Published in final edited form as:
*Proc IEEE Conf Decis Control*. 2014 December ; 2014: 4623–4628. doi:10.1109/CDC.2014.7040111.

# A Subspace Semi-Definite programming-based Underestimation (SSDU) method for stochastic global optimization in protein docking*

**Feng Nan**[†], **Mohammad Moghadasi**[†], **Pirooz Vakili**[‡], **Sandor Vajda**[¶], **Dima Kozakov**[¶], and **Ioannis Ch. Paschalidis**[§]

[†] Division of Systems Engineering, Boston University

[‡] Department of Mechanical Engineering and Division of Systems Engineering, Boston University

[¶] Department of Biomedical Engineering, Boston University

## Abstract

We propose a new stochastic global optimization method targeting protein docking problems. The method is based on finding a general convex polynomial underestimator to the binding energy function in a permissive subspace that possesses a funnel-like structure. We use Principal Component Analysis (PCA) to determine such permissive subspaces. The problem of finding the general convex polynomial underestimator is reduced into the problem of ensuring that a certain polynomial is a Sum-of-Squares (SOS), which can be done via semi-definite programming. The underestimator is then used to bias sampling of the energy function in order to recover a deep minimum. We show that the proposed method significantly improves the quality of docked conformations compared to existing methods.

## I. Introduction

Proteins are macromolecules found in abundance within the cell that play a key role in a variety of cellular functions such as metabolic control, signal transduction, immune response, and gene regulation. To that end, proteins interact with each other and other molecules. At each interaction at least two molecules are involved: a *receptor* and a *ligand*. The receptor is typically a protein within the cell or on the cell membrane. The ligand can be another protein or a smaller molecule (e.g., a drug) that binds to a specific site on the receptor.

The prediction of the 3-dimensional (3-D) structure of a receptor-ligand complex is known as the *protein docking* problem and is an important problem in computational structural biology. It is a critical problem as it is the basis of protein structure design, homology

[*] Research partially supported by the NIH/NIGMS under grants GM093147 and GM061867, by the NSF under grants CNS-1239021, IIS-1237022, and DBI-1147082, by the ARO under grants W911NF-11-1-0227 and W911NF-12-1-0390, and by the ONR under grant N00014-10-1-0952.

[§] Corresponding author. Department of Electrical & Computer Engineering, and Division of Systems Engineering, Boston University, 8 Mary's St., Boston, MA 02215, yannisp@bu.edu, http://ionia.bu.edu/..
{fnan@bu.edu, mohamad@bu.edu, vakili@bu.edu, midas@bu.edu, vajda@bu.edu}

modeling, and helps elucidate protein association. Experimental techniques, such as X-ray crystallography or Nuclear Magnetic Resonance (NMR), do provide 3-D structure information, but they are usually expensive, time-consuming and may not be universally applicable to short-lived molecular complexes. Therefore, computational methods are very much needed and have attracted considerable attention over the last two decades.

We know from the principles of thermodynamics that proteins bind to each other in a way that minimizes the Gibbs free energy of the bound complex. In this regard, the protein docking problem can be seen as a control problem in which one protein – the ligand – is “driven” to approach and dock with the fixed receptor. Considering both molecules as rigid, the control variables that describe the motion of the ligand take values in the space of rigid-body motion represented by the *Special Euclidean group SE*(3). An element of *SE*(3) is of the form $\tilde{\psi}=(\boldsymbol{\rho},\boldsymbol{\Omega})$, where $\boldsymbol{\rho} \in \mathbb{R}^{\mathbf{3}}$ describes the coordinates of a point on the ligand with respect to an inertial frame reference on the receptor, and Ω is a rotation matrix (in *SO*(3) – the special orthogonal group) that specifies the orientation of the ligand with respect to same inertial frame on the receptor. To represent these rotations, one can take the *tangent space* of *SO*(3) at the identity matrix **I**, denoted by *so*(3), and represent a point on the tangent space by $\boldsymbol{\omega} \in \mathbb{R}^{\mathbf{3}}$ – the so called exponential coordinates (see [1], [2] for a more extensive discussion of this representation).

The binding free energy function docking methods seek to minimize can be expressed as a function of $\psi=(\boldsymbol{\rho},\boldsymbol{\omega})$ and denoted by $f:\mathbb{R}^6 \to \mathbb{R}$. This energy function is composed of several force-field energy terms (e.g., Lennard-Jones potential, a solvation term, a hydrogen-bonding term, a Coulomb potential, etc.) that act in different space scales. This leads to having multiple deep funnels and numerous local minima of less depth over its domain, thereby, making global optimization quite challenging.

Our work in this paper develops a new *stochastic global optimization* method which we call Subspace Semi-Definite programming-based Underestimation (SSDU). It targets what is known as the *refinement problem* which amounts to globally minimizing *f* but over a certain limited part of the conformational space. Our approach follows our earlier work [3], [4] and solves a *semi-definite programming problem* to find general convex underestimators approximating the envelope spanned by local minima of the energy function. We use this underestimator to guide us where to continue to randomly search and generate new local minima which are then used to refine the underestimator. The main novelty we introduce in this paper is that optimization over the 6-D space of $\psi=(\boldsymbol{\rho},\boldsymbol{\omega})$ is effectively reduced to a 3-D subspace by using space dimensionality reduction techniques. The underestimator and random sampling of the energy function are constrained in this subspace, hence, the name SSDU. This idea is motivated by our recent work that studied the behavior of two different force-fields and established the same dimensionality-reduced structure [5]. We develop a general form of SSDU that allows for arbitrary convex polynomial underestimators. Our numerical results show that SSDU outperforms existing docking refinement methods.

**Notation**: Vectors will be denoted using lower case bold letters and matrices by upper case bold letters. For economy of space we write **v** = ($v_1$,...,$v_n$) for **v** $\in \mathbb{R}^n$. Prime denotes transpose. For a matrix **P**, $\mathbf{P} \succeq 0$ indicates positive semidefiniteness.

## II. Background on docking methods

The most successful docking methods rely on a multistage procedure that begins with a rigid-body global search on a grid sampling a huge number of docked receptor-ligand conformations. The energy function is approximated by a correlation function and energy evaluation for all these samples is done leveraging the Fast Fourier Transform (FFT). In our work, the initial sampling is conducted using the automated server *ClusPro 2.0* [1] which is based on a docking program called PIPER [6]. PIPER provides a set of conformational clusters – each can be viewed as an ensemble of receptor-ligand conformations that needs to be further refined by a process we call *Protein Docking Refinement*. The main objective of refinement is to locate the global energy minimum within the cluster, sampling and optimizing off-grid but within the conformation space defined by the cluster. Rather than providing a single conformation as an output of the refinement procedure, what is typically done is to fix the size of the cluster but allow members to exit as new conformations with lower energy values get generated and added to the cluster in the process of refinement. In this setting, the common metric used to assess the performance of a refinement algorithm is the percentage of "accurate" predictions (say below 5Å RMSD to the native) within the cluster produced at the end of the refinement. Thus, while refinement does not use the native structure (known using experimental methods such as X-ray crystallography), scoring of refinement algorithms is done by assessing how "close" the cluster is to that structure.

## III. Related work and key contributions

The docking refinement problem has received significant attention since computational methods emerged in the field of structural biology. However, it is still a very challenging problem due to the complexity of the energy landscape of protein-protein interactions. Many docking methods use grid-based search algorithms [7], [8], [9] and assume proteins are rigid. Several other works apply the idea of a progressively improving approximation of the energy function by using a *Monte Carlo Minimization (MCM)*-based approach [10]. In this fashion, the algorithm iteratively performs rigid-body and flexible motions of the ligand towards the receptor for each input conformation. Similar approaches which consider the flexibility of protein interfaces and employ a powerful local minimization step in each iteration of an MCM-based search have been explored in our earlier work [11], [12].

Another recent approach to the docking problem uses the funnel-like shape of the energy function in order to locate the lowest free energy basin and perform a search in that area. The method proposed in [13] considers the dominant driving force-fields of the protein binding process, which allows for an efficient selection of a downhill path on the evolving receptor-ligand-free energy landscape. Later the idea of using *convex canonical quadratic underestimators* to approximate the envelope spanned by the local minima of the energy

[1] http://cluspro.bu.edu/login.php.

function was introduced in the *Convex Global Underestimator (CGU)* method [14]. The *Semi-Global Simplex (SGS)* method [15] uses an exhaustive multistart Simplex search of the protein surface. The main limitation of CGU in higher dimensions has been demonstrated in [3] to be the restricted class of underestimators it uses. In [3] and [4], the Semi-Definite programming-based Underestimation (SDU) method was proposed which addresses all the aforementioned issues. SDU also uses a *quadratic convex function* to underestimate the envelope spanned by the local minima, but it considers the class of general convex quadratic functions for underestimation and uses a biased exploration strategy guided by the underestimator.

The key contributions of our work in this paper are:

- **(i)** Dimensionality reduction: we have shown that optimization over the 6-D space (as in [14] and [15]) or 5-D space (as in [3] and [4]) can be effectively reduced to a 3-D space by applying *Principal Component Analysis (PCA)* to the refinement input structures.
- **(ii)** Introducing a more general class of convex functions (convex polynomials) for underestimation.

## IV. Methodology

SSDU consists of three key components: (1) dimensionality reduction via PCA, (2) underestimation using general convex polynomials, and (3) re-sampling around the global minimum indicated by the convex underestimator. In this section we will discuss these steps in detail.

### A. Dimensionality reduction

As described earlier, a conformation is a 6-D vector $\psi = (\rho, \omega)$ where $\boldsymbol{\rho} \in \mathbb{R}^{\mathbf{3}}$ is the translation vector from the ligand center of mass to the receptor center of mass and $\omega = (w_1, w_2, w_3)$ are the exponential coordinates that define the exponential map from *so*(3) to *SO*(3) (see details in [4], [5]). For the translation vector $\rho = (r_1, r_2, r_3)$ we use spherical coordinates and express it as $(r, a, b)$ where $r = \|\rho\|$ is the length of the vector and $a, b$ specify the exponential coordinates of the azimuth angle (between the projection of $\rho$ on the $r_1 r_2$ plane and the $r_1$ axis) and zenith angle (between $\rho$ and the $r_3$ axis) of the translation vector. In particular, if $\theta$ and $\varphi$ are these angles, $(a,b) = (-\sin\theta \cdot \varphi, \cos\theta\,\varphi)$. With this parametrization, and by re-using the $\psi$ notation, we can write

$$\psi = (\mathbf{r}, \mathbf{a}, \mathbf{b}, \omega_{\mathbf{1}}, \omega_{\mathbf{2}}, \omega_{\mathbf{3}}) . \quad (1)$$

We also re-use $f$ to denote the energy function we wish to minimize, viewed now as a function of these 6 variables.

A first step towards dimensionality reduction is to observe that in low-energy clusters like the ones we wish to refine, the center-to-center distance $r$ between a ligand and the receptor does not exhibit significant variation (see also [4]). Thus, removing $r$ from $\psi$ and redefining $f$, we are left with optimizing $f$(x) over

$$\mathbf{x}=(a,b,\omega_1,\omega_2,\omega_3)\,. \quad (2)$$

It turns out that for protein complexes there exist further directions along which significant variation in coordinates can be ruled out [5]. Essentially, the biophysical explanation is that as the ligand gets in contact with the receptor and parts of the ligand enter in crevices on the receptor, there are fewer directions along which the ligand can move or rotate. We propose to use PCA to identify these restrictive directions and eliminate them from the free energy minimization problem. To that end, let us sample *f* and generate *K* local minima:

$$\left(\mathbf{x}^{(i)}, f^{(i)}=f\left(\mathbf{x}^{(i)}\right)\right), \qquad i=1,\dots,K.$$

Let $\bar{\mathbf{x}}=\frac{1}{K}\sum_{i=1}^{K}\mathbf{x}^{(i)}$ be the mean of the *K* local minima in the 5-D subspace. Let $\mathbf{X}\in\mathbb{R}^{5\times K}$ be the matrix whose columns are $\mathrm{x}^{(\mathrm{i})} - \bar{\mathrm{x}}$, $i = 1,\dots,K$. We compute the eigen decomposition

$$\mathbf{X}\mathbf{X}'=\mathbf{W}\mathbf{\Sigma}\mathbf{W}', \quad (3)$$

where Σ diagonal matrix containing the eigenvectors of $\mathbf{XX'}$ and the column of $\mathbf{W}\in\mathbb{R}^{5\times 5}$ are the corresponding eigenvectors. The transformation $\mathrm{z}^{(i)} = \mathrm{W}' (\mathrm{x}^{(\mathrm{i})} - \bar{\mathrm{x}})$ transforms the *i*th sample point into the principal coordinates.

We then observe [5] that in most protein-protein complexes the first 3 eigenvalues are substantially larger than the remaining two, implying that the principal coordinates $z_1,z_2,z_3$ show much more variation than $z_4,z_5$. Therefore, we will construct the semidefinite underestimator in this 3-D subspace (the *permissive subspace*). Specifically, we let

$$\phi^{(i)}=\left(z_1^{(i)},z_2^{(i)},z_3^{(i)}\right) \quad (4)$$

be the new coordinates of the *i*th sample point in the permissive landscape.

### B. Convex polynomial underestimator

Let *U*(*φ*) be a degree 2*d* polynomial in n variables $\phi\in\mathbb{R}^{\mathbf{n}}$. It has $\begin{pmatrix}2d+n\\2d\end{pmatrix}$ coefficients. Let H(*φ*) = $\nabla^2$*U*(*φ*) be the Hessian matrix evaluated at *φ*. It is well-known that *U*(·) is convex if and only if H(*φ*) is positive semidefinite for all *φ*. As each entry of H(*φ*) is a polynomial, ensuring the semidefiniteness of H(*φ*) is difficult. In fact, even testing whether a degree four polynomial is convex or not is strongly NP-hard [16]. Therefore, we appeal to the notion of SOS-convexity, which is a computationally tractable sufficient condition for convexity [17]. A similar approach for least squares fitting of a convex polynomial to a set of points was used in [18]. In particular, with $\xi\in\mathbb{R}^{\mathbf{n}}$ being a vector of variables, *p*(*φ*,*ξ*) = *ξ*′H(*φ*)*ξ* is a scalar polynomial of degree 2*d* in 2*n* variables (*φ*,*ξ*). Consider the vector of all monomials constructed by multiplying *ξj*'s with *φ*-monomials of degree up to *d* − 1

$$\mathbf{v}=\left(\xi_1,\ldots,\xi_n,\xi_1\phi_1,\ldots,\xi_n\phi_n^{(d-1)}\right). \quad (5)$$

The length of v is $\begin{pmatrix} d+n-1 \\ d-1 \end{pmatrix}\times n$.

The following theorem converts the problem of finding a convex polynomial to that of finding a positive semidefinite matrix.

**Theorem IV.1** *If there exists a matrix* $\mathbf{P}\succeq 0$ *such that* **v′Pv** *=* $p(\varphi,\xi)=\xi'\mathrm{H}(\varphi)\xi$*, then the polynomial* $U(\cdot)$ *is convex.*

*Proof:* Let **P** = **VΛV′** be the spectral decomposition of **P**. Consider the vector of monomials $\mathbf{w}=\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}'\mathbf{v}$. Clearly $p(\varphi,\xi)$ = w′w is now a polynomial in a sum-of-square form and thus nonnegative for all $\varphi,\xi$. It follows that $p(\varphi,\xi)=\xi'\mathrm{H}(\varphi)\xi\geq 0$ for all $\varphi,\xi$, which implies that the Hessian matrix $\mathbf{H}\psi\succeq\mathbf{0}$ for all $\varphi$. We conclude that $U(\cdot)$ is convex.

The condition in the above theorem is equivalent to

$$\boldsymbol{\xi}'\mathbf{H}(\boldsymbol{\phi})\boldsymbol{\xi}\quad \text{is}\quad \textbf{SOS}\quad \text{in}\quad (\boldsymbol{\phi},\boldsymbol{\xi}).$$

A tight convex polynomial underestimating the sample points ($\varphi^{(i)}$, $f^{(i)}=f(\varphi^{(i)})$), $i$ = 1,...,$K$) can thus be found by solving the following semidefinite program:

$$\begin{array}{ll} \underset{U(\cdot)}{min} & \sum_{i=1}^{K}\left[f^{(i)}-U\left(\boldsymbol{\phi}^{(i)}\right)\right] \\ \text{subject to} & f^{(i)}\geq U\left(\boldsymbol{\phi}^{(i)}\right) \\ & \boldsymbol{\xi}'H(\boldsymbol{\phi})\boldsymbol{\xi}\quad \text{is}\quad \textbf{SOS}\quad \text{in}\quad (\boldsymbol{\phi},\boldsymbol{\xi}), \end{array} \quad (6)$$

where $H(\varphi)$ is the Hessian of the polynomial $U(\cdot)$.

In the case of SSDU, $U(\cdot)$ is a polynomial in $n$ = 3 variables and degree 2$d$ = 4 with $\begin{pmatrix} 7 \\ 4 \end{pmatrix}=35$ coefficients:

$$\begin{aligned} U\left(\boldsymbol{\phi}\right)= & a_1+a_2\phi_1+a_3\phi_1^2+a_4\phi_1^3+a_5\phi_1^4+a_6\phi_2+a_7\phi_1\phi_2\\ & +a_8\phi_1^2\phi_2\\ & +a_9\phi_1^3\phi_2\\ & +a_{10}\phi_2^2+a_{11}\phi_1\phi_2^2\\ & +a_{12}\phi_1^2\phi_2^2\\ & +a_{13}\phi_2^3+a_{14}\phi_1\phi_2^3\\ & +a_{15}\phi_2^4+a_{16}\phi_3+a_{17}\phi_1\phi_3\\ & +a_{18}\phi_1^2\phi_3\\ & +a_{19}\phi_1^3\phi_3\\ & +a_{20}\phi_2\phi_3\\ & +a_{21}\phi_1\phi_2\phi_3\\ & +a_{22}\phi_1^2\phi_2\phi_3\\ & +a_{23}\phi_2^2\phi_3\\ & +a_{24}\phi_1\phi_2^2\phi_3\\ & +a_{25}\phi_2^3\phi_3\\ & +a_{26}\phi_3^2+a_{27}\phi_1\phi_3^2\\ & +a_{28}\phi_1^2\phi_3^2\\ & +a_{29}\phi_2\phi_3^2\\ & +a_{30}\phi_1\phi_2\phi_3^2\\ & +a_{31}\phi_2^2\phi_3^2\\ & +a_{32}\phi_3^3+a_{33}\phi_1\phi_3^3\\ & +a_{34}\phi_2\phi_3^3+a_{35}\phi_3^4. \end{aligned}$$

The vector of monomials has a length of $\begin{pmatrix} 4\\ 1 \end{pmatrix} \times 3=12$ elements:

$$\mathbf{v}=\left(\xi_1,\xi_2,\xi_3,\phi_1\xi_1,\phi_2\xi_1,\phi_3\xi_1,\phi_1\xi_2,\phi_2\xi_2,\phi_3\xi_2,\phi_1\xi_3,\phi_2\xi_3,\phi_3\xi_3\right).$$

The constraint $\xi' H(\varphi)\xi$ is SOS in $(\varphi,\xi)$ is equivalent to $\mathbf{P}\succeq 0$, where $\mathbf{v}'\mathbf{P}\mathbf{v} = p(\varphi,\xi) = \xi' H(\varphi)\xi$. By matching the coefficients, we can relate the elements of **P** with the coefficients of $U(\varphi)$. In this case there are 60 such equality constraints. We can now specialize the SDP in (6) as follows:

$$\begin{aligned} \underset{a_1,\ldots,a_{35},\mathbf{P}}{min} & \sum_{i=1}^{K} s^{(i)}\\ \text{s.t.} f^{(i)}- & \left(a_1+a_2\phi_1,\ldots,a_{35}\phi_3^4\right)=s^{(i)},\forall i=1,\ldots,K,\\ & P_{1,1}=12a_5, \qquad P_{4,4}=2a_{12},\ldots,\\ & 2P_{11,12}=2a_{20}, \qquad P_{12,12}=2a_{26},\\ & \mathbf{P}\succeq 0, \qquad s^{(i)}\geq 0, \quad \forall i=1,\ldots,K. \end{aligned} \tag{7}$$

We use the standard semidefinite solver CSDP [19] to solve the above SDP program. Note that unlike the quadratic underestimator, the minimum of the general convex polynomial cannot be computed in closed form; instead, it can be found using Newton's method.

### C. Sampling

Suppose the global minimum of the underestimator obtained in the previous section is $\varphi^P$. Even if the constructed convex underestimator reflects the general funnel structure of $f(\cdot)$ near the native conformation, locating the global minimum is still very difficult because van der Waals interactions create many local minima around the native conformation. Our strategy is to sample more conformations and explore the vicinity of $\varphi^P$. More specifically, let $\sigma_1 \geq ... \geq \sigma_5$ be the diagonal elements of SSS in (3); they are proportional to the variance along each principal direction. We generate a set of $\bar{K}$ random samples $\mathbf{s}^{(l)} \in \mathbb{R}^5,\ l=1,\dots,\bar{K}$, such that each dimension $s_j^{(l)}$ has independent uniform distribution in the range of $(-\frac{1}{2}\beta\sigma_j, \frac{1}{2}\beta\sigma_j)$, $j = 1,...,5$, for some constant $\beta$. Transforming from the principal coordinates to the original coordinates we obtain the new sample points

$$\tilde{\mathbf{x}}^{(l)}=\mathbf{W}\left(\mathbf{z}^P+\mathbf{s}^{(l)}\right)+\bar{\mathbf{x}},$$

where $\mathrm{z}^P$ is obtained by appending the prediction for $z_4^P$, $z_5^P$ to $\varphi^P$. Since the sample variance in the $z_4$,$z_5$ subspace is small, we can obtain a good approximation by taking the sample mean, i.e., $z_4^P=\frac{1}{K}\sum_{i=1}^K z_4^{(i)}$ and $z_5^P=\frac{1}{K}\sum_{i=1}^K z_5^{(i)}$. Finally, to get back to the 6-D conformational space, we append the mean center-to-center distance $r$ in (1) to $\tilde{\mathrm{x}}^{(l)}$ and generate the new sample conformation

$$\boldsymbol{\psi}^{(l)}=\left(\frac{1}{K}\sum_{i=1}^{K} r^{(i)},\tilde{\mathbf{x}}^{(l)}\right). \quad (8)$$

Making the sampling range of $\mathrm{s}^{(l)}$ proportional to the variance ensures good coverage of the conformational space and preserves the sample distribution.

### D. The SSDU algorithm

We summarize the SSDU in Algorithm 1.

## V. Numerical results for test functions

In this section we demonstrate that a higher degree convex polynomial is able to better capture more complex free energy-like functions. The test function we use is similar to that in [15] except the $f_1$ component is now a degree four polynomial:

$$f=f_1+f_2+f_3, \quad f_1=\sum_{i=1}^{3} a_i\phi_i^4, \quad f_2=-\sum_{i=1}^{3} c_i cos\left(b_i\phi_i\right), \quad f_3=A\sum_{i=1}^{3}\left[sin\left(\gamma\phi_i\right)\right].$$

The domain of the test function is 3-D just like the SSDU algorithm. We set the parameters to be **a** = (1,0.1,0.1), **b** = (2$\pi$,3$\pi$,4$\pi$), c = (1,2.5,2.5), $\gamma = 10\pi$. The parameter $A$ determines the amplitude of the high-frequency "noise." Notice the global minimum is at the origin. We sample 100 points uniformly in the cube [−7,2]×[−7,2] × [−7, 2], find a local minimum for each of these sample points, and use them as input to the underestimators. Figure 1 shows an example of such a test function along with degree two and degree four underestimators when $A$ = 30. For different $A$ values, we compare the performance of the degree four and degree two underestimators in Table I. We can see that the degree four underestimator can better capture the funnel structure of the test function in the presence of noise, hence, predicting minimum values that are closer to the true global minimum.

## VI. Protein docking refinement results

In this section we show that our proposed algorithm significantly improves the prediction quality of protein docking refinement. Our test set consists of 10 protein complexes that belong to the category of Others as listed in the first column of Table II. Protein complexes in this category exhibit complex energy surfaces and multiple deep funnels around the native structure and they are known to be difficult cases in protein docking refinement. Our protein docking procedure is applied to the (known) unbound receptor and ligand structures. The input to refinement is a low energy cluster of conformations obtained from PIPER as described in Section I. The number of input conformations to the refinement algorithms is shown in the second column of Table II, which is 1,000 for all complexes. Recall that we consider a conformation "accurate" if it is within 5 Å to the native conformation. The number of "accurate" conformations in the input is shown in the third column of Table II.

The refinement algorithms we compare in this paper are two existing methods: an MCM-based algorithm (MCM) [12] and SDU [3] against two proposed methods: SSDU with quadratic underestimator (SSDU2) and SSDU with degree four underestimator (SSDU4). As part of our implementation, we have added a local minimization subroutine to all algorithms in order to resolve the steric clashes at the protein interfaces and thus obtain more reliable energy evaluation. Specifically, for each sample conformation we first run a *rigid-body energy minimization* algorithm [1] which locally minimizes the position and orientation of the ligand with respect to the receptor. Then we run a *side-chain positioning (SCP)* algorithm [12] that solves a combinatorial optimization problem in order to re-position the amino acid residues at the interface of the receptor-ligand interaction. The SCP algorithm models the flexibility of the protein structures upon binding. Such local minimization steps have been shown to improve protein docking refinement.

For the energy function terms referenced thus far, we have used a state-of-the-art high-accuracy docking energy potential, which combines force-field and knowledge-based energy terms [20], [21], [22]. In particular, interaction energies are computed as a weighted sum ($w$'s are the corresponding weights):

$$E=w_{V\quad DW}E_{V\quad DW}+w_{SOL}E_{SOL}+w_{DARS}E_{DARS}+w_{COUL}E_{COUL}+w_{HB}E_{HB}+w_{RP}E_{RP},$$

where $E_{VDW}$ is the Lennard-Jones potential, $E_{SOL}$ is an implicit solvation term [23], $E_{DARS}$ is a statistical potential [24], $E_{HB}$ is a knowledge-based hydrogen bonding term, $E_{COUL}$ is the Coulomb potential, and $E_{RP}$ is a statistical energy term associated with a specific selection of rotamers from the backbone-dependent rotamer library [25].

In each iteration of SDU, SSDU2 and SSDU4 we re-sample 1,000 conformations around the underestimator minimum. In Table II we report the number of "accurate" conformations out of 1,000 samples for each algorithm upon convergence. In practice we notice SDU, SSDU2 and SSDU4 can generally converge after two iterations whereas MCM typically needs many more iterations. Such ability to quickly make many "accurate" predictions is a tremendous advantage of these stochastic global optimization methods. We can see in Table II that for most of the complexes SSDU2 performs the best among the four methods (The most number of "accurate" prediction for each complex is shaded blue). SSDU4 gives the best prediction quality for 3 complexes. For 1b6c, however, SSDU4 returns no "accurate" prediction. This indicates that using a higher degree polynomial underestimator (SSDU4) can potentially over-fit the sampled data and thus be more susceptible to outliers. No method improves the initial input quality for the second complex 1azs, which represents a case of having multiple funnel structures near the native conformation. We can also see that SSDU2 outperforms SDU in all these complexes, which confirms the advantage of dimensionality reduction. Figure 2 illustrates that SSDU2 can accurately capture the funnel structure by fitting an underestimator in the reduced subspace as compared to SDU.

## VII. Conclusions

We presented a new method for protein docking refinements with two main contributions. The first is dimensionality reduction; we have shown that better underestimators can be constructed to capture the energy landscape in the permissive subspace computed via PCA. The second contribution is to use higher degree SOS-convex polynomials instead of quadratic ones as underestimator; we pose this problem as solving a semi-definite program and demonstrate its advantage in test functions. Finally we show by experimenting on protein complexes that the proposed SSDU algorithms significantly improves refinement quality compared to previous methods.

## Acknowledgments

The authors would like to thank Hanieh Mirzaei and Artem Mamonov for many helpful discussions and Brian Borchers for help with the CSDP solver.

## REFERENCES

1. Mirzaei H, Beglov D, Paschalidis IC, Vajda S, Vakili P, Kozakov D. Rigid body energy minimization on manifolds for molecular docking. Journal of Chemical Theory and Computation. 2012; 8(11):4374–4380. [PubMed: 23382659]
2. Mirzaei, H.; Kozakov, D.; Villar, E.; Mottarella, S.; Beglov, D.; Vajda, S.; Paschalidis, IC.; Vakili, P. Flexible refinement on manifolds for molecular docking. Proceedings of the 52nd IEEE Conference on Decision and Control; Florence, Italy. December 2013; p. 1392-1397.
3. Paschalidis IC, Shen Y, Vakili P, Vajda S. SDU: A semi-definite programming-based underestimation method for stochastic global optimization in protein docking. IEEE Trans. Automat. Contr. 2007; 52(4):664–676. [PubMed: 19759849]

4. Shen Y, Paschalidis IC, Vakili P, Vajda S. Protein Docking by the Underestimation of Free Energy Funnels in the Space of Encounter Complexes. PLoS Computational Biology. 2008; 4(10)

5. Kozakov D, Li K, Hall D, Beglov D, Zheng J, Vakili P, Schueler-Furman O, Paschalidis IC, Clore GM, Vajda S. Encounter complexes and dimensionality reduction in protein-protein association. eLIFE. 2014; 3:e01370. [PubMed: 24714491]

6. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: An FFT-based protein docking program with pairwise potentials. Proteins. 2006; 65:392–406. [PubMed: 16933295]

7. Camacho C, Gatchell D, Kimura S, Vajda S. Scoring docked conformations generated by rigid-body protein-protein docking. J. Proteins: Struct. Funct. Genet. 2000; 40:525–537.

8. Norel R, Sheinerman F, Petre D, Honig B. Electrostatic contributions to protein protein interactions: fast energetic filters for docking and their physical basis. J. Protein Sci. 2001; 10:2147–2161.

9. Palma P, Krippahl L, Wampler J, Moura J. Bigger: a new (soft) docking algorithm for predicting protein interactions. J. Proteins: Struct. Funct. Genet. 2000; 39:372–384.

10. Gray J, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl C, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Molecular Biol. 2003; 331:281–299.

11. Moghadasi, M.; Kozakov, D.; Mamonov, A.; Vakili, P.; Vajda, S.; Paschalidis, IC. A message passing approach to side chain positioning with applications in protein docking refinement. Proceedings of the 51st IEEE Conference on Decision and Control; Maui, Hawaii. December 2012; p. 2310-2315.

12. Moghadasi, M.; Kozakov, D.; Vakili, P.; Vajda, S.; Paschalidis, IC. A new distributed algorithm for side-chain positioning in the process of protein docking. Proceedings of the 52nd IEEE Conference on Decision and Control; Florence, Italy. December 2013; p. 739-744.

13. Camacho C, Vajda S. Protein docking along smooth association pathways. Proc. Natl. Acad. Sci. USA. 2001; 98:10, 636–10, 641.

14. Phillips, A.; Rosen, J.; Dill, K. From Local to Global Optimization. In: Pardalos, PM., et al., editors. ch. Convex Global Underestimation for Molecular Structure Prediction. Kluwer Academic Publishers; 2001. p. 1-18.

15. Dennis S, Vajda S. Semi-global simplex optimization and its application to determining the preferred solvation sites of proteins. J. Comp. Chem. 2002; 23(3):319–334. [PubMed: 11908495]

16. Ahmadi AA, Olshevsky A, Parrilo PA, Tsitsiklis JN. NP-hardness of deciding convexity of quartic polynomials and related problems. CoRR. 2010 abs/1012.1908.

17. Ahmadi AA, Parrilo PA. A complete characterization of the gap between convexity and SOS-convexity. SIAM Journal on Optimization. 2013; 23(2):811–833.

18. Magnani A, Lall S, Boyd S. Tractable fitting with convex polynomials via sum-of-squares. Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on. 2005:1672–1677.

19. Borchers B. CSDP, a C library for semidefinite programming. Optimization Methods and Software. 1999; 11(1-4):613–623. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10556789908805765.

20. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Molecular Biology. 2003; 331:281–299.

21. Andrusier N, Nussinov R, Wolfson HJ. Firedock: Fast interaction refinement in molecular docking. Proteins: Structure, Function, and Bioinformatics. 2007; 69(1):139–159. [Online]. Available: http://dx.doi.org/10.1002/prot.21495.

22. Pierce B, Weng Z. Zrank: Reranking protein docking predictions with an optimized energy function. Proteins: Structure, Function, and Bioinformatics. 2007; 67(4):1078–1086. [Online]. Available: http://dx.doi.org/10.1002/prot.21373.

23. Schaefer M, Karplus M. A comprehensive analytical treatment of continuum electrostatics. J Phys Chem. 1996; 100:1578–1599.

24. Chuang G-Y, Kozakov D, Brenke R, Comeau SR, Vajda S. Dars (decoys as the reference state) potentials for protein-protein docking. Biophysical journal. 2008; 95(9):4217–27. [PubMed: 18676649]

25. Shapovalov M, Dunbrack Jr R. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure. 2011; 19(6):844–858. [PubMed: 21645855]

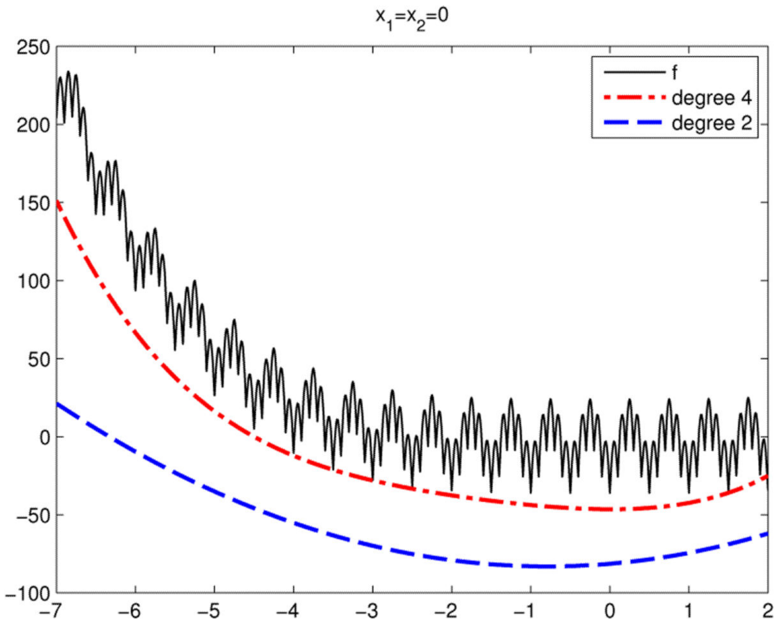

**Fig. 1.**

A slice at the $x_1 = x_2 = 0$ plane of the test function and underestimators of degree four and two. The degree four underestimator approximates the test function more closely and predicts a global minimum value closer to the true value, which is the origin.

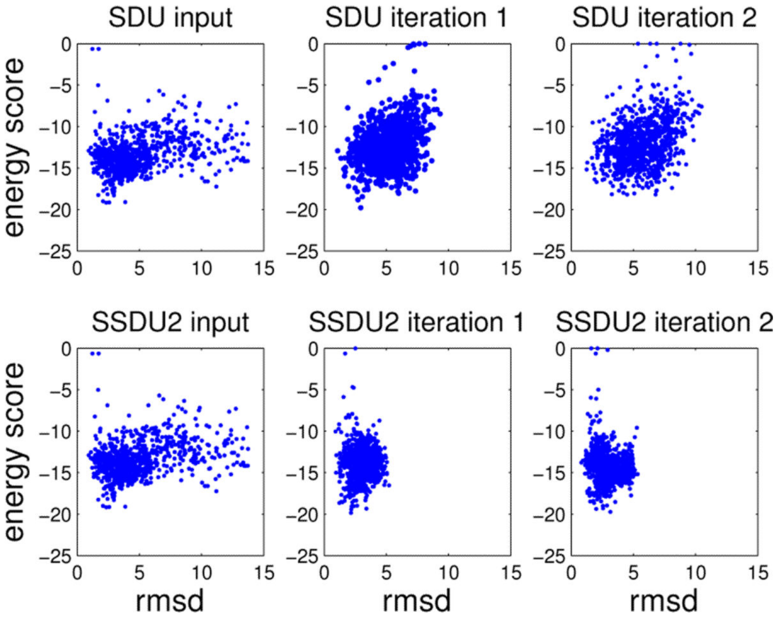

**Fig. 2.**

Refinement comparison of SDU and SSDU2 for 3d5s. The upper three plots show the evolution of SDU for 2 iterations. The bottom three plots show the evolution of SSDU2 for 2 iterations. Each point in the figure is a conformation plotted by its RMSD to the native on the X-axis and its energy value on the Y-axis. The input conformations are the same for both methods. SSDU2 gives more “accurate” predictions as measured by RMSD.

### TABLE I

Distances from the true global minimum to the predicted minimum by degree four and degree two underestimators at different noise levels. Averaged over 20 runs. The degree 4 underestimator predicts a minimum closer to the true value.

| | ***A* = 10** | ***A* = 30** | ***A* = 50** |
|---|---|---|---|
| Degree 4 | 1.1628 | 1.4381 | 1.4760 |
| Degree 2 | 1.6526 | 1.9589 | 2.6754 |

## TABLE II

Refinement results for 10 systems with the number of accurate conformations out of the output samples for four different methods. For each complex the best performing method is shaded in blue.

| Complex | Total | Initial | MCM | SDU | SSDU2 | SSDU4 |
|---|---|---|---|---|---|---|
| 1akj | 1000 | 184 | 148 | 92 | 185 | 170 |
| 1azs | 1000 | 721 | 570 | 176 | 510 | 270 |
| 1b6c | 1000 | 435 | 511 | 81 | 655 | 0 |
| lgrn | 1000 | 234 | 170 | 90 | 558 | 184 |
| lxqs | 1000 | 149 | 101 | 66 | 222 | 52 |
| 1e96 | 1000 | 241 | 321 | 104 | 386 | 41 |
| lsyx | 1000 | 215 | 174 | 250 | 292 | 465 |
| 1xd3 | 1000 | 335 | 316 | 53 | 383 | 419 |
| 2cfh | 1000 | 515 | 584 | 46 | 596 | 795 |
| 3d5s | 1000 | 666 | 741 | 365 | 993 | 268 |

## Algorithm 1

SSDU Algorithm

```
1: **Initialization:** Starting from $\bar{K}$ sample points perform local minimization to obtain $K$ distinct local minima $\psi^{(1)},\ldots,\psi^{(K)}$ of $f(\cdot)$.
2: **Dimensionality Reduction:** For each sample point $i$ reduce $\psi^{(i)}$ to $\mathbf{x}^{(i)}$ in (2) then transform to $\varphi^{(i)}$ in (4) using PCA.
3: **Underestimation:** Solve the SDP in (7) to obtain the convex polynomial underestimator $U(\varphi)$. Set the predictive point $\varphi^P$ to be the minimizer of $U(\varphi)$. Transform $\varphi^P$ to $\psi^P$ in the 6-dimensional conformational space.
4: **Exploration:** Generate random samples $\psi^{(l)}$; $l = 1,\ldots,\bar{K}$, in (8) based on the predictive point $\varphi^P$. Perform local minimization starting from $\psi^{(l)}$; $i = l,\ldots,\bar{K}$, to obtain a set of distinct local minima $\mathrm{L} = \{\hat{\psi}^{(1)}, \cdots, \hat{\psi}^{(K)}\}$ and let $\psi^G = \operatorname{argmin}_{\hat{\psi}\in\mathrm{L}} f(\hat{\psi})$.
5: **Termination:** If $\|\psi^G - \psi^P\| < \varepsilon$ or if there is no progress in reducing $f(\psi)$ over the last few iterations then stop; otherwise go to Step 2.
```