JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

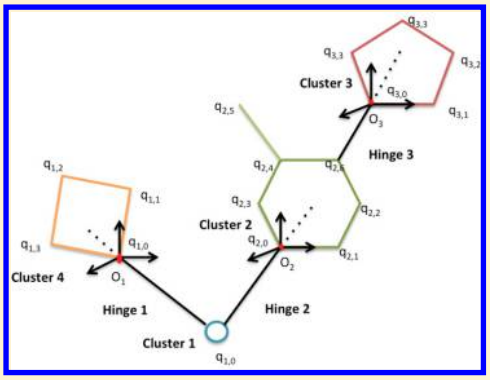
Energy Minimization on Manifolds for Docking Flexible Molecules

Hanieh Mirzaei,[†] Shahrooz Zarbafian,[‡] Elizabeth Villar,[§] Scott Mottarella,[∥] Dmitri Beglov,[⊥] Sandor Vajda,[⊥] Ioannis Ch. Paschalidis,[#] Pirooz Vakili,[▽] and Dima Kozakov*[,⊥]

[†]Division of Systems Engineering, [‡]Department of Mechanical Engineering, [§]Department of Chemistry, [∥]Program in Bioinformatics, [⊥]Department of Biomedical Engineering, [#]Division of Systems Engineering & Department of Electrical and Computer Engineering, [▽]Division of Systems Engineering & Department of Mechanical Engineering, Boston University, Boston, Massachusetts 02215, United States

Supporting Information

ABSTRACT: In this paper, we extend a recently introduced rigid body minimization algorithm, defined on manifolds, to the problem of minimizing the energy of interacting flexible molecules. The goal is to integrate moving the ligand in six dimensional rotational/translational space with internal rotations around rotatable bonds within the two molecules. We show that adding rotational degrees of freedom to the rigid moves of the ligand results in an overall optimization search space that is a manifold to which our manifold optimization approach can be extended. The effectiveness of the method is shown for three different docking problems of increasing complexity. First, we minimize the energy of fragment-size ligands with a single rotatable bond as part of a protein mapping method developed for the identification of binding hot spots. Second, we consider energy minimization for docking a flexible ligand to a rigid protein receptor, an approach frequently used in existing methods. In the third problem, we account for flexibility in both the ligand and the receptor. Results show that minimization using the manifold optimization algorithm is substantially more efficient than minimization using a traditional all-atom optimization algorithm while producing solutions of comparable quality. In addition to the specific problems considered, the method is general enough to be used in a large class of applications such as docking multidomain proteins with flexible hinges. The code is available under open source license (at http://cluspro.bu.edu/Code/Code_Rigtree.tar) and with minimal effort can be incorporated into any molecular modeling package.



INTRODUCTION

The challenge for predictive docking is to computationally obtain a model of the bound complex from the coordinates of component molecules.[1−7] Docking methods generally search for the minima of a scoring function that is based on molecular mechanics and may include empirical solvation terms. In this work, we focus on determining the binding pose and do not consider the problem of calculating the binding free energy. The scoring functions generally have large numbers of local minima, resulting in extremely rugged energy landscapes. Therefore, independently of the algorithm used for sampling the conformational space, virtually all docking algorithms include some type of *local continuous minimization* of the energy function in order to remove steric clashes and obtain more reliable energy values.[3] For example, both the popular Monte Carlo minimization[8−10] and the medium-range semidefinite underestimator[11−13] (SDU) methods proceed through series of local minima. Thus, developing more efficient local minimization algorithms has a direct impact on the overall efficiency of docking protocols. It is worth noting that local optimization of molecular energies are used in other areas of computational biology and chemistry and more efficient local optimization algorithms are beneficial in those domains as well.[14]

One possible formulation of local energy minimization as an optimization problem is to assume that all atoms can move freely and then rely on the minimization of the energy function to enforce the structural constraints due to covalent and noncovalent interactions. We refer to this approach as *all-atom (AA) optimization*. In this case, the conformational space is simply the $3n$-dimensional Euclidean space $\mathbb{R}^{3n}$ where n is the total number of atoms of the complex. Variations of all-atom optimization are also possible: for example, in some applications of protein small-molecular docking, one can assume the receptor protein is rigid while all atoms of the ligand can move freely.

In contrast to the all-atom optimization approach, one can explicitly take the partial or complete rigidity of the receptor and ligand into account when defining the conformational space. For example, if both receptor and ligand are assumed to be rigid, then the conformational space is the space of movements of the ligand relative to the receptor, namely, the 6-dimensional space of rigid body motion. This search space is no longer a Euclidean space but a *manifold*, and we refer to optimization on such search spaces as *manifold optimization (MO)*.

The advantage of an all-atom formulation is that the search space is always a Euclidean space with a well-known geometry for which various efficient and well-understood optimization

Received: February 20, 2014

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

algorithms are available. On the other hand, it has the drawback that the dimension of the search space is often large, leading to slow convergence of optimization algorithms. By formulating the energy minimization problem as a manifold optimization, we can often arrive at a search space with the smallest possible dimension. However, the geometry of the resulting manifold may present challenges for optimization.[15−17] For example, assume that we wish to develop a manifold optimization algorithm similar to the Euclidean steepest descent algorithm; in this case, to mimic a line search, once a steepest descent direction is determined, we need to identify a geodesic of the manifold in that direction, namely a curve that is the shortest distance between two points; it is important to note that not moving along a geodesic is tantamount to searching along an arbitrary curve in the Euclidean space rather than performing a line search, clearly not an advisible move. Therefore, manifold optimization requires computing or approximating the geodesics of the manifold and, for second order optimization methods, the Hessian of the energy function. These computations are not easily implementable in all cases.[15]

We have recently introduced a novel approach to rigid body minimization on manifolds that addresses these issues.[18,19] The novel element of our approach is to select an alternative manifold and group of rigid body transformations that correspond to the direct product of the groups of rotations and translations. We have shown that the new formulation avoids the difficulties associated with optimization on the so-called Special Euclidean group, $SE(3)$, which is often used to represent rigid motions,[15] better matches the moves of the optimization with the nature of molecular interactions, and provides the opportunity to improve the performance of the optimization algorithm. Indeed, we have also shown that the resulting algorithm substantially outperforms the state-of-the-art local minimization algorithms used for rigid body minimization.[18]

In this paper, we extend our manifold optimization approach to flexible minimization by allowing for some ligand and receptor flexibility. We represent these flexibilities using the internal coordinates of the ligand and receptor[20−22] and combine the internal coordinate representation with our manifold representation of the rigid moves of the ligand with respect to the receptor. We show that the resulting overall search space is also a manifold for which geodesics can be efficiently computed and approximated, and thus, our manifold optimization method developed for rigid body minimization can be extended to problems of finding the relative spatial orientation of two molecules with internal degrees of freedom. We note that a manifold approach involving exponential map parametrization has been recently presented for the global structure optimization of single molecules consisting of rigid fragments linked by rotatable bonds,[23] thus without the free rotational/translational freedom of one molecule (the ligand) relative to the other (the receptor). Since our approach combines these external degrees of freedom with the internal ones, it can be considered as the generalization of both our method and the one described by Kusumaatmaja et al.[23]

Although our combined method is very general and applies to a broad range of problems such as the efficient minimization of two multidomain proteins that have flexible hinges linking domains that can be considered rigid, here we focus on applications related to docking small ligands to proteins. The method is applied to three problems of increasing complexity. First, we minimize fragment-size ligands with a single rotatable bond as part of a protein mapping method developed for the identification of binding hot spots. Second, we consider energy minimization for docking a flexible ligand to a rigid protein receptor, an approach frequently used in existing methods. Third, we account for flexibility in both the ligand and the receptor. Results show that energy minimization using our manifold optimization algorithm is substantially more efficient than energy minimization using a traditional all-atom optimization algorithm while producing comparable low energy solutions. Furthermore, we present a detailed comparison of the traditional all-atom and of the manifold-based optimization algorithms.

MATERIALS AND METHODS

To avoid cumbersome notation and to simplify the discussion, in what follows, we limit the internal flexibilities of the ligand to torsional moves along rotatable bonds. We explicitly identify the manifold of the conformational space in this context. This manifold can be viewed as combining internal coordinates of the ligand with that of rigid motion of the entire ligand with respect to a reference coordinate frame. Next, we specify our local optimization algorithm. This algorithm uses a local parametrization of the manifold involving exponential maps, similarly to the algorithm we have developed for rigid body minimization.[18] As emphasized below, our particular formulation of rigid motions of the ligand makes the computation of this exponential parametrization possible.

Ligand Flexibility and Representation. We assume that the ligand is composed of a set of rigid clusters, that is, arbitrary sets of atoms that may be grouped as rigid bodies. In the case of small ligands such clusters can be formed by chemical groups in which the atoms move very little relative to each other, for example, in phenyl or naphthalene groups (see Figure 1). However, in protein−protein docking, individual domains, connected by flexible hinges, may be frequently represented as rigid clusters, and hence, we will refer to the bonds connecting rigid clusters as *hinges*.[21,22]

In general, freely rotating and translating clusters will have six degrees of freedom; however, as a first approximation, it is reasonable to assume that bond lengths and bond angles do not change significantly upon binding and can be initially assumed fixed (see, e.g., refs 9 and 24). By fixing bond lengths and bond angles, the only internal flexibilities of the ligand are torsional moves around rotatable bonds. A bond for which only changes in the torsional angle are permitted is modeled as a one degree-of-freedom rotational hinge. Following refs 21, 22, and 25, we use a *torsion tree* to represent the rigid and rotatable parts of the ligand.

Tree Topology Model. We form a topology graph $G = (V,E)$ of the ligand such that each node of the graph corresponds to a rigid cluster of the molecule. Two nodes are connected by an edge if and only if there is a rotatable covalent bond between the corresponding rigid clusters in the molecule. We assume that the resulting graph does not have any cycles and is a connected graph; in other words, it is a tree. We select one particular node/cluster of the tree as the *root* cluster. Once the root cluster is selected, the parent of each node in the tree is uniquely and completely determined. For example, in Figure 1, cluster 1 is chosen as the root cluster; cluster 1 is the parent of clusters 4, 5, and 2, cluster 4 is the parent of cluster 6, and cluster 2 is the parent of cluster 3. Each hinge is defined between a pair of parent−child clusters and connects an atom in the parent cluster to an atom in the child cluster. As shown in Figure 2, we assign a coordinate frame, that is initially parallel to a fixed reference coordinate frame, to each hinge. The center of the coordinate frame is the end atom of the hinge in the

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

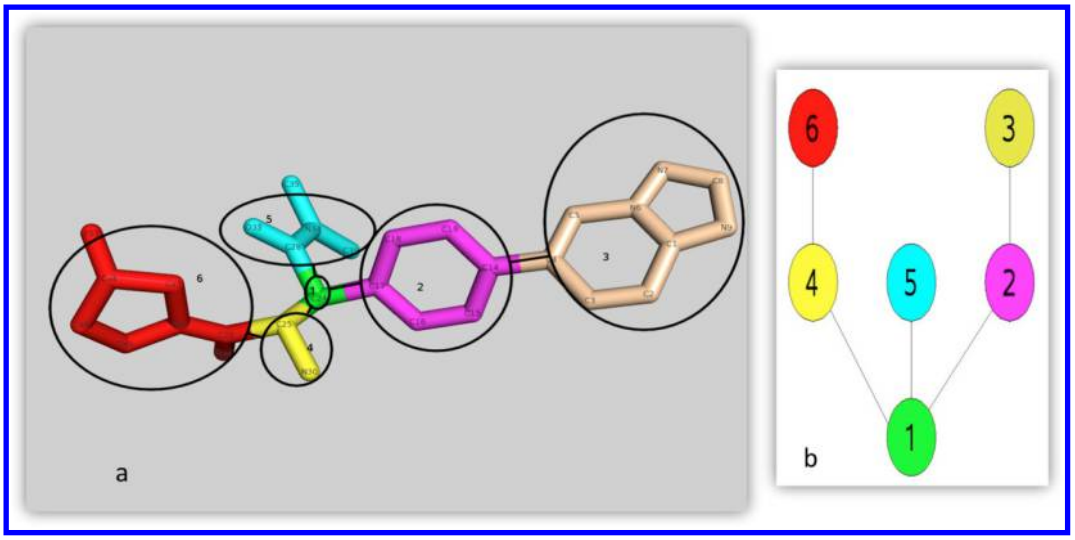


Figure 1. (a) Atoms of the ligand of 2FJP complex decomposed into rigid clusters; each cluster is colored differently and a number is assigned to it. (b) The tree structure of the corresponding rigid decomposition. Two rigid clusters are connected if there is a covalent bond between the corresponding clusters in the molecule.

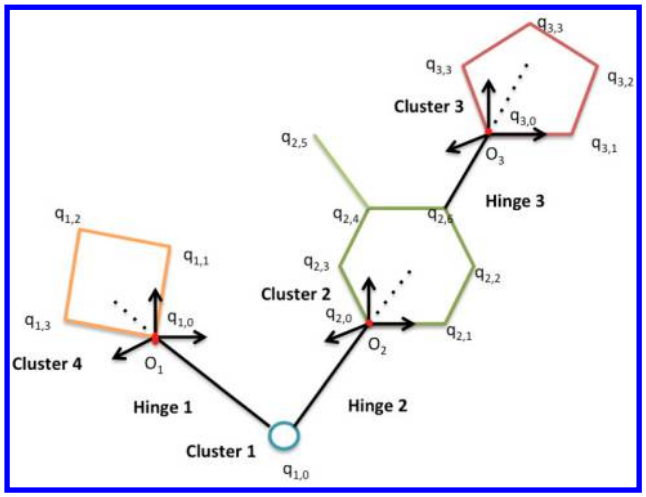


Figure 2. Illustration of the tree model of a molecule. The black lines represent hinges between parent−child clusters. For each hinge, we attach a coordinate frame to its child cluster. For example, hinge 3 is between atom $q_{2,6}$ from cluster 2 and atom $q_{3,0}$ from cluster 3. Cluster 2 is the parent of cluster 3, and there is a coordinate frame corresponding to hinge 3 attached to cluster 3. The center of the coordinate frame is atom $q_{3,0}$, which is the end atom of hinge 3.

corresponding child cluster. The motion of hinges is characterized by the motion of these frames with respect to the reference coordinate frame.

General Equation for the Ligand Displacement. We use the following notations:

- R denotes a rotation matrix, and t denotes a translation vector for the *whole ligand*.
- O denotes the center of mass of the ligand.
- We denote atoms in cluster A having m_A atoms as $(A,0),\cdots,(A,m_A-1)$ and their Cartesian coordinates with respect to the *fixed reference frame* as $\mathbf{q}_{A,0},\cdots,\mathbf{q}_{A,m_A-1}$.
- Let hinge k, denoted by h_k, be the hinge between parent cluster A and child cluster B. Assume hinge k is between atoms $\mathbf{q}_{A,ak} \in A$ and $\mathbf{q}_{B,0} \in B$.
  - $O_k = \mathbf{q}_{B,0}$ is the center of coordinate frame corresponding to hinge k.
  - $\mathbf{u}_k$ is the unit vector in the direction $\mathbf{q}_{B,0} - \mathbf{q}_{A,ak}$.
  - θ_k is the torsion angle around hinge k.
  - R_k is the rotation matrix corresponding to θ_k torsion rotation around hinge k.

We recall that if an axis-angle parametrization is used to describe a rotation, say the unit vector $\mathbf{u} = (u_1, u_2, u_3)$ and rotation parameter θ, then the rotation matrix of the corresponding rotation is given by

$$R = e^{\theta[\mathbf{u}]}$$

where

$$[\mathbf{u}] = \begin{bmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{bmatrix}$$

Therefore, with the notation introduced above, we have

$$R_k = e^{\theta_k[\mathbf{u}_k]} \tag{1}$$

A rotation around hinge k would move the atoms in cluster A if and only if hinge k appears on the path from cluster A to the root cluster. To find the position of atom $(A,i) \in A$ after hinge rotations, we need to first find the path from cluster A to the root cluster. Let $P = \{h_p,..,h_0\}$ be the hinges on the path from cluster A to the root cluster, listed in the order they appear on the path. We can apply hinge rotations in different order: for example, we can apply them in the order that they appear on path P or in the reverse order of their position on path P. To compute the gradient of the energy function with respect to torsional parameters, it is most advantageous to apply the hinge rotations in the order that they appear on path P. While this assertion is not surprising, we discuss the mathematical consequences of following the path and provide an example to show that any alternative strategy leads to more complex computations (also, see, e.g., ref 21).

Assume that $\mathbf{q}_{A,i}$ denotes the Euclidean coordinates of atom i in cluster A with respect to the fixed reference frame; then, after rotation with rotation matrix R_{h_k}, we have

$$\mathbf{q}_{A,i}^{\text{new}} = R_{h_k}(\mathbf{q}_{A,i} - O_{h_k}) + O_{h_k}$$

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

where $\mathbf{q}_{A,i}^{\text{new}}$ is the new position of atom i in cluster A and O_{h_k}, as mentioned before, is the center of the coordinate frame assigned to hinge h_k. If we apply torsional rotations on the path from cluster A to the root cluster, then the new position of atom i in cluster A would be

$$\mathbf{q}_{A,i}^{\text{new}} = R_{h_0}(\ldots(R_{h_p}(\mathbf{q}_{A,i} - O_{h_p}) + O_{h_p}) + \ldots - O_{h_0}) + O_{h_0}$$

The closed form formula for $\mathbf{q}_{A,i}^{\text{new}}$, which can be easily verified by induction on the length of the path, is as follows:

$$\mathbf{q}_{A,i}^{\text{new}} = (\Pi_{j=0}^{p} R_{h_j})(\mathbf{q}_{A,i} - O_{h_p}) + \sum_{k=0}^{p-1} (\Pi_{j=0}^{k} R_{h_j})(O_{h_{k+1}} - O_{h_k}) + O_{h_0}$$

After applying the internal motions of the ligand, we also need to rotate and translate the ligand as a rigid body. Recall that R and t represent, respectively, a rotation and a translation of the whole ligand. We denote the manifold (and group) of orientation-preserving rotations of the three-dimensional Euclidean space, $\mathbb{R}^3$, the so-called Special Orthogonal group, by SO(3), that is,

$$\text{SO}(3) = \{R \in \mathbb{R}^{3\times 3};\ R^T R = I;\ \det(R) = 1\}$$

The translation vector t is simply an element of the three-dimensional Euclidean space, $\mathbb{R}^3$.

We use our formulation of rigid body motion introduced in refs 18 and 19 for this purpose. In this formulation of rigid body motion, we allow the user to choose a center of rotation. Let $\mathbf{p}$ denote this center of rotation ($\mathbf{t}$ and $\mathbf{p}$ are defined with respect to the fixed reference frame). Then, the rigid motion of the atoms of the ligand corresponding to R, $\mathbf{t}$, and $\mathbf{p}$, is defined as follows. Let $\mathbf{q}$ denote the three-dimensional coordinates of an atom of the ligand with respect to the fixed reference frame. Then, after the rigid motion, the new coordinates of this atom are given by

$$R(\mathbf{q} - \mathbf{p}) + \mathbf{p} + \mathbf{t}$$

Furthermore, with this rigid motion, the center of rotation, $\mathbf{p}$, also moves to a new center of rotation given by $\mathbf{p} + \mathbf{t}$; in other words, it is translated by $\mathbf{t}$. In words, instead of rotating atoms of the ligand with respect to a fixed center of rotation, as is the case in rigid body motions represented by the manifold SE(3), we rotate atoms of the ligand with respect to a moving center of rotation that keeps a fixed relative distance to the atoms of the ligand. It turns out that this new formulation, arguably a more natural motion in the context of molecular docking problems, resolves some of the mathematical problems associated with the SE(3) manifold: the manifold $\text{SO}(3) \times \mathbb{R}^3$ associated with the new formulation is a direct product of its components $SO(3)$ and $\mathbb{R}^3$ both as groups and as Riemannian manifolds; in this case, there is no mismatch between the group and the natural Riemannian structures of $\text{SO}(3) \times \mathbb{R}^3$, and we do not face the complications that are associated with SE(3) rigid body transformations. (For more details, see refs 19) Thus, in what follows, we select the center of rotation to be the *center of mass of the ligand*. After a rigid body move, the new position of (A,i) atom, namely $\mathbf{q}_{A,i}^{\text{new}}$, will be

$$\mathbf{q}_{A,i}^{\text{new}} = R(\mathbf{q}_{A,i} - O) + O + \mathbf{t}$$

Order of Selecting Rotations in Ligand Displacement. As has been noted in the literature, the order of selecting rotations has an impact on the efficiency of computations.[21,22] We use an example to illustrate this issue in the context of local optimization. In Figure 2, the path from cluster 3 to the root cluster is $P = \{3,2\}$. Let $q_{3,i}$ be the coordinates of an atom in cluster 3. If we move along path P, we first rotate hinge 3 and then rotate hinge 2. By rotation around hinge 3, we will not touch atoms (2,0) and (1,0), which are the atoms corresponding to hinge 2. Therefore, the axis of rotation corresponding to hinge 2 would be $\mathbf{q}_{2,0} - \mathbf{q}_{1,0}$. The new position of atom $(3,i)$ can be easily calculated as follows:

$$\mathbf{q}_{3,i}^{\text{new}} = R_2(R_3(\mathbf{q}_{3,i} - O_3) + O_3 - O_2) + O_2$$

On the other hand, if we move in the reverse order along path P and move the atoms in cluster 3 by first rotating hinge 2, we would move atoms (2,6) and (3,0) which are the two ends of hinge 3. As a result, the axis of rotation corresponding to hinge 3 is no longer $\mathbf{u}_2 = \mathbf{q}_{3,0} - \mathbf{q}_{2,6}$, but $R_2(\mathbf{q}_{3,0} - \mathbf{q}_{2,6})$; therefore, the rotation matrix corresponding to a rotation around hinge 3 would be $e^{\theta_3 R_2[\mathbf{u}_2]}$, which is no longer a function of θ_3 only; in fact, in general, it is a function of all previous hinge rotations along the path. By contrast, when we move from the cluster to the root, each rotation matrix R_i corresponding to θ_i rotation about hinge i is only a function of θ_i.

For local minimization of the energy function, we need to calculate the gradient of the Cartesian coordinates of atoms with respect to torsional rotation parameters. By selecting to move on the path from the cluster toward the root, the energy functions will have a simpler form and the gradients can be more easily computed.

Manifold of Conformational Space: SO(3) $\times \mathbb{R}^3 \times T^d$. As stated earlier, the manifold of the conformational space corresponding to our representation of rigid-body is $\text{SO}(3) \times \mathbb{R}^3$.[19] To specify the manifold associated with torsional flexibilities, assume that the ligand has d rotatable bonds/hinges and let θ_i be the rotational parameter associated with the ith hinge. Allowing θ_i values to be any real number, $\theta = (\theta_1,\cdots\theta_d) \in \mathbb{R}^d$ represents an internal ligand motion. On the other hand, for any two θ and θ' such that

$$\theta - \theta' = 2\pi(i_1, i_2, \cdots, i_d)\ i_1, \cdots, i_d \text{ integers}$$

they represent the same torsional motion of the ligand. The above relationship defines an *equivalence relation* on $\mathbb{R}^d$. The quotient space of $\mathbb{R}^d$ with respect to this equivalence relation is a so-called d-dimensional *torus* manifold

$$T^d = \underbrace{S^1 \times S^1 \ldots \times S^1}_{d \text{ times}}$$

See, for example, ref 26 Section 3.2. The conformational space of the overall movements of the ligand including rigid and internal motions is, therefore, given by the direct product of $\text{SO}(3) \times \mathbb{R}^3$ and T^d. Thus, as in the case of rigid body optimization considered in ref 19, the search space of the optimization is a direct product manifold whose geodesics and exponential coordinates, as we show in the next section, are easy to compute. As a result, the basic local optimization approach of ref 19 can be adopted in this case as well.

Local Optimization Algorithm. We use a local parametrization of the conformational space $\text{SO}(3) \times \mathbb{R}^3 \times T^d$ using the exponential map defined on the tangent space of the manifold at $(I,\mathbf{0},\mathbf{0})$. Our motivation for using a local parametrization is to be able to take full advantage of the well-developed Euclidean optimization algorithms. Furthermore, local parametrization based

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

on the exponential map is particularly well-suited for the local optimization we consider because, in the vicinity of $(I,\mathbf{0},\mathbf{0})$, lines on the tangent space are mapped onto geodesics on the manifold; therefore, line search on the tangent space corresponds to moving along geodesics on the manifold.

Exponential Parametrization of SO(3) × $\mathbb{R}^3$ × T^d. Given that the manifold of the conformational space is a direct product of its component manifolds, the geodesics of the product manifold is the product of geodesics of the factor manifolds, and the exponential map on the product manifold is the product of the exponential maps on the factor manifolds. This fact simplifies the computation of the exponential map. (For a brief review of product Riemannian manifolds and optimization on them, see, e.g., ref 27 Appendix A). $\mathbb{R}^3$ is a trivial manifold, and its exponential map is the identify function of $\mathbb{R}^3$; the exponential map on T^d is the product of exponential maps defined for each coordinate, defined on $\mathbb{R}$, and given by $\exp(\theta_i) = (\cos(\theta_i), \sin(\theta_i))$; only the exponential map on SO(3) requires more detailed description.

The *tangent space* of SO(3) at I is denoted by so(3) and can be identified with the space of 3 × 3 skew-symmetric matrices. For $\omega = (\omega_1,\omega_2,\omega_3)^T \in \mathbb{R}^3$, let

$$[\omega] = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}$$

We can identify the tangent space, so(3), with $\mathbb{R}^3$; in this case, the standard Euclidean norm on $\mathbb{R}^3$ defines a Riemannian metric on SO(3): for ω, $\omega' \in \mathbb{R}^3$, the inner product of ω and ω' is defined by

$$\omega\cdot\omega' = -\frac{1}{2}tr([\omega][\omega']) = \frac{1}{2}tr([\omega]^T[\omega'])$$

where $tr(A)$ denotes the trace of matrix A.

The *exponential map* at identity $I \in$ SO(3) maps the tangent space at identity, that is, so(3), to SO(3). It is defined by

$$\exp_I(\omega) = e^{[\omega]}$$

where the term on the right is a matrix exponential. The right-hand side simplifies to give what is known as the Rodrigues formula (see, e.g., ref 26 Section 4.4.1)

$$e^{[\omega]} = I + \frac{\sin(\|\omega\|)}{\|\omega\|}[\omega] + \frac{(1-\cos(\|\omega\|))}{\|\omega\|^2}[\omega]^2$$

where $\|\omega\|$ is the Euclidean norm of ω. The exponential map at $R \in$ SO(3) is simply defined as $\exp_R(\omega) = Re^{[\omega]}$. Geodesics of SO(3) are given by $R(u) = Re^{[\omega]s}$, $\omega \in \mathbb{R}^3$, and $s \in \mathbb{R}$, and they correspond to the projection by the exponential map of lines going through the origin on the tangent space.

Gradient of the Energy Function with Respect to Exponential Map Parametrization. Consider the exponential coordinate parametrization of SO(3) × $\mathbb{R}^3$ × T^d described above, and let $(\omega,\mathbf{t},\theta) \in \mathbb{R}^{6+d}$ be a point in the tangent space of SO(3) × $\mathbb{R}^3$ × T^d at $(I,\mathbf{0},\mathbf{0})$, where $\omega = (\omega_1,\omega_2,\omega_3)$, $\mathbf{t} = (t_1,t_2,t_3)$, and $\theta = (\theta_1,\cdots,\theta_d)$ represents the orientation, translation, and torsional parameters, respectively. Then, the energy function can be viewed as a function of $(\omega,\mathbf{t},\theta)$. Assume that the ligand consists of m atoms and $\mathbf{q} = (\mathbf{q}_1,\cdots,\mathbf{q}_m)$ is the initial position of the ligand, where $\mathbf{q}_i = (x_i,y_i,z_i)$, $i = 1,\cdots, m$, and let ν represent one of the scalar components of $(\omega,\mathbf{t},\theta)$, namely, one of ω_j, $j = 1,2,3$, or t_j, $j = 1,2,3$, or θ_j, $j = 1,\cdots d$; then, the derivative of the energy function with respect to ν can be written as

$$\frac{\partial E}{\partial \nu} = \sum_{i=1}^{m} \frac{\partial E}{\partial x_i}\frac{\partial x_i}{\partial \nu} + \frac{\partial E}{\partial y_i}\frac{\partial y_i}{\partial \nu} + \frac{\partial E}{\partial z_i}\frac{\partial z_j}{\partial \nu}$$

The terms $(\partial x_i/\partial E)$ and $(\partial y_i/\partial E)(\partial z_i/\partial E)$ are easy to compute; see, for example, ref 28. We provide details for computing $(\partial x_i/\partial \nu)$, $(\partial y_i/\partial \nu)$, and $(\partial z_i/\partial \nu)$ as Supporting Information.

Optimization Algorithm. Given the exponential map parametrization, our energy minimization problem is defined on the $(6 + d)$-dimensional Euclidean space $\mathbb{R}^{6+d}$. From among the many deterministic algorithms available to solve local minimization problems on a Euclidean space, we have selected the limited memory Broyden−Fletcher−Goldfarb−Shanno (LBFGS) method,[29] quasi-Newton type approach. In our parametrization, the gradient and the Hessian of the energy function with respect to the parameters of optimization can be explicitly calculated. However, these are costly operations, evaluating the Hessian being significantly more costly than evaluating the gradient. Our choice of LBFGS is based on the fact that it uses only gradient information to obtain second order information about the energy function.

Denoting the elements of $\mathbb{R}^{6+d}$ by x, and the energy function by E, the LBFGS method consists of the following iterations[29]

$$x_{k+1} = x_k + \alpha_k d_k \tag{2}$$

where

$$d_k = -H_k \nabla E_k \tag{3}$$

where ∇E_k is the gradient of the energy function, H_k is the LBFGS approximation of the inverse of the Hessian of the energy function as described in ref 29, and α_k is an appropriately selected step-length as described in ref 30.

As pointed out in ref 29, the choice of H_0 influences the behavior of the algorithm. When the diagonal entries of the Hessian are all positive, it is recommended to let H_0 be a diagonal matrix with the diagonal entries of the inverse of the Hessian. Given that, in our problem, the diagonal entries of the Hessian are sometimes negative, we use the identity matrix as the initial H_0.

Scoring Function. In all the simulations described in the following section, we have used CHARMM potential with CHARMM19 parameters and with Analytical Continuum Electrostatics (ACE) solvation model.[31] The nonbonded cutoff was 13 Å, but we also explored the use of 12 and 14 Å cutoff values. The results, showing little sensitivity to the cutoff parameter, are provided in the Supporting Information Section D. The parametrization of small molecules was performed as described in ref 32.

RESULTS AND DISCUSSIONS

In this section, we describe three applications of our manifold optimization algorithm to local minimization in the context of small molecular docking. The performance of the manifold optimization algorithm is compared to the performance of all-atom optimization, the commonly used algorithm for refinement in protein−small molecule docking. The overall conclusion from these results is that in all cases the overall quality of solutions produced by the manifold optimization algorithm is equal or better than that of all-atom optimization while its computational efficiency is superior. Furthermore, in a separate section, we compare the behavior of the two

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

algorithms in some detail and draw conclusions that provide valuable insight about their performance.

Application to Protein Mapping. In the first set of experiments, we apply the proposed manifold optimization algorithm in the context of the protein mapping program FTMap.[33] The objective of mapping is to identify potentially favorable binding sites of the protein called *hot spots*. The FTMap mapping algorithm places molecular probes, small organic molecules that vary in size and shape, on a dense grid around the protein and finds favorable positions of the probes using empirical free energy functions. For each probe type, the first step of FTMap consists of global sampling of the 6D space of translations and rotations of the probe using a Fast Fourier Transform (FFT) correlation approach. In the next step, FTMap performs an off-grid local minimization of the docked structures. The resulting structures are then clustered, and the consensus clusters formed by clusters of several probes are identified as hot spots of the protein. In the current version of FTMap, the off-grid local minimization step is implemented using the CHARMM[28] potential and all-atom minimization. We compare the performance of all-atom (AA) minimization and our manifold optimization (MO) algorithm for off-grid local minimization of the probes.

Of the 16 probes considered in FTMap, 10 probes have no rotatable bonds and are fully rigid whereas the other 6 are flexible, each having a single C−O rotatable bond allowing for the rotation of the H atom of an OH group. In an earlier paper, we have reported the results of the comparison of our rigid-body manifold optimization algorithm with all-atom optimization for the rigid probes.[18] Here, we compare our flexible manifold optimization (MO) algorithm with all-atom (AA) minimization based on the remaining 6 flexible probes. The flexibility of these probes is captured by a single torsional angle.

To compare the two algorithms, 14 protein structures were selected from the Protein Data Bank (PDB);[34] 7 of these proteins have been the subject of a recent hot spot study.[35] All ligand and bound water molecules were removed prior to mapping. Each complex is evaluated using an energy expression that includes van der Waals and electrostatic interaction energy terms as well as solvation effects. In the current version of FTMap, the 2000 most favorable docked positions of each probe are refined by all-atom minimization of the CHARMM energy function. During this minimization the probe molecules are considered fully flexible, but the atoms of the receptor protein are taken as fixed.

We compare the two minimization algorithms based on the *quality of their solutions* and their *computational efficiency* as follows. Those cases where (i) the energy difference of the local minima found by the two algorithms is less than 0.01 kcal/mol or (ii) the local minima are within 0.05 Å RMSD distance of each other are considered as ties. In all other cases, the quality of the solution of one algorithm relative to the other is considered superior if it has a lower energy. It is worth noting that, in the context of mapping applications, including condition (ii) for ties is meaningful. In other applications, it may be more appropriate to use only closeness of energy, that is, condition (i), as a basis for declaring ties. To verify that including condition (ii) has not biased our conclusions, in the Supporting Information (Section C), we have used condition (i) only; the experimental results show that our conclusions about the quality of solutions and efficiency comparisons between the two algorithms remain unchanged.

We have selected *the number of energy function evaluations* needed to converge to a local minimum as the measure of computational efficiency of each algorithm. Given that energy function evaluations are the most costly operations and the same energy function is used for both algorithms, the number of energy function evaluations is a fair and appropriate basis for comparing the runtime of the two algorithms. As the convergence rate of all-atom minimization is low, we stop the algorithm after a maximum of 500 energy function evaluations. Also, to obtain a more reliable energy value for flexible manifold optimization and to compare the results obtained by the two algorithms in terms of energy values, we relax the bond lengths and bond angles of the conformations after manifold minimization by performing 20 steps of all-atom minimization. (In Section B of the Supporting Information, we provide a more detailed discussion of our choices for the maximum number of energy evaluations for all-atom minimization and the relaxation phase of manifold optimization.)

The results of the comparison of the two algorithms based on the six flexible probes are reported in Table 1. Note that, for each protein, the comparison is based on about 12 000 cases of local minimization. Proteins are identified by their four-letter PDB code[34] in the first column of the table. The second column gives the number of ties between the two algorithms, $E_{AA} = E_{MO}$. The third column is the number of conformations for which manifold optimization (MO) converged to a local minimum with a lower energy (i.e., $E_{MO} < E_{AA}$). The fourth column is the number of conformations for which all-atom minimization (AA) produced a better result, $E_{AA} < E_{MO}$. The fifth column is the average number of energy function evaluations of manifold optimization and the last column is the average number of energy function evaluations of all-atom minimization. As can be seen from the table, the manifold optimization yields equal or lower energies than the all-atom optimization in 80.98% of the 168 000 minimization runs

Table 1. Comparison of the Quality of Solutions and Computational Efficiency of Manifold Optimization (MO) with All-Atom Minimization (AA) for Flexible Probes[a]

	quality of solutions: which performs better			computational efficiency: avg. no. of steps	
protein	$E_{AA} = E_{MO}$	$E_{MO} < E_{AA}$	$E_{AA} < E_{MO}$	MO	AA
2CAB	5683	3691	1874	82	451
1IVG	5752	3319	2135	86	423
1BBC	4929	4033	1855	67	454
1O8A	5428	2103	2088	93	421
1F5L	5926	3195	2066	79	440
1S3E	5392	2013	1932	94	425
2O8T	5866	2951	2239	78	434
1W50	6003	2880	2685	94	420
1J2E	4833	1775	2001	93	389
1YES	5281	3797	1914	82	438
1HCL	5594	2955	1711	84	429
1THS	5810	3009	1896	87	420
1BN5	6295	2952	1849	77	419
1PUD	6371	3080	2137	72	427
	53.02%	**27.96%**	**19.02%**	**5.1**	**1**

[a]The fifth column is the average number of energy function evaluations of manifold optimization and the last column is the average number of energy function evaluations of all-atom minimization.

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

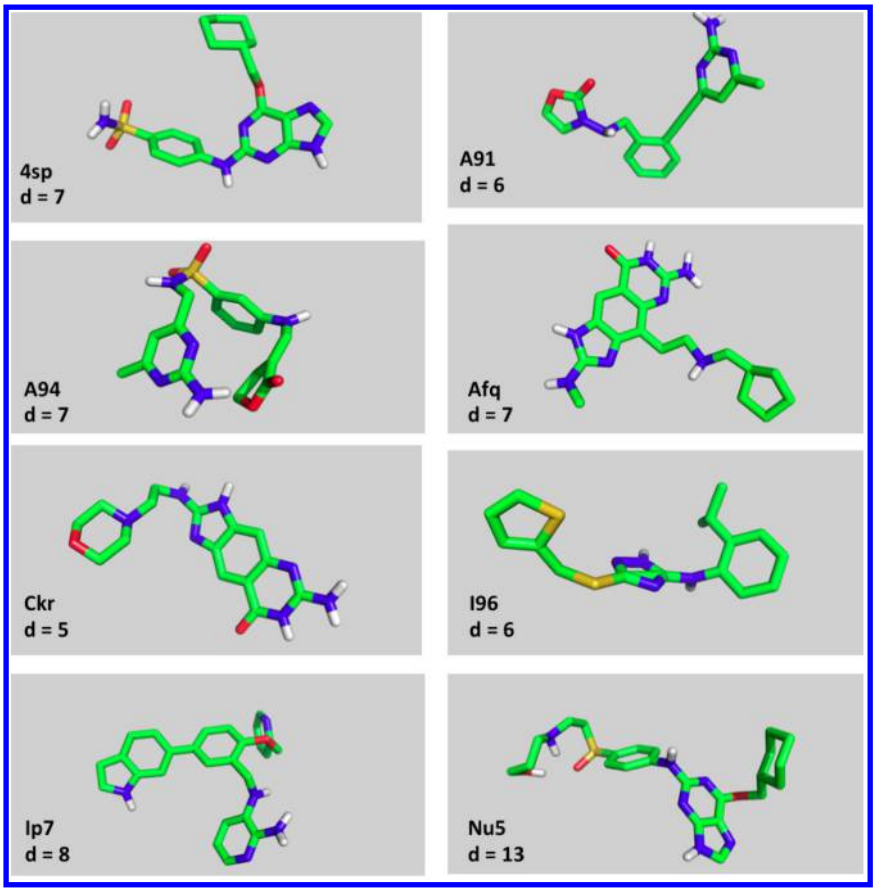


Figure 3. Small molecule ligands used in the second and third set of experiments. The number of rotatable bonds of each ligand is shown by *d*.

(12 000 for each of the 14 proteins), while the manifold optimization is 5.1 times faster than the all-atom approach.

Local Minimization with Fixed Receptor Protein. In our second and third set of experiments, by comparison with the first set, the number of rotatable bonds are substantially increased, and in the third set of experiments, we also allow movement of some of the side chains on the interface of the receptor. It is important to highlight the fact that the search space of our manifold optimization does not include adjustments to bond lengths and bond angles of the ligand while these flexibilities are included in all-atom optimization. Therefore, to make the comparison of the two algorithms more meaningful, after the completion of our manifold optimization, we apply a fixed number of relaxation steps using all-atom optimization. See the Supporting Information section for additional discussion of these choices. To compare the performance of the two algorithms we include the additional all-atom optimization steps in assessing the computational cost of our manifold optimization.

The small molecule ligands we consider, have been the subject of a study of our group in the context of Fragment Based Drug Design (FBDD).[35] They are presented in Figure 3. The number of rotatable bonds of these ligands varies between 6 and 13. In each case, unbound structures of the components of the complex, selected from the Protein Data Bank,[34] were docked using the docking program AutoDock.[25] For each complex, 150 structures with the lowest scores were selected and they were refined by minimizing their CHARMM[28] energy using all-atom minimization and our manifold optimization algorithm. We have used the following standard LBFGS stopping rule

$$\frac{|\nabla E|}{\max(1, \|x\|)} \leq \delta$$

as the primary stopping rule where ∇E is the gradient of the energy function and δ is a small threshold value. In our experiments, for relatively strict thresholds ($\delta = 10^{-4}$), all-atom (AA) minimization took a very long time to converge, and for more relaxed thresholds, the quality of its solution was inferior to manifold optimization followed by all-atom relaxation (MO). Therefore, as is often done in practice, we selected a cap on the number of relaxation steps and a cap on the number of all-atom optimization steps; if the energy gradient criterion above was not satisfied by the maximum number of iterations allowed, the algorithm was stopped and the results were reported. The maximum values were selected in such a way as to make the quality of the solutions provided by AA and MO comparable. (See Supporting Information (Section B) for more details.) We used 250 relaxation steps (using all-atom minimization) after the completion of the manifold optimization. In the case of all-atom minimization, due to the low rate of convergence we selected a cap of the 1000 minimization steps, after which the algorithm was stopped.

The two algorithms are compared based on 1200 test optimization runs (150 for each of the 8 complexes). As in the case of our protein mapping experiments, we compare the two off-grid minimization algorithms based on the quality of their solutions and their computational efficiency. The results of the comparison are shown in Table 2. Each protein is specified by

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 2. Comparison of the Quality of Solutions and Computational Efficiency of Manifold Optimization (MO) with All-Atom Minimization (AA) for Protein−Small Molecule Test Cases

complex description			quality of solutions: which performs better			computational efficiency: avg. steps	
complex	protein	ligand	$E_{AA} = E_{MO}$	$E_{MO} < E_{AA}$	$E_{AA} < E_{MO}$	MO	AA
1H1S	1HCL	4sp	0	36	114	386	1000
2G9X	1HCL	nu5	0	77	73	514	1000
2OG2	1YES	a91	16	54	80	348	1000
2QG0	1YES	a94	5	87	58	398	1000
3GE7	1PUD	afq	11	62	77	360	1000
3C2N	1PUD	ckr	3	105	42	363	1000
2OAZ	1BN5	i96	30	58	62	364	1000
2OHU	1W50	ip7	9	89	52	378	1000
			6%	**47%**	**47%**	**2.57**	**1**

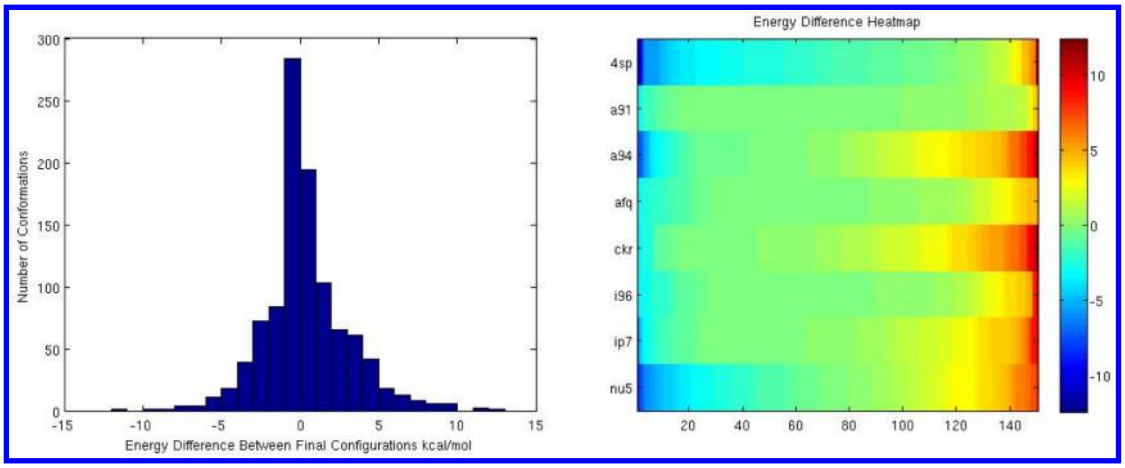


Figure 4. Bar diagram and heatmap plots of the energy differences between the final conformations of all-atom minimization (AA) and manifold optimization (MO) for protein−small molecule test cases.

its four-letter PDB code in the first column of the table. The second and third columns identify the receptor and the ligand of the complex, respectively. The fourth column reports the number of ties between the two algorithms, $E_{AA} = E_{MO}$. The fifth column is the number of conformations in which manifold optimization (MO) reached lower energy (i.e., $E_{MO} < E_{AA}$). The sixth column is the number of conformations for which all-atom minimization (AA) produced a better result, $E_{AA} < E_{MO}$. The seventh column is the average number of energy function evaluations of the manifold optimization algorithm and the eighth column represents that average for the all-atom optimization algorithm. According to these results, the manifold optimization algorithm has only slightly better performance based on the quality of solution criterion; however, it is on average 2.6 times faster than all-atom minimization.

Figure 4 provides a more refined comparison between the two algorithms based on the minimum energy values they find. The frequency graph gives the overall number of cases (across all ligands) where *the difference* between the minimum energy that each algorithm finds falls within a specified range. The heatmap plot in Figure 4 is a graphical representation of the difference between minimum energies for each ligand. The overall conclusion from Figure 4 is that the minimum energies that the two algorithms find are overall close to each other; that is, the two algorithms locate minimum conformations with comparable energies.

Finally, we take a closer look at the two phases of manifold optimization: the actual manifold optimization phase and the relaxation phase. Recall that in computing the computational cost of the manifold optimization algorithm, we have taken the cost of both phases into account. The average cost of manifold optimization phase across all ligands is about 113 function evaluations; the cost of relaxation phase, on the other hand, is more than twice that, namely 250 function evaluations.

Table 3 gives the average amount of movement for each phase in terms of the RMSD between the structures at the start and the completion of each phase. As can be seen, the bulk of the movement of the ligand (on average 1.86 Å RMSD) takes place during manifold optimization; changes during the relaxation phase most likely correspond to bond length and angle adjustments that lower the energy of the complex but do not move the ligand significantly (on average 0.25 Å RMSD).

Table 3. Average Change in RMSD during Manifold Optimization Phase and the Relaxation Phase

complex description			avg RMSD changes	
complex	protein	ligand	MO	relaxation
1H1S	1HCL	4sp	1.75	0.49
2G9X	1HCL	nu5	2.38	0.45
2OG2	1YES	a91	2	0.24
2QG0	1YES	a94	1.5	0.23
3GE7	1PUD	afq	1.72	0.18
3C2N	1PUD	ckr	1.88	0.11
2OAZ	1BN5	i96	1.68	0.14
2OHU	1W50	ip7	2	0.17
			1.86	**0.25**

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 4. Comparison of the Quality of Solutions and Computational Efficiency of Manifold Optimization (MO) with All-Atom Minimization (AA) for Protein−Small Molecule Test Cases with Receptor Side Chain Flexibility

complex description			quality of solutions: which performs better			computational efficiency: avg. steps	
complex	protein	ligand	$E_{AA} = E_{MO}$	$E_{MO} < E_{AA}$	$E_{AA} < E_{MO}$	MO	AA
1H1S	1HCL	4sp	5	40	105	645	1478
2G9X	1HCL	nu5	6	76	68	776	1500
2OG2	1YES	a91	73	38	39	608	1398
2QG0	1YES	a94	25	80	45	662	1461
3GE7	1PUD	afq	43	55	52	620	1487
3C2N	1PUD	ckr	20	83	47	627	1500
2OAZ	1BN5	i96	69	40	41	619	1426
2OHU	1W50	ip7	21	67	62	634	1500
			22%	**40%**	**38%**	**2.3**	**1**

One implication of this observation is that in some applications, such as identifying hot spots, it may be possible to forego the relaxation phase since the manifold optimization phase may be sufficient to locate the hot spots. In such cases, manifold optimization may be an order of magnitude more efficient when compared to all-atom optimization. We elaborate further on this issue in the section on the comparison of the behavior of manifold and all-atom optimization algorithms.

Local Minimization with Receptor Side Chain Flexibility. The only difference between our third and second set of experiments is that we now allow adjustments to some of the side chains on the interface of the receptor. For each ligand, following the heuristic protocol of Firedock,[36] we consider side chains of the receptor that are within 5 Å RMSD from the ligand, and for which the repulsive van der Waals between them and the rest of the complex is greater than 6 kcal/mol, as flexible/movable side chains. The manifold optimization search space in this case includes torsional rotations of the flexible side chains.

In the third set of experiments, we perform 500 steps of all-atom minimization during the relaxation phase of manifold optimization. For the all-atom minimization algorithm, we put a cap of 1500 steps after which the algorithm is stopped. The results of the comparison of the two algorithms are shown in Table 4. The format of the table is the same as Table 3: Each complex is identified by its four-letter PDB code in the first column of the table; the second and third columns identify the receptor and the ligand of the complex, respectively; the fourth column gives the number of ties between the two algorithms; the fifth column is the number of conformations in which manifold optimization converged to a local minimum with lower energy; the sixth column is the number of conformations in which all-atom minimization produced better result; the seventh and eighth columns report the average number of energy function evaluation for, respectively, the manifold and the all-atom optimization algorithms. Table 4 shows that the manifold optimization algorithm produces slightly better results, based on the quality of solution criterion; however, on average, it completes this task about 2.3 times faster than the all-atom minimization algorithm.

Figure 5 provides a more refined comparison of the two algorithms based on the lowest energy they identify. The frequency graph gives the overall number of cases (across all ligands) where the difference between the minimum energy that each algorithm finds falls within a specified range. The heatmap plot in Figure 5 is a graphical representation of the difference between the minimum energies for each ligand. The overall conclusion from Figure 5 is that the energies of the minimum conformations the two algorithms find are close to each other; that is, the two algorithms locate minimum conformations with comparable energies.

Finally, Table 5 gives the average amount of movement for the manifold optimization and relaxation phases of the manifold optimization algorithm in terms of the RMSD between the structures at the start and the completion of each phase. As in the second set of experiments, the bulk of the movement of the ligand (on average 1.83 Å RMSD) takes place during manifold optimization; the movement during the relaxation phase is smaller (on average 0.37 Å RMSD). The average cost of the manifold optimization algorithm across all ligands is about 150 function evaluations; the cost of relaxation phase, on the other hand, is more than three times that, namely 500 function evaluations. In cases when we can forego the relaxation phase, the manifold optimization algorithm is an order of magnitude more efficient when compared to all-atom optimization algorithm.

The overall conclusions from the third set of experiments are the same as those we arrived at in the second set of experiments: Manifold optimization and all-atom optimization produce conformations that have comparable energies; however, manifold optimization performs this task more efficiently. Furthermore, if we eliminate the relaxation phase of manifold optimization, justifiable in a variety of applications, manifold optimization is an order of magnitude more efficient than all-atom optimization.

Comparing the Behavior of Manifold and All-Atom Optimization Algorithms. In this section, we examine more closely the nature of the manifold and all-atom optimization algorithms. To this end, we plot (i) the energy of the intermediate conformations in the course of optimization as a function of the number of energy evaluations of each algorithm and (ii) the amount of displacements (in RMSD) of the ligand relative to the initial conformation, again as a function of the number of energy evaluations of each algorithm.

In Figure 6, the energy of the intermediate complex at each energy evaluation is plotted (on the vertical axis) against the number of energy evaluations (on the horizontal axis) for each optimization algorithm. The plots represent a single trajectory for each optimization algorithm. The trajectory of the manifold optimization algorithm is represented by the red curve and that of the all-atom optimization by the green curve.

Recall that the manifold optimization algorithm has two phases of (a) a manifold optimization phase and (b) a relaxation phase implemented by all-atom optimization. It is easy to

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

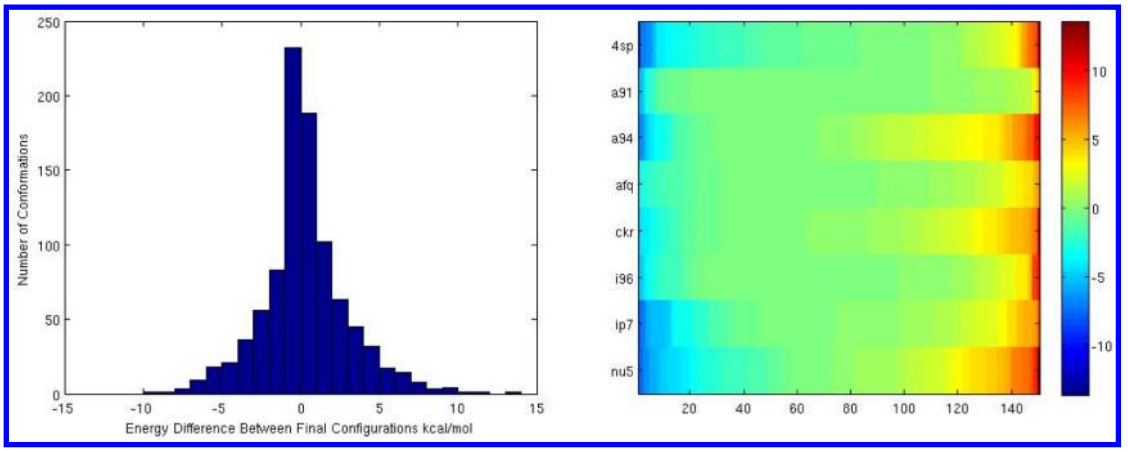


Figure 5. Bar diagram and heatmap plots of the energy differences between the final conformations of all-atom minimization (AA) and manifold optimization (MO) for protein−small molecule optimization with receptor side chain flexibility.

Table 5. Average Change in RMSD during Manifold Optimization Phase and the Relaxation Phase for the Third Set of Experiments

complex description			avg. RMSD changes	
complex	protein	ligand	MO	relaxation
1H1S	1HCL	4sp	1.78	0.76
2G9X	1HCL	nu5	2.29	0.66
2OG2	1YES	a91	1.99	0.37
2QG0	1YES	a94	1.43	0.33
3GE7	1PUD	afq	1.71	0.26
3C2N	1PUD	ckr	1.82	0.15
2OAZ	1BN5	i96	1.69	0.14
2OHU	1W50	ip7	1.96	0.29
			1.83	**0.37**

identify these two phases on the red trajectories for each ligand. As we have noted, and as is visible in Figure 6, the manifold optimization phase is often completed within 100−200 energy evaluations. Figure 6 shows that the energy of the resulting configuration is reduced further and significantly in the relaxation phase; the sharp drop of the energy in the early part of the relaxation phase suggests that this reduction is most likely achieved by a simple adjustment of the bond lengths and bond angles of the ligand, adjustments that were not within the search space of the manifold optimization algorithm. The rapid reduction of the energy in the early part of the all-atom optimization algorithm (green curve) reinforces this conclusion. After this initial period, as can be seen, the all-atom optimization algorithm reduces the energy in a significantly more gradual fashion.

Figure 7 is the result of averaging 150 optimization trajectories for each ligand, each trajectory corresponding to an initial conformation generated by the docking program AutoDock. Again, the red curve represents the manifold optimization algorithm and the green curve the all-atom optimization. As can be seen, the same behavior that was observed in the single trajectory case in Figure 6 is observable in the average case as well, suggesting that the conclusions we arrived at above are more broadly valid and that they have not been the artifact of a single nonrepresentative trajectory.

To test our hypothesis that the energy reduction during the relaxation phase is mostly due to adjustments of bond lengths

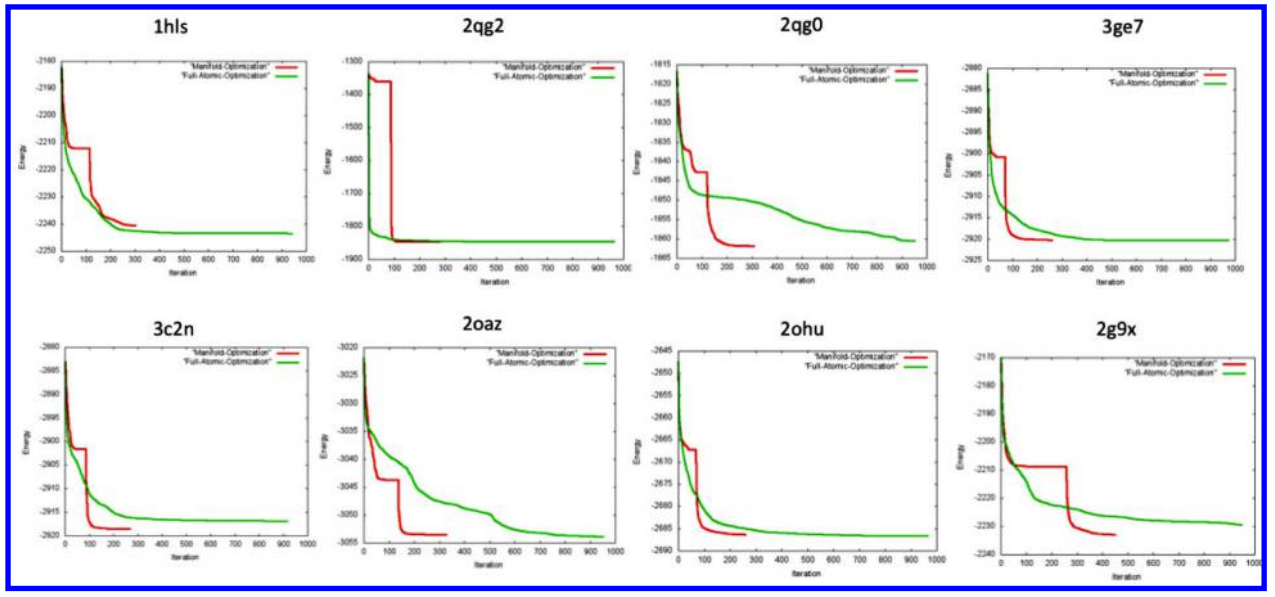


Figure 6. Comparing the behavior of the manifold optimization and all-atom optimization for a single conformation of the small molecule.

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

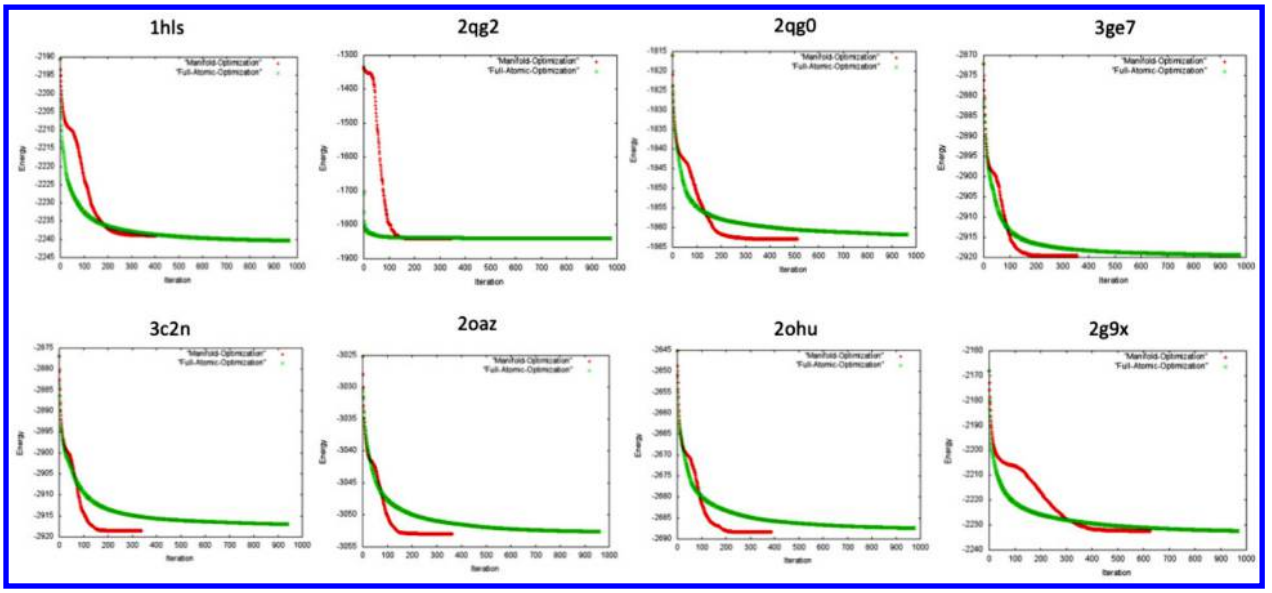


Figure 7. Comparing the average behavior of the manifold optimization and all-atom optimization over all conformations of small molecules.

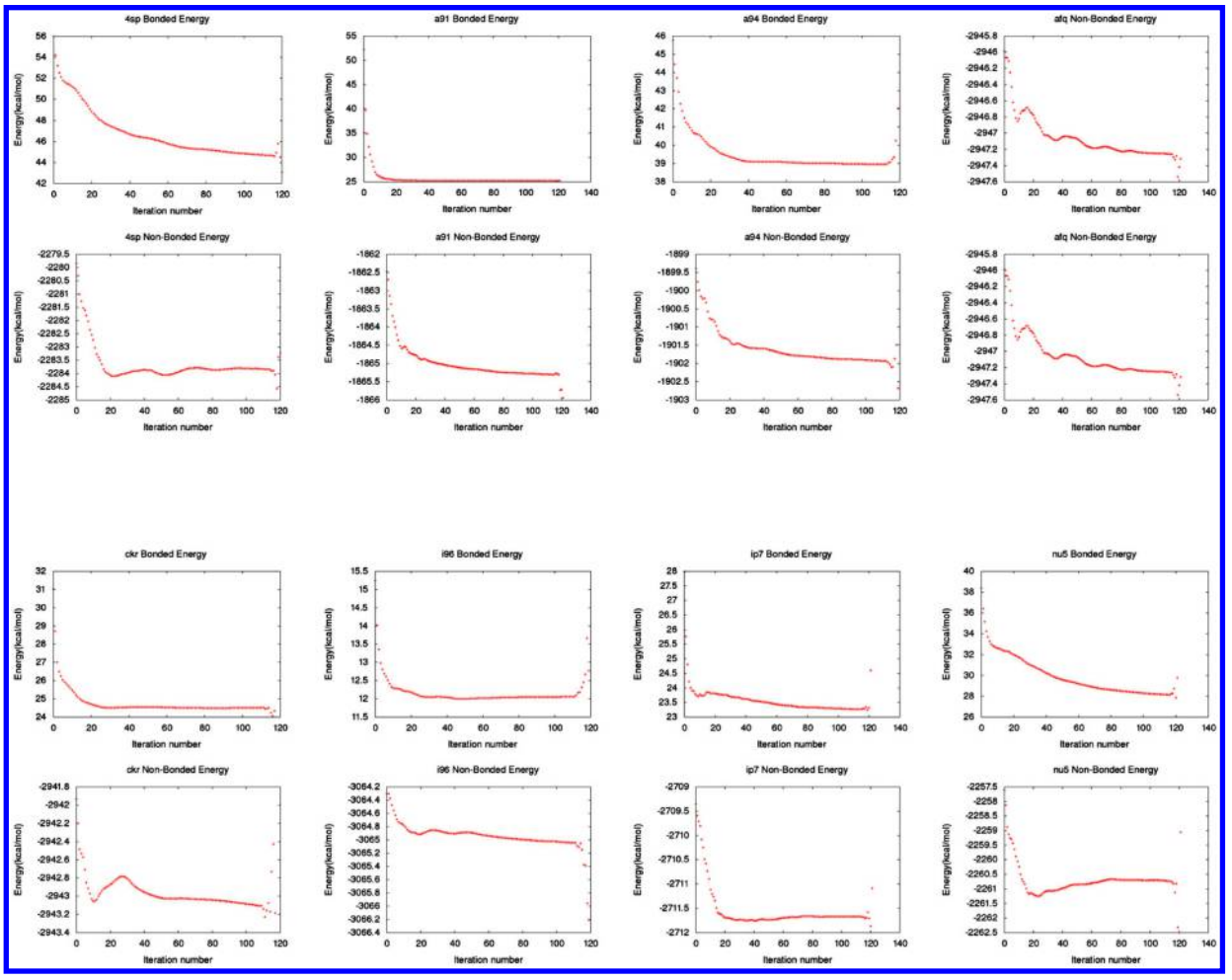


Figure 8. Magnitude of change in bonded and nonbonded energy terms during the relaxation phase of manifold optimization.

and bond angles of the ligand, we plot changes in the bonded energy terms (bond length, bond angle, torsion, and impropers) and nonbonded energy terms (van der Waals and electrostatic energy terms, the latter calculated by the Analytical Continuum Electrostatics (ACE) model as implemented in CHARMM) during the relaxation phase of MO for each of the probes in Figure 8. Table 6 presents the contributions of bonded and nonbonded energy terms as a percentage of the overall reduction in the total energy during the relaxation phase of MO. Figure 8 and Table 6 confirm our hypothesis that the

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 6. Percentage Change in Total Energy Due to Changes in Bonded and Non-Bonded Energy Terms during the Relaxation Phase of Manifold Optimization

complex description			percentage of change in total energy	
complex	protein	ligand	bonded terms	nonbonded terms
1H1S	1HCL	4sp	74%	26%
2G9X	1HCL	nu5	77%	23%
2OG2	1YES	a91	91%	9%
2QG0	1YES	a94	73%	27%
3GE7	1PUD	afq	86%	14%
3C2N	1PUD	ckr	85%	15%
2OAZ	1BN5	i96	81%	19%
2OHU	1W50	ip7	67%	33%
			79%	**21%**

energy reduction during the relaxation phase is mostly due to adjustments of bond lengths and bond angles of the ligand.

We now turn to the question of the displacement of the ligand by the two optimization algorithms. Figure 9 includes eight graphs, one for each ligand. Each graph shows the displacements of the ligand after 50, 100, 150, 200, 400, 500, 750, and 1000 steps of the all-atom optimization algorithm. At each of these points, the box plot of the displacements of the ligand, relative to its initial conformation, is plotted. The box plot is for 150 different optimization trajectories corresponding to different initial conformations and the displacements are measured in Ångstrom RMSD.

Figure 10 presents the same displacements for the manifold optimization algorithm. Note that for the manifold optimization algorithm, displacements of the ligand are shown after 20,

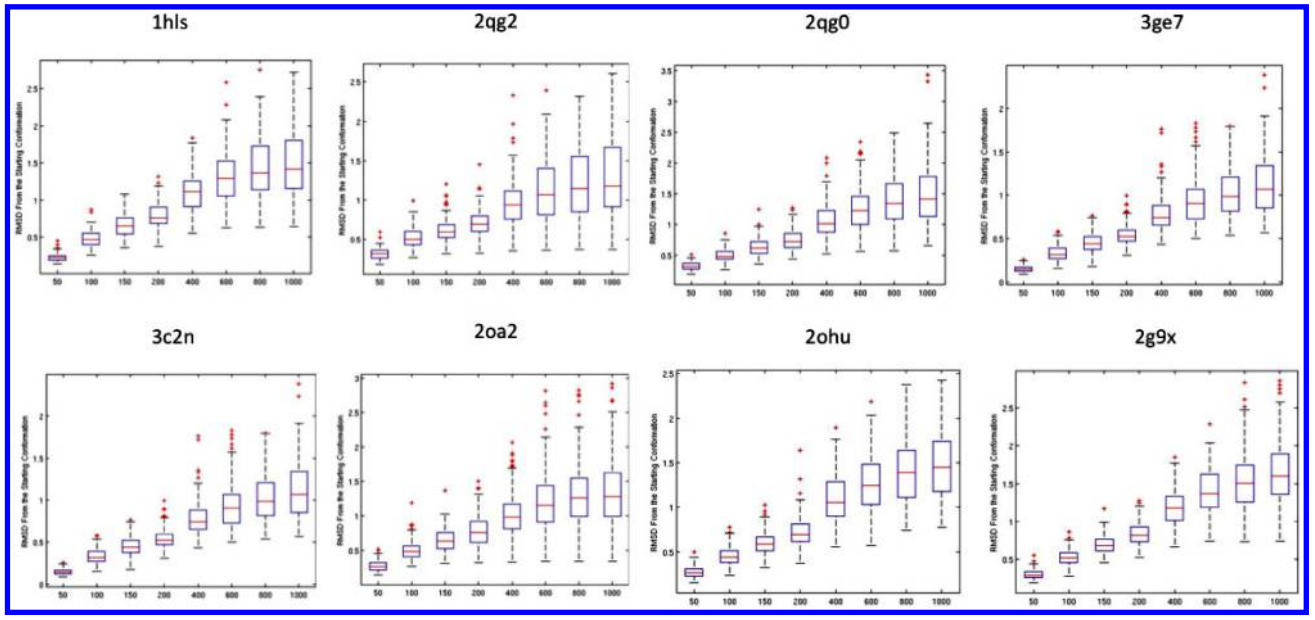


Figure 9. Displacement of the ligand by the all-atom optimization algorithm.

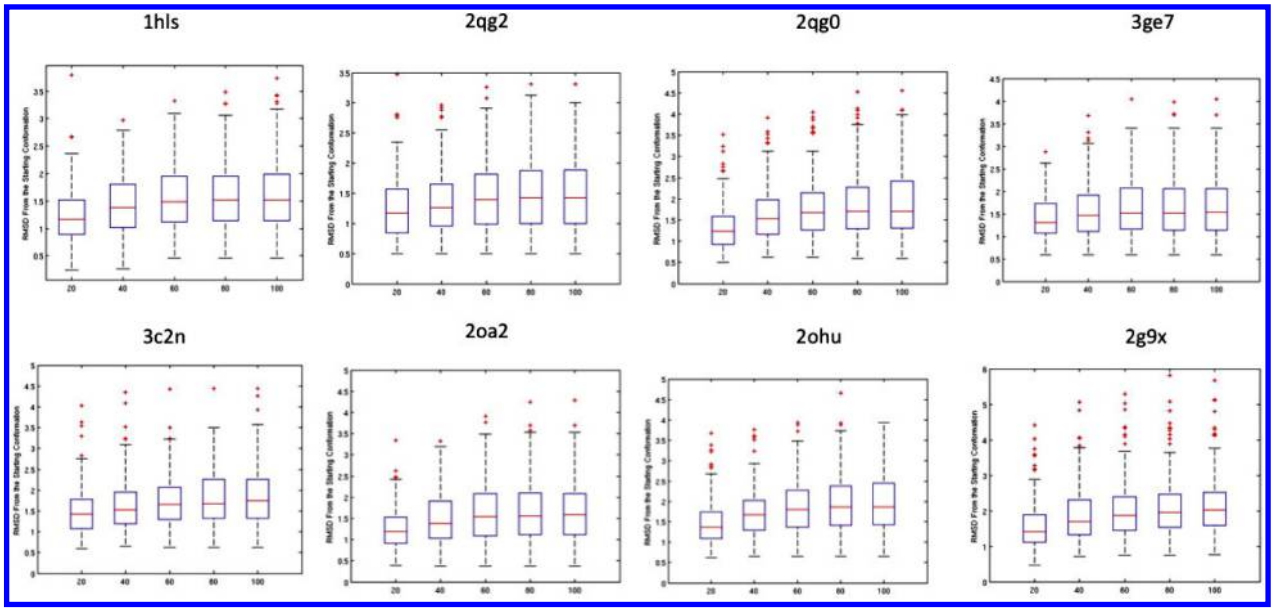


Figure 10. Displacement of the ligand by the manifold optimization algorithm.

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

40, 60, 80, and 100 steps of the algorithm. A comparison of the graphs for the two optimization algorithms shows that the manifold optimization algorithm moves the ligand more rapidly toward a local minimum while the movement of the all-atom optimization is more gradual. This pattern is observed across all ligands and can only be attributed to the inherent difference between the behavior of the two algorithms: the significantly more rapid movement of the manifold optimization algorithm, possibly due to its lower dimensional search space, explains its superior computational efficiency to the slower moving all-atom optimization algorithm.

CONCLUSIONS

In this paper, we have extended the manifold optimization approach, recently developed for rigid body minimization, to minimization of flexible molecules in application to docking. For this extension, we combine a representation of the rigid moves of the ligand by an appropriate manifold with a manifold representation of the flexibility of the ligand. The second manifold corresponds to the internal coordinate representation of flexible moves of the ligand. The structure of the resulting combined manifold is such that its geodesics are simple to compute when an exponential parametrization of manifold is selected. We have given detailed description of the exponential parametrization and algorithms for computing the gradient of the energy function with respect to the parameters of the exponential parametrization. We have provided a comparison of the performance of the proposed local optimization algorithm with all-atom optimization, and showed that the manifold optimization algorithm is substantially more efficient than the all-atom algorithm while producing solutions of comparable quality. We have also provided a detailed comparison of the nature of the two optimization algorithms that provides valuable insight about the superior performance of the approach based on manifold representation, which results in substantial reduction in the dimensionality of the search space. We have presented three applications with flexibility restricted to rotation around an increasing number of rotatable bonds. However, the manifold formalism presented here and the method described also apply to wide variety of docking molecules that include rigid regions connected by flexible linkers, for example, rigid secondary structure elements linked by flexible loops or multidoman proteins that with some level of hinge motion. The code is available to the community under open source license (at http://cluspro.bu.edu/Code/Code_Rigtree.tar) and with minimal effort can be incorporated into any molecular modeling package.

ASSOCIATED CONTENT

Supporting Information

Computing derivatives with respect to parameters of exponential parametrization. Choice of stopping criterion and parameters for manifold optimization and all-atom optimization. Tie criteria for the performance of manifold optimization and all-atom optimization. Sensitivity of the solutions and efficiency of manifold optimization to the cutoff parameter of the implicit solvent model. This material is available free of charge via the Internet at http://pubs.acs.org/

AUTHOR INFORMATION

Corresponding Author

*E-mail: midas@bu.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Research supported in part by National Institutes of Health/National Institute of General Medical Sciences (NIH/NIGMS) under grants R01-GM093147, GM064700, and GM061867, by the National Science Foundation (NSF) under grants CNS-1239021, DBI1147082, and IIS-1237022, by the Army Research Office (ARO) under grants W911NF-11-1-0227 and W911NF-12-10390, and by the Office of Naval Research (ONR) under grant N00014-10-1-0952.

REFERENCES

(1) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 409−443.
(2) Smith, G.; Sternberg, M. *Curr. Opin. Struct. Biol.* **2002**, *12*, 28−35.
(3) Vajda, S.; Kozakov, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 164−170.
(4) Ghemtio, L.; Perez-Nueno, V. I.; Leroux, V.; Asses, Y.; Souchet, M.; Mavridis, L.; Maigret, B.; Ritchie, D. W. *Comb. Chem. High Throughput Screening* **2012**, *15*, 749−69.
(5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−49.
(6) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. *J. Med. Chem.* **2006**, *49*, 5851−5.
(7) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. *J. Chem. Inf. Model.* **2008**, *48*, 2214−25.
(8) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. *J. Mol. Biol.* **2003**, *331*, 281−299.
(9) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 538−548.
(10) Davis, I. W.; Baker, D. *J. Mol. Biol.* **2009**, *385*, 381−392.
(11) Paschalidis, I. C.; Shen, Y.; Vajda, S.; Vakili, P. *IEEE Trans. Autom. Control* **2007**, *52*, 664−676.
(12) Shen, Y.; Vakili, P.; Vajda, S.; Paschalidis, I. C. Optimizing noisy funnel-like functions on the Euclidean group with applications to protein docking. *Proc. 46th IEEE Conf. Decision Control (CDC)* **2007**, 4545−4550.
(13) Shen, Y.; Paschalidis, I. C.; Vakili, P.; Vajda, S. *PLoS Comput. Biol.* **2008**, *4*, e1000191.
(14) Kusumaatmaja, H.; Whittleston, C. S.; Wales, D. J. *J. Chem. Theory Comput.* **2012**, *8*, 5159.
(15) Gwak, S.; Kim, J.; Park, F. C. *IEEE Trans. Robot. Autom.* **2003**, *19*, 65−74.
(16) Smith, S. T. Optimization techniques on riemannian manifolds. Proc. Fields Inst. Workshop on Hamiltonian and Gradient Flows. *Algorithms, and Control.* **1994**, arXiv:1407.5965.
(17) Absil, P. A.; Mahony, R.; Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*; Princeton University Press: Princeton, NJ, 2008.
(18) Mirzaei, H.; Beglov, D.; Paschalidis, I. C.; Vajda, S.; Vakili, P.; Kozakov, D. *J. Chem. Theory Comput.* **2012**, *8*, 4374−4380.
(19) Mirzaei, H.; Kozakov, D.; Beglov, D.; Paschalidis, I. C.; Vajda, S.; Vakili, P. A new approach to rigid body minimization with application to molecular docking. *Proc. 51th IEEE Conf. Decision Control (CDC)* **2012**, 2983−2988.
(20) Abagyan, R.; Totrov, M.; Kuznetsov, D. *J. Comput. Chem.* **1994**, *15*, 488−506.
(21) Jain, A.; Vaidehi, N.; Rodriguez, G. *J. Comput. Phys.* **1993**, *106*, 258−268.
(22) Schwieters, C. D.; Clore, G. *J. Magn. Reson.* **2001**, *152*, 288−302.
(23) Kusumaatmaja, H.; Whittleston, C. S.; Wales, D. J. *J. Chem. Theory Comput.* **2012**, *8*, 5159−5165.
(24) MacKerel Jr., A.; Brooks, C., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. *CHARMM: The Energy Function and Its Parameterization*

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

with an Overview of the Program; The Encyclopedia of Computational Chemistry; John Wiley & Sons: Chichester, 1998; Vol. *1*, pp 271−277.
(25) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Olson, D. S. G. A. J. *J. Comput. Chem.* **2009**, *16*, 2785.
(26) Selig, J. M. *Geometric Fundamentals of Robotics*; Springer: New York, 2005.
(27) Ma, Y.; Kosecka, J.; Sastry, S. *Int. J. Comput. Vision* **2001**, *44*, 219−249.
(28) Brooks, B. R.; Bruccoleri, R. E.; Olafson, D. J.; States, D.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(29) Liu, D. C.; Nocedal, J. *Math. Prog.* **1989**, *45*, 503−528.
(30) Xie, D.; Schlick, T. *Optim. Method Softw.* **2002**, *17*, 683−700.
(31) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578−1599.
(32) Ngan, C. H.; Bohnuud, T.; Mottarella, S. E.; Beglov, D.; Villar, E. A.; Hall, D. R.; Kozakov, D.; Vajda, S. *Nucleic Acids Res.* **2012**, *40*, W271−5.
(33) Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. *Bioinformatics* **2009**, *25*, 621−627.
(34) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.
(35) Hall, D. R.; Ngan, C.; Zerbe, B.; Kozakov, D.; Vajda, S. *J. Chem. Inf. Model.* **2012**, *52*, 199−209.
(36) Andrusier, N.; Nussinov, R.; Wolfson, H. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 139−159.

DOI: 10.1021/ct500155t
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX