, 122.

[29]Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689.

[30]Jorgensen, W. L.; Tirado-Rives, J. *MCPRO*, Version 2.05, Yale University, New Haven, CT, 2006.

[31]Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.

[32]Thomas, L. L.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2008**, *ASAP.*

[33]Barreiro, G.; Guimarães, C. R. W.; Tubert-Brohman, I.; Lyons, T. M.; Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Inf. Model.* **2007**, *47*, 2416

[34]Barreiro, G.; Kim, J. T.; Grimarães, C. R. W.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. *J. Med. Chem.*, **2007**, *50*, 5324

[35]Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864

[36]Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133

[37]Gaussian 03, Revision B.04, Frisch, M. J.; Pople, J. A. *et al.* Gaussian, Inc.. Pittsburgh PA, 2003

[38] Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601

[39] Tubert-Brohman, I.; Guimarães, C. R. W.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2005**, *1*, 817

[40] Price, M. L. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 9455.