

# Extracting knowledge-based energy functions from protein structures by error rate minimization: Comparison of methods using lattice model

Yu Xia<sup>a)</sup> and Michael Levitt

Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305

(Received 30 June 2000; accepted 1 September 2000)

We describe a general framework for extracting knowledge-based energy function from a set of native protein structures. In this scheme, the energy function is optimal when there is least chance that a random structure has a lower energy than the corresponding native structure. We first show that subject to certain approximations, most current database-derived energy functions fall within this framework, including mean-field potentials, Z-score optimization, and constraint satisfaction methods. We then propose a simple method for energy function parametrization derived from our analysis. We go on to compare our method to other methods using a simple lattice model in the context of three different energy function scenarios. We show that our method, which is based on the most stringent criteria, performs best in all cases. The power and limitations of each method for deriving knowledge-based energy function is examined. © 2000 American Institute of Physics. [S0021-9606(00)51844-3]

## I. INTRODUCTION

Energy functions empirically extracted from a database of known native protein three-dimensional structures are referred to as “knowledge-based.” Due to the rapid increase in the number of known protein structures and the limited success of physical potentials to discriminate native structure from misfolded structures (also known as decoys), knowledge-based energy functions have been widely used in fold recognition, comparative modeling, and *ab initio* structure prediction.<sup>1</sup> Many different methods have been proposed for extracting knowledge-based energy function based on very different principles. They can be roughly grouped into three categories.

The first category involves methods based on statistical mechanics or Bayesian statistics. First proposed by Tanaka and Scheraga<sup>2</sup> and later refined by Miyazawa and Jernigan<sup>3,4</sup> and Sippl,<sup>5,6</sup> these methods, based on statistical mechanics, all rely on Boltzmann statistics:

$$e_j = -kT \ln \frac{N_j^n}{N_j^{\text{ref}}}, \quad (1)$$

where  $e_j$  is the energy parameter for  $j$ th-type interaction,  $N_j^n$  is the total number of  $j$ th-type interactions observed in the database of native structures, and  $N_j^{\text{ref}}$  is the total number of  $j$ th-type interactions expected in the reference state. Existing methods differ mainly in their choices of reference states.<sup>7</sup> An alternative approach based on Bayesian statistics using a simple log-odds score also yields the same formalism.<sup>8,9</sup>

In spite of its simplicity and widespread use, several problems are associated with this approach.

- (a) The physical nature of the reference state is obscure, making it difficult to choose a reference state properly.

- (b) It is not clear why energy function parameters derived from Boltzmann statistics should give the best performance in threading tests and *ab initio* prediction. Even though supported by the Random Energy Model,<sup>10</sup> Boltzmann statistics has been shown to introduce systematic errors in recovering energy function parameters.<sup>11</sup>

The second category involves methods based on direct optimization of some measurement of the performance of the energy function. Very different optimization schemes have been proposed: maximizing  $T_f/T_g$ , the ratio of the folding temperature to the glass transition temperature,<sup>12</sup> maximizing the average probability of successful prediction,<sup>13</sup> minimizing the free energy of the native state,<sup>14</sup> etc. Many of these methods reduce to the optimization of the Z-score, the energy difference between average decoy structure and native structure in units of standard deviation.<sup>12,13,15</sup>

The third category, proposed by Crippen,<sup>16–18</sup> involves constraint satisfaction. A library of incorrect structures is explicitly constructed and energy function parameters are tuned to satisfy all the inequalities that ensure the native structure has lower energy than any of the incorrect decoy structures.

*A priori*, it would appear that energy parameters derived from less well-justified Boltzmann statistics should perform significantly worse in threading and *ab initio* tests relative to parameters derived from more sound optimization and constraint satisfaction methods. Surprisingly, the performances are quite similar to each other. It is intriguing that methods based on very different formalisms generate energy parameters with similar performances.

In this work we show that all three categories of methods can indeed be derived as approximations in a unified theoretical framework of minimizing the error rate, i.e., the probability that a random structure has a lower energy than the corresponding native structure. We describe a simple imple-

<sup>a)</sup> Author to whom correspondence should be addressed; electronic mail: yuxia@csb.stanford.edu

mentation of error rate minimization, and compare our method with two other popular methods, Z-score optimization and mean-field approximation using a simple lattice model in the context of three different energy function scenarios.

## II. A UNIFIED FRAMEWORK FOR KNOWLEDGE-BASED ENERGY FUNCTION

In this section we present a unified framework for deriving knowledge-based energy functions. First we consider the problem of extracting the energy function from a single native structure, and show that all three above-mentioned approaches can be obtained from the same assumption under different approximations. Then we extend the result to include multiple native structures. Some details of the derivation are presented in the Appendix.

### A. Problem specification for a single protein

Given a protein with known native structure and a library of random and uniformly sampled decoy structures, we start from the assumption that the native structure has the lowest energy compared with any other structure:

$$E^n < E_i, \quad \text{for all } i = 1, \dots, M, \quad (2)$$

where  $E^n$  is the energy of native structure,  $E_i$  is the energy of the  $i$ th decoy, and  $M$  is the number of decoys. This is a natural assumption in that the main use of the energy functions we are seeking is to distinguish the native state from the decoys. We decompose the energy into individual interaction components in the following way:

$$E_i = \sum_{j=1}^N c_{ij} e_j = \mathbf{c}_i \cdot \mathbf{e}, \quad \text{for all } i = 1, \dots, M, \quad (3)$$

where  $N$  is the number of interaction types,  $c_{ij}$  is the number of occurrences of the  $j$ th interaction type in the  $i$ th decoy, and  $e_j$  is the interaction energy for the  $j$ th interaction type. In this formulation, the energy is assumed to be linear with respect to the interaction counts. Many existing energy functions can be defined in this way, including contact potentials, distance-dependent potentials, and many-body interactions.

The energy for the native structure is also a linear function of native interaction counts:

$$E^n = \sum_{j=1}^N c_j^n e_j = \mathbf{c}^n \cdot \mathbf{e}. \quad (4)$$

$E$  is usually interpreted as the solvent-averaged effective potential energy of the protein. It is well justified to decompose potential energy into individual interactions; however, Eq. (2) becomes a *minimal* requirement for a perfect energy function, because the native state must not only be the global minimum in the potential energy surface, it must also be a *pronounced* minimum, with significantly lower energy than any other state. Here our hope is that with a large number of known protein structures, even a minimal requirement such as Eq. (2) is restrictive enough to recover the “true” potential energy parameters with precision. In the next section we will discuss when this hypothesis breaks down and how this can be possibly remedied.

When the number of decoys,  $M$ , is small enough that a solution  $\mathbf{e}$  exists to satisfy all the inequalities in Eq. (2), this approach reduces to the constraint satisfaction method.<sup>16</sup> However, since the energy function formulation in Eqs. (3) and (4) is only an approximation to the true potential, the inequalities in Eq. (2) may not be all satisfied when  $M$  is very large, at least in the case of the popular residue-residue contact potential.<sup>19</sup> In order to get the best approximation to the true potential, we can simply minimize the error rate, i.e., the percentage of inequalities in Eq. (2) that are not satisfied. In this way, we define energy function parametrization as the optimization of the following error rate function,  $R(\mathbf{e})$ :

$$R(\mathbf{e}) = \langle \theta(E^n - E_i) \rangle_i, \quad (5)$$

where  $\theta(x)$ , the Heaviside step function, is 1 when  $x$  is positive, and 0 when  $x$  is negative.  $\langle \dots \rangle_i$  denotes averaging over subscript  $i$ .

Intuitively,  $R(\mathbf{e})$  is the probability that a randomly selected decoy in the decoy set has lower energy than the native structure. A value of 0 means that the native structure has the lowest energy compared with decoy structures; a value of 1 means it has the highest energy. Therefore, energy parameters that minimize  $R(\mathbf{e})$  should give the best discriminating performance. Let us denote the optimal energy parameter set as  $\mathbf{e}^0$ . In what follows we derive an approximate analytical solution for  $\mathbf{e}^0$  and augment it with a geometrical interpretation.

### B. General solution and geometrical interpretation

We first explore the geometrical meaning of the  $R(\mathbf{e})$  function. Equation (5) can be rewritten as

$$R(\mathbf{e}) = \int_{(\mathbf{c} - \mathbf{c}^n) \cdot \mathbf{e} < 0} p_c(\mathbf{c}) d\mathbf{c}, \quad (6)$$

where  $p_c(\mathbf{c})$  is the joint probability density function of the random interaction count vector  $\mathbf{c}$ . We see that  $R(\mathbf{e})$  is a partial integral of  $p_c(\mathbf{c})$  with a boundary of an  $N - 1$ -dimensional hyperplane that goes through the point  $\mathbf{c}^n$ . The normal to the hyperplane that minimizes the integral determines the optimal energy parameter  $\mathbf{e}^0$  (see Fig. 1). Under certain approximations, a simple analytical solution can be derived for the optimal energy parameter  $\mathbf{e}^0$ :

$$\mathbf{e}^0 = \vec{\nabla} p_c(\mathbf{c}^n). \quad (7)$$

Details of the derivation are presented in the Appendix (Sec. 1). The geometrical interpretation of this solution is intuitive in that  $\mathbf{e}^0$  is perpendicular to the contour of interaction count distribution through point  $\mathbf{c}^n$ . It is clear from this equation that the optimal energy parameters  $\mathbf{e}^0$  depend on the decoy interaction count distributions  $p_c(\mathbf{c})$ , especially near the interaction count vector for the native structure  $\mathbf{c}^n$ . Furthermore, by assuming that  $p_c(\mathbf{c})$  follows an independent normal distribution or an independent Poisson distribution, the above solution leads to Z-score optimization and mean-field statistics, respectively. Details of the derivation are presented in the Appendix (Secs. 2 and 3) and illustrated in Fig. 2.

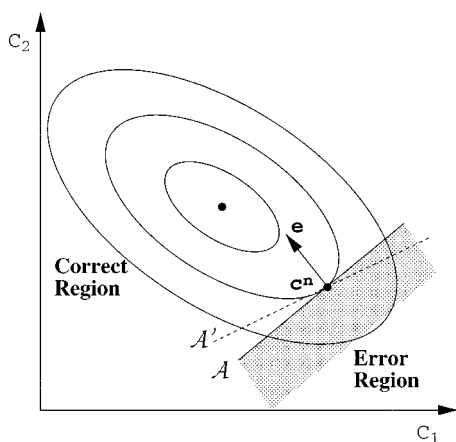


FIG. 1. Geometrical interpretation of the energy function. This is a 2D representation of the high-dimensional space of interaction counts. Each point in the space represents structures with particular interaction counts. Concentric ellipsoids represent the contour map of the interaction count distribution.  $c_i$  is the count for the  $i$ th interaction type. Point  $\mathbf{c}^n$  is the native structure. Energy function parameters are represented by a vector  $\mathbf{e}$ .  $\mathcal{A}$  is the hyperplane that goes through  $\mathbf{c}^n$  with  $\mathbf{e}$  as its normal vector. The region above the hyperplane  $\mathcal{A}$  is called ‘‘Correct Region’’ because points (structures) within this region have higher energy than the native structure. The region below  $\mathcal{A}$ , shown shaded, is called ‘‘Error Region’’ because points (structures) within this region have lower energy than the native structure.  $\mathcal{A}'$  corresponds to a hyperplane defining an alternative set of energy parameters.

### C. Extension to multiple proteins

The above discussion can be extended to multiple proteins, where the native structures for many sequences are assumed to be known. The optimal energy function parameter  $\mathbf{e}^0$  should minimize the average prediction error rate function  $R(\mathbf{e})$ :

$$R(\mathbf{e}) = \langle \theta(E_k^n - E_{ki}) \rangle_{i,k}, \quad (8)$$

where  $E_k^n$  is the energy for the native structure of the  $k$ th protein,  $E_{ki}$  is the energy for the  $i$ th decoy of the  $k$ th protein,

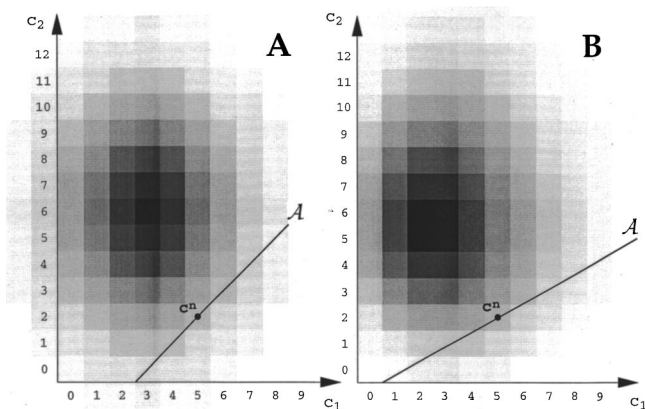


FIG. 2. Comparison between (A) the normal distribution and (B) the Poisson distribution. Each point is shaded according to the density of the interaction count distribution.  $\mathbf{c}^n$  is the native structure, and hyperplane  $\mathcal{A}$  corresponds to the optimal energy parameters. Even though the two cases share the same native structure and the same average interaction counts, the optimal energy parameters, indicated by the hyperplane  $\mathcal{A}$ , are different due to the different interaction count distributions.

and  $\langle \dots \rangle_{i,k}$  denotes averaging over subscripts  $i$  and  $k$  (decoys and proteins). We can rewrite this equation in terms of interaction counts:

$$R(\mathbf{e}) = \langle \theta((\mathbf{c}_k^n - \mathbf{c}_{ki}) \cdot \mathbf{e}) \rangle_{i,k}. \quad (9)$$

This equation can also be interpreted geometrically. Let us define the relative interaction count  $\mathbf{c}'_{ki}$  as

$$\mathbf{c}'_{kij} = c_{kij} - c_{kj}^n \quad \text{for all } j = 1, \dots, N, \quad (10)$$

where for each interaction type  $j$ ,  $c_{kij}$  is the interaction count in the  $i$ th decoy of the  $k$ th protein, and  $c_{kj}^n$  is the interaction count in the native structure of the  $k$ th protein. We then have

$$R(\mathbf{e}) = \int_{\mathbf{c}' \cdot \mathbf{e} < 0} p_{\mathbf{c}'}(\mathbf{c}') d\mathbf{c}', \quad (11)$$

where  $p_{\mathbf{c}'}(\mathbf{c}')$  is the joint probability density function of relative interaction count for randomly chosen protein sequence and decoy. We see that  $R(\mathbf{e})$  is a partial integral of  $p_{\mathbf{c}'}(\mathbf{c}')$  with a boundary of  $N-1$ -dimensional hyperplane that goes through the origin. The normal to the hyperplane that minimizes the integral determines the optimal energy parameter  $\mathbf{e}^0$ .

Similar to the single protein case, under different approximations on the decoy interaction count distribution  $p_{\mathbf{c}'}(\mathbf{c}')$ , the above optimization problem leads to both Z-score optimization and mean-field statistics. Details of the derivation are presented in the Appendix (Sec. 4).

### III. DERIVING ENERGY PARAMETERS FROM PROTEIN FOLDS

In the previous section, we derived general principles for knowledge-based energy function extraction. In this section, we describe a simple implementation of our error rate minimization. We also describe our implementation of two other popular methods, Z-score optimization and mean-field approach.

#### A. Method I: Minimize the error rate function

Our goal is to minimize the error rate function  $R(\mathbf{e})$ , the probability that a random decoy of a randomly picked protein has a lower energy than the corresponding native structure, without any assumption on the interaction count distribution. Our procedure is a variation of the perceptron learning algorithm known as the pocket algorithm.<sup>20</sup> We start with a set of normalized random energy parameters  $\mathbf{e}$  with zero mean and iteratively update them [since adding or multiplying a constant to all energy parameters  $e_j$  does not change the value of  $R(\mathbf{e})$  in our lattice study, we can normalize all energy parameters to have zero mean and unit norm]. At each step we randomly choose a protein sequence from the training set and its corresponding native structure,  $\mathbf{c}^n$ , as well as a random alternative structure,  $\mathbf{c}$ . We update the energy parameters in the following way:

$$\mathbf{e}^{\text{new}} = \mathbf{e} + \lambda(\mathbf{c} - \mathbf{c}^n) \theta(E^n - E), \quad (12)$$

where  $\theta(x)$  is the Heaviside step function. Namely, we keep the old energy parameters when the energy for the native structure is lower than the alternative structure, and only up-

date the energy parameters when this constraint is violated. The update to the energy parameters is chosen in such a way as to correct this violation. The degree of correction is controlled by a tunable parameter  $\lambda$ , which linearly decreases from  $\lambda_0$  to 0 as optimization proceeds ( $\lambda_0$  is set to 0.1 in our study). The new energy parameters are again normalized and the procedure continues.

We retain energy parameters that have survived unchanged for the longest number of steps during the optimization run (the pocket algorithm). It can be shown that, for a sufficiently long training time, this gives, with probability arbitrarily close to unity, the set of energy parameters which produces the smallest possible number of misclassifications.<sup>20</sup>

This is the optimization procedure that we use when there is no large energy gap between the native structures and the alternative structures (see energy function scenarios I and II described in the next section). However, for the folding potential (energy function scenario III described in the next section), the energy gap between the native structure and alternative structures is large enough that the simple constraint that the native structure has lower energy than alternative structures is no longer sufficiently restrictive. As we will see in the results section, the above optimization is unstable and the results depend strongly on the initial conditions. To solve this problem we apply a more restrictive constraint,

$$E > E^n + \delta, \quad (13)$$

where  $\delta$ , a tunable nonnegative number, represent the energy gap between the native structures and all other alternative structures. Since all energy parameters  $\mathbf{e}$  are normalized,  $\delta$  is a unitless constant. A large energy gap  $\delta$  corresponds to a pronounced folding funnel as well as fast folding kinetics.<sup>21</sup> We minimize the error rate for the above constraints using the same learning procedure as above, with a new updating rule,

$$\mathbf{e}^{\text{new}} = \mathbf{e} + \lambda(\mathbf{c} - \mathbf{c}^n - \delta\mathbf{e})\theta(E^n + \delta - E). \quad (14)$$

Again the new energy parameters are normalized. We show how to determine the best value for  $\delta$  in the results section.

## B. Method II: Optimize the average Z-score

We define the average Z-score as the arithmetic average over different proteins of the individual Z-scores for the corresponding native structures,

$$\bar{Z} = \frac{1}{P} \sum_{k=1}^P \frac{\langle E_{ki} \rangle_i - E_k^n}{\sqrt{\langle E_{ki}^2 \rangle_i - \langle E_{ki} \rangle_i^2}}, \quad (15)$$

where for protein  $k$ ,  $E_k^n$  is the energy for the native structure,  $E_{ki}$  is the energy for the  $i$ th decoy, and  $\langle \dots \rangle_i$  denotes averaging over subscript  $i$ . Mirny and Shakhnovich pointed out<sup>15</sup> that the above definition can be rewritten in the following way:

$$\bar{Z} = \frac{1}{P} \sum_{k=1}^P \frac{\sum_{j=1}^N a_{kj} e_j}{\sqrt{\sum_{j=1}^N \sum_{j'=1}^N e_j \mathbf{B}_{kjj'} e_{j'}}}, \quad (16)$$

where vector  $\mathbf{a}_k$  and covariance matrix  $\mathbf{B}_k$  are defined in terms of relative interaction counts  $c'_{kij}$  from Eq. (10):

$$a_{kj} = \langle c'_{kij} \rangle_i \quad \text{for all } k=1, \dots, P, \quad j=1, \dots, N, \quad (17)$$

$$\mathbf{B}_{kjj'} = \langle c'_{kij} c'_{kij'} \rangle_i \quad \text{for all } k=1, \dots, P, \quad j, j'=1, \dots, N. \quad (18)$$

We further adopt the assumption of Mirny and Shakhnovich that the covariance matrix  $\mathbf{B}_k$  only depends on protein size. In our lattice model study, all lattice proteins have the same size. As a result, we can replace all  $\mathbf{a}_k$  and  $\mathbf{B}_k$  by their average  $\bar{\mathbf{a}}$  and  $\bar{\mathbf{B}}$ , and rewrite  $\bar{Z}$  as

$$\bar{Z} = \frac{\sum_{j=1}^N \bar{a}_j e_j}{\sqrt{\sum_{j=1}^N \sum_{j'=1}^N e_j \bar{\mathbf{B}}_{jj'} e_{j'}}}. \quad (19)$$

We evaluate  $\bar{\mathbf{a}}$  and  $\bar{\mathbf{B}}$  by explicitly constructing alternative structures for all lattice proteins.

To obtain optimal energy parameters  $\mathbf{e}^0$  one would normally perform numerical optimization on the above equation. There is, however, an analytical solution that optimizes  $\bar{Z}$ . We first rewrite the above equation in matrix form,

$$\bar{Z} = \frac{\mathbf{e} \bar{\mathbf{a}}^T}{\sqrt{\mathbf{e} \bar{\mathbf{B}} \mathbf{e}^T}}, \quad (20)$$

where  $\bar{\mathbf{a}}$  and  $\mathbf{e}$  are row vectors, and  $\mathbf{e}^T$  is the transpose of  $\mathbf{e}$ .  $\bar{\mathbf{B}}$  is a real symmetric covariance matrix, and is therefore diagonalizable with

$$\bar{\mathbf{B}} = \mathbf{V} \mathbf{D} \mathbf{V}^T, \quad (21)$$

where  $\mathbf{D}$  is a diagonal matrix that contains all the eigenvalues  $0 = D_{11} < D_{22} \leq \dots \leq D_{NN}$ , and  $\mathbf{V}$  is an orthogonal matrix that satisfies  $\mathbf{V} \mathbf{V}^T = \mathbf{V}^T \mathbf{V} = \mathbf{I}$ , where  $\mathbf{I}$  is the unitary matrix. All the eigenvalues are nonnegative since the average covariance matrix  $\bar{\mathbf{B}}$  is positive semidefinite. In our lattice model study, all the alternative lattice structures have the same total number of contacts as the native structure. As a result, adding a constant energy term to each of the contact energy parameters will not change the value of  $\bar{Z}$ .  $\bar{\mathbf{B}}$  is therefore not full rank, and the smallest eigenvalue  $D_{11}$  is 0.

Note that  $\mathbf{e} \bar{\mathbf{a}}^T = \mathbf{e} (\mathbf{V} \mathbf{V}^T) \bar{\mathbf{a}}^T$ , and  $\mathbf{e} \bar{\mathbf{B}} \mathbf{e}^T = \mathbf{e} (\mathbf{V} \mathbf{D} \mathbf{V}^T) \mathbf{e}^T = (\mathbf{e} \mathbf{V}) \mathbf{D} (\mathbf{e} \mathbf{V})^T$ . If we combine the variables in the following ways:  $\bar{\mathbf{a}}' = \bar{\mathbf{a}} \mathbf{V}$ , and  $\mathbf{e}' = \mathbf{e} \mathbf{V}$ , we can rewrite  $\bar{Z}$  as follows:

$$\bar{Z} = \frac{\mathbf{e}' \bar{\mathbf{a}}'^T}{\sqrt{\mathbf{e}' \mathbf{D} \mathbf{e}'^T}}. \quad (22)$$

Since  $\mathbf{D}$  is a diagonal matrix, this equation is similar to Eq. (A9) in the Appendix (Section 2) and has the following solution  $\mathbf{e}'^0$  that optimizes  $\bar{Z}$ :

$$e_j'^0 = k \frac{\bar{a}_j'}{D_{jj}}, \quad \text{for all } j=2, \dots, N, \quad (23)$$

where  $k$  is an arbitrary positive constant.  $e_1'^0$  is an arbitrary number since  $D_{11}$  is 0, and we set  $e_1'^0$  to be 0. This is equivalent to setting the average of all the energy parameters  $e_j$  to be 0.

Finally we obtain the optimal energy parameters,

$$\mathbf{e}^0 = \mathbf{e}' \circ \mathbf{V}^T. \quad (24)$$

For proteins with different sizes, this solution is no longer exact but still serves as a good first-order approximation and a reliable starting point for further numerical optimization.

### C. Method III. Derive mean-field, Bayesian statistics

In the mean-field/Bayesian approach, the energy parameters  $\mathbf{e}^0$  is estimated in the following way:

$$e_j^0 = -k \ln \frac{\sum_{k=1}^P c_{kj}^n}{\sum_{k=1}^P \langle c_{ki} \rangle_i}. \quad (25)$$

In this formulation the reference state is defined explicitly by sampling alternative lattice structures. Sometimes the numerator is 0 due to insufficient sequence sampling. In this case the numerator is set to be 1/2.

## IV. A TEST CASE

In this section, we describe a simple lattice model for protein structures. We identify three scenarios based on the accuracy of the energy function formulation, and for each describe how we generate a database of native protein sequences, native structures, and alternative structures (decoys). This database of model proteins is then used to evaluate the performance of the three methods for deriving energy parameters.

### A. Sets of lattice folds for training and testing

We use a simple  $6 \times 6$  two-dimensional square lattice model to construct test proteins. Protein conformations are represented by self-avoiding compact walks that occupy all lattice vertices. There are a total of 57 337 such conformations that are not symmetry related. Each lattice vertex represents a protein residue and is labeled by one of the 20 amino acids. As a result, all our model proteins are 36 residues in size. We set the energy function formulation to be a pairwise residue–residue contact energy function: the total energy of a conformation is calculated as the product of energy parameter for each residue–residue contact type and the number of occurrences for that contact type in the given conformation, summed over all contact types:

$$E = \sum_{j=1}^N e_j c_j. \quad (26)$$

Two residues are said to be in contact if (i) they are non-bonded, and (ii) they occupy adjacent vertices in the lattice. The total number of energy parameters  $e_j$  that need to be determined is 210 (the number of unique residue pairs ALA–ALA, ALA–CYS, etc.).

To test how effective different methods extract knowledge-based potentials, we first choose a set of arbitrary energy parameters  $e_j$  as the reference potential. In this work we choose the mean-field residue–residue contact energy function derived from real proteins by Hinds and Levitt<sup>22</sup> as the reference potential. We construct a database of protein sequences and the corresponding native structures according to the reference potential, then use different methods to extract knowledge-based potentials from this database of model

proteins. We compare the performance of each method by evaluating how accurately the extracted potential can recover the reference potential.

In our study, we generate 20 000 protein sequences and their corresponding native structures as the training set. Knowledge-based energy function parameters are extracted from this set. We then generate another 5000 protein sequences and their corresponding native structures using different starting random seeds as the test set to evaluate the statistical reliability of the extracted energy parameters.

### B. Three different energy function scenarios

How do we generate proper protein sequences and their corresponding native structures for the reference potential? This question is closely related to how accurate the energy function formulation mimics the true physical potential formulation. Since we do not know the exact forces that are responsible for protein folding, energy function formulations with different accuracy have been used traditionally. This will affect the accuracy of the extracted knowledge-based energy parameters. We would like to take this factor into consideration in our lattice model study. We identify three scenarios based on the accuracy of energy function formulation.

In scenario I, the energy function formulation is a rough approximation to the true potential. The reference energy parameters assign a low energy to the native structure, but not necessarily the lowest. This energy function is useful as a purification potential for protein structure prediction methods. It is capable of selecting a small subset from a large ensemble of alternative structures with higher concentration of near-native structures. Residue–residue contact potential is an example of such formulation for real proteins.

In scenario II, the energy function formulation mimics the true potential more accurately. The native structure is at the global energy minimum using the reference energy parameters. This energy function is an example of a perfect discriminatory potential for protein structure prediction methods. It is capable of discriminating the native structure from all other alternative structures.

In scenario III, the energy function formulation is the closest to the true potential. The native structure is at a *pronounced* global energy minimum using the reference energy parameters. This energy function is a folding potential and should be able to fold proteins in dynamics simulations.

These three scenarios are summarized in Fig. 3. In what follows we show how we simulate these three scenarios by choosing proper model protein sequences and structures.

#### 1. Scenario I: Purification potential

We first generate random protein sequences. For each protein sequence, we enumerate all possible conformations and define their true energies  $E'$  in the following way:

$$E' = E + N(0, \sigma^2), \quad (27)$$

where  $E$  is the energy as computed in Eq. (26),  $N(0, \sigma^2)$  is an additional Gaussian noise term with zero mean and standard deviation  $\sigma$  to reflect the approximate nature of the energy function formulation. In our study  $\sigma$  is set to 0.1. The struc-

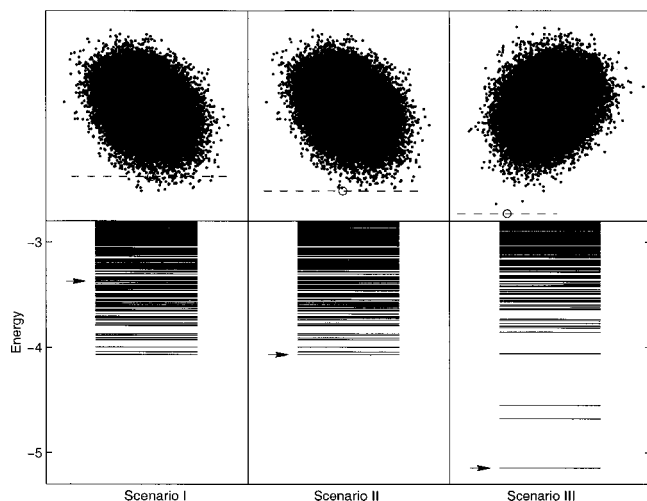


FIG. 3. Three scenarios based on the accuracy of energy function formulation. Top three plots show the two-dimensional projection of the 210-dimensional contact count vector. Dots (●) represent alternative conformations for a protein, and circles (○) represent the native structure. Dotted lines represent the reference energy parameters. Bottom three plots show the low energy tails of the energy spectra evaluated using the reference energy parameters. The native structure is indicated by the arrow. In scenario I, the potential is a purification potential, i.e., the energy of the native structure is among the lowest of all possible structures as judged by the reference energy parameters. In scenario II, the potential is a perfect discriminatory potential, i.e., the energy of the native structure is at the global minimum. In scenario III, the potential is a folding potential and the energy of the native structure is much lower than all alternative structures.

ture with the lowest energy  $E'$  is chosen to be the native structure; this structure will not have the lowest energy with the true energy,  $E$ .

## 2. Scenario II: Discriminatory potential

We generate random protein sequences. For each protein sequence, we enumerate all possible conformations and compute their energies as defined in Eq. (26). We then choose the conformation with the lowest energy to be the native conformation. All sequences with degenerate native structures are discarded.

## 3. Scenario III: Folding potential

We first generate protein sequences and native structures in the same way as in scenario II. This guarantees that the native structure is at the global energy minimum. In order to achieve a pronounced energy minimum for the native structure, we perform sequence design on the native structure.<sup>23</sup> We change the sequence to minimize the energy of the native structure using a simulated annealing procedure. At each step two residues are swapped and the energy difference between the new sequence and the old sequence is evaluated as  $\Delta E$  (this maintains the native amino acid composition). The new sequence is accepted if the following condition is satisfied:

$$e^{-\Delta E/kT} > p, \quad (28)$$

where  $p$  is a random number between 0 and 1, and  $kT$  is decreased linearly from 0.1 to 0 during the optimization. We perform  $10^5$  steps of sequence design and take the final se-

quence as the designed sequence for the native structure. We ensure that the target structure does have the lowest energy for the designed sequence. Again, all sequences with degenerate native structures are discarded.

## C. Decoys for recovering energy parameters

Decoys, either realistic or artificial, are needed to recover energy function parameters from a set of native protein structures. In this study we generally use realistic decoys, but also consider the effect of using artificial decoys constructed by shuffling the native protein sequence or contacts.

### 1. Realistic decoys on a lattice

Realistic decoys are alternative structures with the correct topological constraints. In our lattice model, there are a total of 57 336 such decoys excluding the native structure. Random subsets of these decoy conformations for each protein are used to recover the energy parameters by the three methods.

### 2. Shuffling sequence

For sequence shuffling, we generate random sequences with the same amino acid composition as in the native sequence, and thread them onto the native structure. These decoys do not have the same sequential chain connectivity as in the native protein and therefore are artificial.

### 3. Shuffling contacts

For contact shuffling, we generate random contact maps as decoys with the same chain connectivity and the same total number of contacts as in the native structure. The decoys generated are artificial and usually do not correspond to any physical lattice structures.

## D. Comparing energy parameters using correlation coefficient

We compute the linear correlation coefficient between derived and reference energy parameters as the main indicator of how effectively different methods extract knowledge-based energy functions. The linear correlation coefficient  $r$  for pairs of quantities  $(x_i, y_i)$  is given by the formula

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}. \quad (29)$$

We also compute the correlation between derived energy parameters with random starting points as an indication of the robustness of the extracting procedure. Since the parameter fluctuations are expected to be large, we use the more robust Spearman rank-order correlation coefficient  $r_s$ :<sup>24</sup>

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2} \sqrt{\sum_i (S_i - \bar{S})^2}}, \quad (30)$$

where  $R_i$  is the rank of  $x_i$  among the other  $x$ 's,  $S_i$  is the rank of  $y_i$  among the other  $y$ 's.

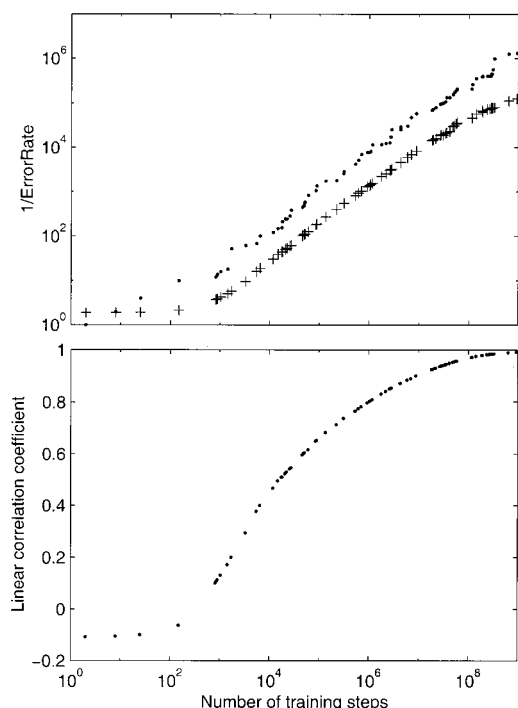


FIG. 4. Optimization recovers the reference discriminatory potential for scenario II by minimizing the error rate. During the optimization run, we retain copies of those energy parameters that have survived unchanged for the longest number of steps. Dots (●) in the top plot represent the number of correct assignments until the next error for these retained energy parameters. Pluses (+) in the top plot give the reciprocal error rate for the energy parameters evaluated over the test set. Dots (●) in the bottom plot give the linear correlation coefficients between the retained energy parameters and the reference potential.

## V. RESULTS

### A. Error rate minimization

Starting from random energy parameters, we perform error rate minimization using the pocket algorithm for  $10^9$  steps. The results are shown in Fig. 4 for energy function scenario II (discriminatory potential). We see that keeping the energy parameters that have so far survived unchanged for the longest number of steps is a very good approximation to the energy parameters with minimal error rates. During the optimization run, the error rate decreases gradually, and the correlation coefficient between the extracted and the reference energy parameters approaches asymptotically to 1. In fact, less than ten million steps are needed for extracted parameters that are 90% accurate.

For energy function scenario III (folding potential), we use the tunable parameter  $\delta$  described in Eqs. (13) and (14), which represents the energy gap between the native structures and all other alternative structures. We run shorter optimization procedure ( $10^8$  steps) using Eq. (14) with two different sets of random starting energy parameters, and compute the rank-order correlation coefficients between the two sets of extracted energy parameters. This measures the robustness of the optimization procedure. We plot the correlation against different  $\delta$  parameter in Fig. 5. As shown in the plot, the optimization procedure is not robust when  $\delta$  is 0 and no energy gap is imposed: the correlation between two

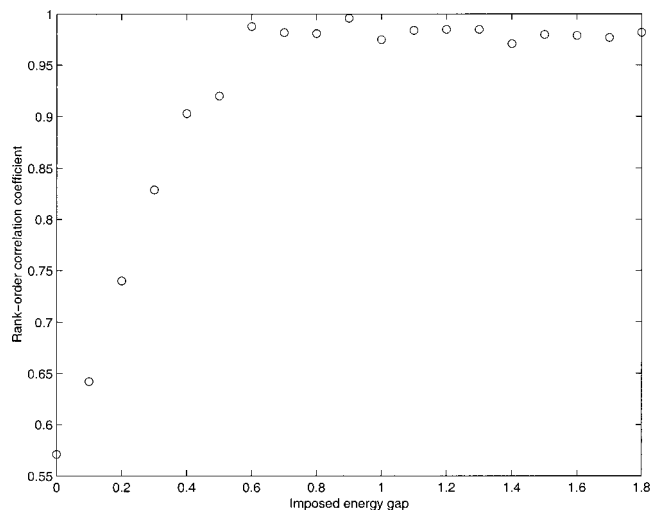


FIG. 5. Rank-order correlation coefficient between two sets of independently derived energy parameters using error rate minimization procedure for scenario III (folding potential), plotted against imposed energy gap [ $\delta$  parameter from Eq. (13)].

sets of independently derived energy parameters is less than 0.6. When we increased  $\delta$  and impose larger and larger energy gap, the correlation increases and reaches a steady maximum when  $\delta$  is around 0.7. We pick this value for  $\delta$  to run full optimization.

### B. Z-score optimization

The performance of Z-score optimization is summarized in Fig. 6 and Table I. In all three energy scenarios this procedure is able to find energy parameters with an average

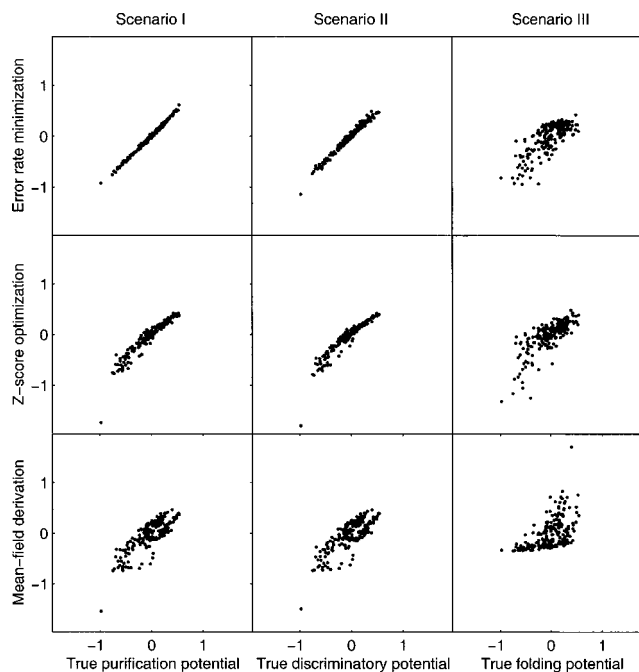


FIG. 6. Comparison of derived potentials vs reference potentials. Three different derivation methods (error rate minimization, Z-score optimization, and mean-field derivation) are tested in three different scenarios. Derived energy parameters are rescaled to have the same mean and standard deviation as the reference energy parameters.

TABLE I. Performance of potentials derived by different methods.

	True potential	Potential derived by		
		Error rate minimization	Z-score optimization	Mean-field statistics
Scenario I (purification potential)				
c.c. <sup>a</sup>	1.000	0.997	0.953	0.814
Z-score <sup>b</sup>	3.65	3.58	3.80	3.45
Error rate <sup>c</sup>	$4.6 \times 10^{-4}$	$4.8 \times 10^{-4}$	$9.8 \times 10^{-4}$	$3.3 \times 10^{-3}$
Scenario II (discriminatory potential)				
c.c. <sup>a</sup>	1.000	0.991	0.951	0.799
Z-score <sup>b</sup>	4.06	4.12	4.31	3.90
Error rate <sup>c</sup>	0.0	$7.6 \times 10^{-6}$	$6.4 \times 10^{-5}$	$4.6 \times 10^{-4}$
Scenario III (folding potential)				
c.c. <sup>a</sup>	1.000	0.782	0.775	0.525
Z-score <sup>b</sup>	5.37	6.50	6.90	3.83
Error rate <sup>d</sup>	$7.4 \times 10^{-3}$	$4.4 \times 10^{-4}$	$1.2 \times 10^{-3}$	$5.1 \times 10^{-2}$

<sup>a</sup>Linear correlation coefficient with the reference energy parameters.

<sup>b</sup>Average Z-score for the energy of the native structures evaluated over the test set.

<sup>c</sup>Error rate for the constraints  $E > E^n$ , evaluated over the test set.

<sup>d</sup>Error rate for the constraints  $E > E^n + 0.7$ , evaluated over the test set.

Z-score even higher than that of the reference potential. This is a sign of overfitting that reflects the differences between the normal distribution and the true interaction count distribution, especially at the low count limit that holds for repulsive forces, as discussed previously in this work. As a result, derived energy parameters systematically underestimate repulsive forces, as shown in Fig. 6. Nevertheless, the correlation between extracted potential and reference potential is excellent and the variance is low.

### C. Mean-field statistics

The performance of mean-field statistics is summarized in Fig. 6 and Table I. The extracted energy parameters are approximately unbiased, but the variance is high compared to the reference potential. This is likely due to the key approximation that ignores correlations between interaction count distributions in deriving the mean-field statistics. Mean-field statistics provides a rough but easily derived estimate to the energy parameters for any number of parameters. Mean-field statistics is easy to calculate even for a very large number of energy parameters and protein sequences, which is a clear advantage over the other two methods.

### D. Comparison of three methods under three energy function scenarios

We compare the extracted energy parameters with the reference energy parameters using three different methods (error rate minimization, Z-score optimization, and mean-field derivation) under three different scenarios. All derived energy parameters are rescaled to have the same mean and standard deviation as the reference parameters. We assess the quality of the extracted energy parameters by its correlation with the reference energy parameters as shown in Fig. 6. Our method of error rate minimization performs consistently best across all three scenarios. Z-score optimization also does very well, and mean-field potentials finish up third.

Each method has similar performances for scenario I and scenario II, but degraded performance for scenario III. For example, the best method of error rate minimization can recover >99% reference potential for scenarios I and II, but can only recover 78% for scenario III. The reason is that, in scenario III, there is a large energy gap between the lowest energy native structure and all other structures. This feature is unlikely to be changed by a small perturbation in the energy parameters and the error associated with the extracted energy parameters is large. For scenarios I and II, the native structure is close to alternative structures in the contact count space, and the energy parameters can be determined to a higher degree of accuracy.

We also show in Table I the Z-scores and error rates for the derived potentials evaluated on the test set. Mean-field potentials perform well judging from both Z-scores and error rates. Energy parameters derived from Z-score optimization perform better, but can still have two to ten times more errors than parameters derived from error rate minimization, which performs best. The relatively good performance of mean-field potentials explains their widespread use in protein structure prediction.

We observe that even though all three methods fail to provide a near-perfect correlation coefficient for scenario III, the extracted parameters perform very well as judged by both Z-score and error rate. This shows that with a good formulation of the energy function, even a rough parametrization can be very effective in separating the native structures from other misfolded structures.

### E. Energy parameters derived from sequence and contact shuffling

While our theory suggests that alternative, decoy structures are needed to recover energy parameters, for real proteins it is often time consuming to generate such structures. Indeed, many current methods for deriving knowledge-based energy functions do not use realistic decoys with correct to-

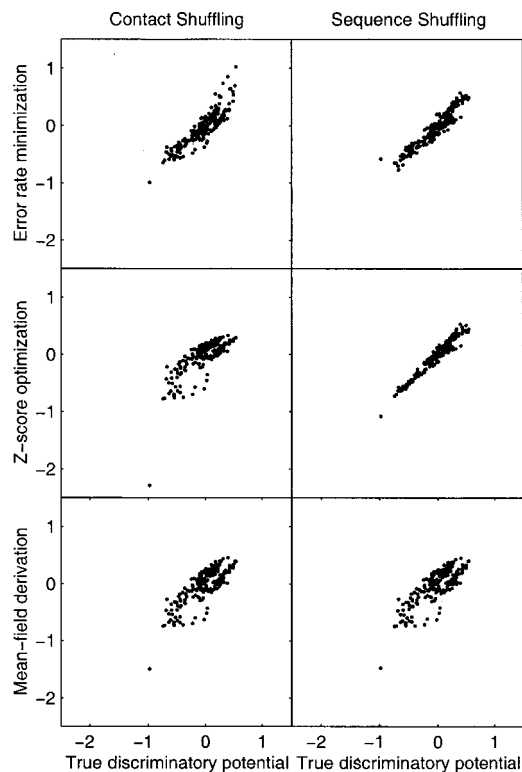


FIG. 7. Comparison of derived potentials vs true potentials for scenario II without explicit decoy construction. Three different methods (error rate minimization, Z-score optimization, and mean-field derivation) are tested. Artificial decoys are generated by either contact shuffling or sequence shuffling of the native structure. Derived energy parameters are rescaled to have the same mean and standard deviation as the reference energy parameters.

ological constraints. Rather, they use artificial decoys that are derived from the native structure by sequence or contact shuffling. These artificial alternative structures are easy to generate because they only involve random permutations of the native structure. On the other hand, they are not real alternative conformations for the given protein sequence. It is not obvious that energy parameters extracted from these structures will recover the reference potential.

We address this problem using our lattice model. We attempt to recover the reference energy parameters for energy function scenario II (discriminatory potential) by using three different methods. Instead of generating realistic decoy structures, we simply shuffle the sequence or contacts of the native structure to obtain alternative structures.

We compare derived energy parameters with reference energy parameters in Fig. 7. In both cases for sequence shuffling and contact shuffling, we are able to recover reference energy parameters to a degree, but not as well as the same procedure with realistic alternative structures. The only exception is for Z-score optimization combined with sequence shuffling, which recovers reference energy parameters better than the same procedure with realistic decoys. Even in this case, the performance is still not as good as our optimal method of error rate minimization with realistic decoy structures.

These results provide justification for using sequence shuffling or contact shuffling in energy function parametrization. However, these results also suggest that sequence

shuffling or contact shuffling on the native structure does not replace constructing alternative structures explicitly if accurate energy parameters are to be derived.

## VI. DISCUSSION

### A. Issues in deriving energy functions: Formulation vs parametrization

The complete process of deriving knowledge-based energy function consists of two steps: first selecting the proper formulation for the energy function, for example residue-residue contact, distance dependent, etc., and second, determining the best values for the parameters, i.e., parametrization. In this work we focus on energy function parametrization, but we also emphasize the importance of proper energy function formulation. As shown in Fig. 6 and Table I, formulation with different accuracy has a large impact on the subsequent parametrization step. It is easier to parametrize an effective energy function if we have a more accurate formulation.

While energy function parametrization can be stated in a mathematical way in this work, the process of energy function formulation is highly domain-specific and requires more insights into the nature of protein energetics and geometry. Our results suggest that in addition to refining the techniques for energy parametrization, we need to devote more efforts to exploring more accurate energy function formulations.

### B. Do mean-field potentials have a sound physical basis?

It has been argued that mean-field potentials are invalid because they lack justification from statistical physics.<sup>25</sup> In this article we provide justifications for mean-field potentials by means of parameter optimization. We do this with two physically based hypotheses: (a) the native structure occurs at the global minimum of free energy, and (b) protein energetics can be decomposed primarily into pairwise interactions. Assumption (a) is simply the thermodynamic hypothesis, and assumption (b) is also supported by physics: protein energetics involves van der Waals interactions, electrostatic forces, and solvation interactions. Both van der Waals and electrostatic interactions are pairwise. Solvation interactions are not pairwise; however, efforts have been made to approximate solvation interactions by pairwise terms.<sup>26</sup>

After determining the energy function formulation and constraints, energy parameters are subsequently optimized to fit the constraints. We have shown that under certain approximations, mean-field potentials can be derived as a result of optimization. The physical meaning of pairwise mean-field potential is therefore clear: it represents the best pairwise approximation of protein energetics. We note that energy functions derived from statistical mechanics are powerful as discriminatory potentials precisely because they are *implicitly* optimized for discriminating power. Indeed, many recent *ab initio* protein structure prediction methods used mean-field potentials and have proven to be effective.<sup>27–29</sup>

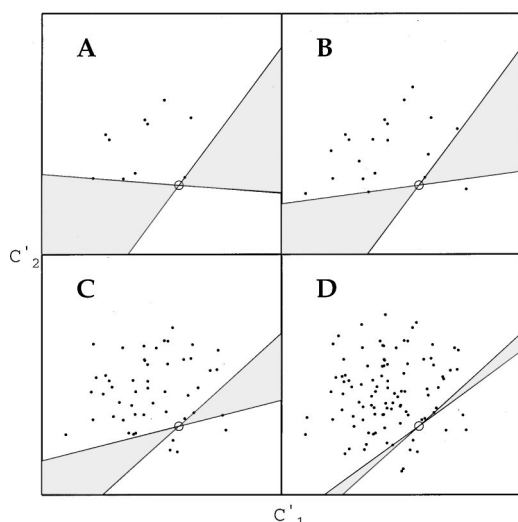


FIG. 8. Dependence of precision and accuracy of optimal energy parameters on the number of structures sampled. In this hypothetical case, only two interaction types exist:  $c'_1$  and  $c'_2$  are the relative counts in decoys compared with native structure for interaction type 1 and 2, respectively. Each dot represents a decoy structure. Dots are randomly generated following the normal distribution. The circle represents the collection of all native structures in the database, which is located at the origin [Eq. (11)]. The difference between the four plots is the number of points sampled,  $n$ . All hyperplanes within the shaded regions correspond to optimal energy function. (A)  $n=10$ . Decoy structures form a feasible set of constraints, but the error of optimal energy parameters is large due to insufficient sampling. (B)  $n=20$ . Decoy structures form an *infeasible* set; i.e., all constraints cannot be simultaneously satisfied. The error of the optimal parameters is still very large. This means that the infeasibility is only a necessary requirement: we need much more points to determine optimal energy parameters. (C)  $n=50$ . The error is reduced, but still quite large. (D)  $n=100$ . With only a two-fold increase in the number of samples, we see a dramatic improvement in the precision of the optimal energy parameters. This indicates that we now have good sampling.

### C. Importance of proper sampling

Proper sampling is crucial for energy function parametrization. First, there should be enough sampled points to determine the energy parameters with reasonable precision. Second, the sampled distribution should reflect the true interaction count distribution to determine the energy parameters with reasonable accuracy. Improper sampling will result in over-learning and large errors, especially in a high dimensional space with many parameters (see Fig. 8).

In particular, we need to make sure that enough structures are sampled in the vicinity of the native structures. These structures are needed to accurately position the hyperplane that represents the optimal energy parameters. In other words, quality of the decoys has a large effect on the accuracy of the derived energy parameters.

### D. Equivalent description in linear programming terms

In our approach, we work on the space of structures: we treat structures (constraints) as points, and energy parameters as normal vectors to hyperplanes. However, other people that use linear programming also concentrate on the dual space.<sup>17,30</sup> They work on the space of energy parameters: each point is a set of energy parameter, and each decoy struc-

ture is a constraint represented by a half-space. In linear programming language, minimizing the error rate function  $R(\mathbf{e})$  is the same as solving the maximum cardinality feasible set problem, i.e., finding a maximum cardinality set of constraints that is feasible. We choose to use our formalism because our geometrical interpretation is clearer and more intuitive in representing the error rate function.

### E. Will the methods work for a parameter space with higher-dimensions?

We only test energy function extraction for the 210-parameter residue–residue contact potential. In order to describe the protein energetics in a more accurate way, energy functions with many more parameters are needed. Will these three methods work for the more complicated energy function formulations?

Mean-field potential will clearly be feasible for any number of parameters because it does not involve any explicit optimization step. Indeed, mean-field potentials with more than 200 000 parameters have been proposed.<sup>8</sup> Both error rate minimization and Z-score optimization involve an explicit optimization step and therefore are likely to suffer from the curse of dimensionality. In our current study, we are able to recover the reference potential to a very high degree using both error rate minimization and Z-score optimization in the 210-dimensional parameter space. This suggests that the performance of the optimization procedure depends more on the shape of the decoy interaction count distribution than on the dimensionality. In our study, the shape of the decoy interaction count distribution is likely to be very smooth: it is roughly a mixture of simple (near-Gaussian) distributions with at most  $P$  clustered maxima, where  $P$  is the number of proteins. Because of this the search complexity may not be strongly dependent on dimensionality, allowing us to locate the global minima in a high dimensional space with 210 parameters. As a result, we expect to be able to extend both error rate minimization and Z-score optimization to a parameter space with dimensionality higher than 210.

## VII. CONCLUSION

In this article we presented a unified scheme of knowledge-based energy function parametrization from which most current approaches can be derived as approximations. We first start with a constraint satisfaction formulation, and then transform it into optimization formulation. The solution to this optimization problem only depends on the statistical distribution of interaction counts in the decoy set as compared to the native structures. By approximating the interaction count distribution to be either a normal or Poisson distribution, we are able to derive both Z-score optimization and mean-field statistics.

We implement our formalism of error rate minimization using the pocket algorithm and compare our method with both Z-score optimization and mean-field statistics using simple lattice models under three different energy function scenarios. Our method consistently performs best across all scenarios.

In this article, we outline the theoretical treatment of extracting knowledge-based energy functions as well as lattice model studies. We are currently deriving energy parameters for real protein structures using off-lattice models. That work will complement our present treatment and will provide further insights into the nature of knowledge-based energy functions.

## ACKNOWLEDGMENTS

This work is supported by NIH Grant GM 45514. Yu Xia is a Howard Hughes Medical Institute Predoctoral Fellow. We thank Patrice Koehl and Ram Samudrala for helpful discussions.

## APPENDIX

### 1. General solution for the optimal energy parameters

Here we present the detailed derivation of Eq. (7), a general solution for the optimal energy parameters  $\mathbf{e}^0$ . We start with Eq. (6) and its geometrical interpretation in Fig. 1.  $p_c(\mathbf{c})$  is the joint probability density function of the random interaction count vector  $\mathbf{c}$ .  $R(\mathbf{e})$  is a partial integral of  $p_c(\mathbf{c})$  with a boundary of an  $N-1$ -dimensional hyperplane that goes through point  $\mathbf{c}^n$ . The normal to the hyperplane that minimizes the integral determines the optimal energy parameter  $\mathbf{e}^0$ , and we denote the corresponding optimal hyperplane as  $\mathcal{A}$  (see Fig. 1). Assuming  $p_c(\mathbf{c})$  to be an arbitrary probability density function with a single maximum,  $R(\mathbf{e})$  is optimal when the following equation is satisfied:

$$\left\{ \frac{d}{d\mathbf{e}} \int_{(\mathbf{c}-\mathbf{c}^n) \cdot \mathbf{e} < 0} p_c(\mathbf{c}) d\mathbf{c} \right\}_{\mathbf{e}=\mathbf{e}^0} = 0. \quad (\text{A1})$$

The meaning of the equation is the following. Let us perturb the normal vector of  $\mathcal{A}$  infinitesimally and get a new  $N-1$ -dimensional hyperplane  $\mathcal{A}'$ . Both  $\mathcal{A}$  and  $\mathcal{A}'$  go through point  $\mathbf{c}^n$ , and their intersection is an  $N-2$ -dimensional hyperplane,  $\mathcal{B}$ . The above equation requires that the integral of  $p_c(\mathbf{c})$  over  $V$ , the region in between hyperplane  $\mathcal{A}$  and  $\mathcal{A}'$ , is zero. We note the following relationship between integral over  $V$  and integral within the hyperplane  $\mathcal{A}$ :

$$\int_V p_c(\mathbf{c}) dV = \int_{\mathcal{A}} p_c(\mathbf{c}) r(dS, \mathcal{B}) d\theta dS, \quad (\text{A2})$$

where  $d\theta$  is the infinitesimal angle between  $\mathcal{A}$  and  $\mathcal{A}'$ ,  $r(dS, \mathcal{B})$  is the signed distance between  $dS$  and hyperplane  $\mathcal{B}$  (see Fig. 9). Therefore from Eqs. (A1) and (A2) we have

$$\int_{\mathcal{A}} p_c(\mathbf{c}) r(dS, \mathcal{B}) dS = 0. \quad (\text{A3})$$

This equation holds true for an arbitrary hyperplane  $\mathcal{B}$  within  $\mathcal{A}$  that goes through point  $\mathbf{c}^n$ . Therefore, point  $\mathbf{c}^n$  is the mean of the distribution  $p_c(\mathbf{c})$  within hyperplane  $\mathcal{A}$ .

Now we assume that the mean of the distribution  $p_c(\mathbf{c})$  within hyperplane  $\mathcal{A}$  is also approximately the point that maximizes  $p_c(\mathbf{c})$  within  $\mathcal{A}$ . This is exact for both normal and Poisson distribution, and is a good first-order approximation for other distributions as well. Under this assumption, we see

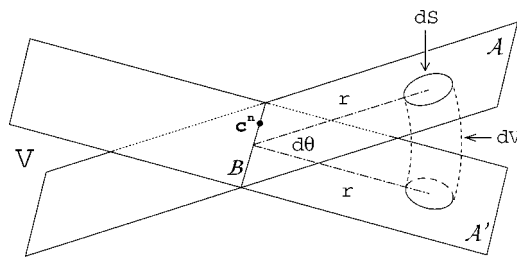


FIG. 9. Blowup and 3D representation of the region near the native structure in Fig. 1. Point  $\mathbf{c}^n$  is the native structure.  $N-1$ -dimensional hyperplanes  $\mathcal{A}$  and  $\mathcal{A}'$  correspond to two sets of energy parameters that differ infinitesimally. The intersection of  $\mathcal{A}$  and  $\mathcal{A}'$  is an  $N-2$ -dimensional hyperplane  $\mathcal{B}$ .  $V$  is the region in between hyperplane  $\mathcal{A}$  and  $\mathcal{A}'$ .  $d\theta$  is the angle between  $\mathcal{A}$  and  $\mathcal{A}'$ .  $dS$  is an infinitesimal area in  $\mathcal{A}$ .  $r$  is the signed distance between  $dS$  and hyperplane  $\mathcal{B}$ .

that the hyperplane  $\mathcal{A}$  is the tangent plane of contour map of  $p_c(\mathbf{c})$  through point  $\mathbf{c}^n$ . This leads to the simple solution for  $\mathbf{e}^0$  in Eq. (7).

In what follows we assume two approximations for  $p_c(\mathbf{c})$ : an independent normal distribution and an independent Poisson distribution. We show that these two approximations lead to  $Z$ -score optimization and mean-field statistics, respectively.

### 2. Case I: Independent normal distribution

Let us assume that  $p_c(\mathbf{c})$  follows independent normal distribution:

$$p_c(\mathbf{c}=\mathbf{c}_i) = \prod_{j=1}^N \frac{1}{\sqrt{2\pi}\sigma_j} e^{-(c_{ij}-\bar{c}_j)^2/2\sigma_j^2}. \quad (\text{A4})$$

Since a linear combination of independent normally distributed random variables is also normally distributed, we know that  $p_E(E)$  is also a normal distribution:

$$p_E(E) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(E-\bar{E})^2/2\sigma^2}, \quad (\text{A5})$$

where

$$\bar{E} = \sum_{j=1}^N \bar{c}_j e_j, \quad (\text{A6})$$

$$\sigma = \sqrt{\sum_{j=1}^N \sigma_j^2 e_j^2}. \quad (\text{A7})$$

By variable substitution, Eq. (6) can be transformed to the following equation:

$$\begin{aligned} R(\mathbf{e}) &= \int_{E < \sum_{j=1}^N c_j^n e_j} p_E(E) dE \\ &= \int_{Z(\mathbf{e})}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-E'^2/2} dE', \end{aligned} \quad (\text{A8})$$

where  $Z(\mathbf{e})$  is the  $Z$ -score for the native structure given by

$$Z(\mathbf{e}) = \frac{\sum_{j=1}^N (\bar{c}_j - c_j^n) e_j}{\sqrt{\sum_{j=1}^N \sigma_j^2 e_j^2}}. \quad (\text{A9})$$

Since each term in the integral in Eq. (A8) is positive, minimizing  $R(\mathbf{e})$  is equivalent to maximizing  $Z(\mathbf{e})$ .

We solve for  $\mathbf{e}^0$  and get

$$e_j^0 = k \frac{\bar{c}_j - c_j^n}{\sigma_j^2} \quad \text{for all } j, \quad (\text{A10})$$

where  $k$  is an arbitrary positive constant. This result can also be obtained from Eq. (7).

With the optimal value of  $\mathbf{e}^0$ , we can then calculate the optimal Z-score value as

$$Z^0 = \sqrt{\sum_{j=1}^N \left( \frac{\bar{c}_j - c_j^n}{\sigma_j} \right)^2}. \quad (\text{A11})$$

### 3. Case II: Independent Poisson distribution

Next we consider an alternative approximation of the interaction count distribution,  $p_c(\mathbf{c})$ , as an independent Poisson distribution:

$$p_c(\mathbf{c} = \mathbf{c}_i) = \prod_{j=1}^N \frac{e^{-\lambda_j} \lambda_j^{c_{ij}}}{c_{ij}!}. \quad (\text{A12})$$

The unsymmetrical Poisson distribution is a better approximation than the symmetrical normal distribution to the interaction count distribution, since interaction counts can never be negative. The tail of the interaction count distribution differs significantly from the tail of the normal distribution, especially at the limit of low count, and it is the tail of the distribution that is important in defining the optimal energy function parameters (see Fig. 2 for comparison between the two distributions and the resulting optimal energy parameters).

Using Sterling's approximation to approximate the Poisson distribution by a continuous function,

$$\ln x! \doteq x \ln x - x. \quad (\text{A13})$$

Substituting Eq. (A12) into Eq. (7), we get an analytical solution for the optimal energy parameter set  $\mathbf{e}^0$ :

$$e_j^0 = k \ln \frac{\lambda_j}{c_j^n} = k \ln \frac{\bar{c}_j}{c_j^n} \quad \text{for all } j, \quad (\text{A14})$$

where  $k$  is an arbitrary positive constant. This is equivalent to Boltzmann statistics. For the low count limit, the Sterling approximation is no longer valid. We account for this difference by allowing one adjustable parameter  $\alpha$ :

$$p_c(\mathbf{c} = \mathbf{c}_i) \propto \prod_{j=1}^N \frac{e^{-\lambda_j} \lambda_j^{c_{ij} + \alpha}}{e^{(c_{ij} + \alpha) \ln(c_{ij} + \alpha) - (c_{ij} + \alpha)}}, \quad (\text{A15})$$

which corresponds to the following solution:

$$e_j^0 = k \ln \frac{\bar{c}_j}{c_j^n + \alpha} \quad \text{for all } j. \quad (\text{A16})$$

The difference between the true Poisson distribution [Eq. (A12)] and the approximation [Eq. (A15)] is smallest when  $\alpha$  is about 1/2.

### 4. Extension to multiple proteins

Here we show that for energy functions extracted from multiple native protein structures (see Sec. II C), both Z-score optimization and mean-field statistics can also be derived under different approximations.

We first start with Eq. (11), and the optimal energy parameter  $\mathbf{e}^0$  should minimize the error rate  $R(\mathbf{e})$ . By assuming independent normal distributions for decoy interaction counts for each sequence, this optimization problem can be reduced to optimization of the average Z-score in a way similar to one outlined above in Sec. 2. In what follows, we show that under certain conditions, this optimization problem will also lead to Bayesian statistics.

First we introduce an approximation to Eq. (8):

$$R(\mathbf{e}) = \left\langle \theta \left( \sum_{k=1}^P E_k^n - \sum_{k=1}^P E_{k i_k} \right) \right\rangle_{i_1, \dots, i_P}, \quad (\text{A17})$$

where  $P$  is the number of proteins with known native structures. What we are doing here is to regard all  $P$  native structures as domains of one single imaginary protein, with no interactions between domains, and define the error rate  $R(\mathbf{e})$  based on the "native" structure and decoys for this imaginary protein. In this way we approximate the multiple protein problem as the single protein problem. This is an approximation. Nevertheless, minimizing this approximate error rate function is a necessary condition for a perfect energy function.

Now that the multiple protein problem is reduced to the single protein problem, by assuming independent Poisson distributions for decoy interaction counts in a way similar to one outlined above in Sec. 3, we get Boltzmann-like statistics as the first-order approximation result for optimal energy parameter.

- <sup>1</sup>J. Moult, *Curr. Opin. Struct. Biol.* **7**, 194 (1997).
- <sup>2</sup>S. Tanaka and H. A. Scheraga, *Macromolecules* **9**, 945 (1976).
- <sup>3</sup>S. Miyazawa and R. L. Jernigan, *Macromolecules* **18**, 534 (1985).
- <sup>4</sup>S. Miyazawa and R. L. Jernigan, *J. Mol. Biol.* **256**, 623 (1996).
- <sup>5</sup>M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl, *J. Mol. Biol.* **216**, 167 (1990).
- <sup>6</sup>M. J. Sippl, *J. Mol. Biol.* **213**, 859 (1990).
- <sup>7</sup>A. Godzik, A. Kolinski, and J. Skolnick, *Protein Sci.* **4**, 2107 (1995).
- <sup>8</sup>R. Samudrala and J. Moult, *J. Mol. Biol.* **275**, 895 (1998).
- <sup>9</sup>K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, *J. Mol. Biol.* **268**, 209 (1997).
- <sup>10</sup>A. V. Finkelstein, A. M. Gutin, and A. Y. Badretdinov, *Proteins: Struct., Funct., Genet.* **23**, 151 (1995).
- <sup>11</sup>P. D. Thomas and K. A. Dill, *J. Mol. Biol.* **257**, 457 (1996).
- <sup>12</sup>R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 4918 (1992).
- <sup>13</sup>T. L. Chiu and R. A. Goldstein, *Folding Des.* **3**, 223 (1998).
- <sup>14</sup>M. H. Hao and H. A. Scheraga, *J. Phys. Chem.* **100**, 14540 (1996).
- <sup>15</sup>L. A. Mirny and E. I. Shakhnovich, *J. Mol. Biol.* **264**, 1164 (1996).
- <sup>16</sup>V. N. Maiorov and G. M. Crippen, *J. Mol. Biol.* **227**, 876 (1992).
- <sup>17</sup>G. M. Crippen, *J. Mol. Biol.* **260**, 467 (1996).
- <sup>18</sup>G. M. Crippen, *Folding Des.* **2**, S58 (1997).
- <sup>19</sup>M. Vendruscolo and E. Domany, *J. Chem. Phys.* **109**, 11101 (1998).
- <sup>20</sup>C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, New York, 1995).
- <sup>21</sup>A. Sali, E. Shakhnovich, and M. Karplus, *Nature (London)* **369**, 248 (1994).

- <sup>22</sup>D. A. Hinds and M. Levitt, *J. Mol. Biol.* **243**, 668 (1994).
- <sup>23</sup>E. I. Shakhnovich and A. M. Gutin, *Protein Eng.* **6**, 793 (1993).
- <sup>24</sup>W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. (Cambridge University Press, Cambridge, England, 1997).
- <sup>25</sup>A. Ben-Naim, *J. Chem. Phys.* **107**, 3698 (1997).
- <sup>26</sup>J. Weiser, P. S. Shenkin, and W. C. Still, *J. Comput. Chem.* **20**, 217 (1999).
- <sup>27</sup>K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker, *Proteins: Struct., Funct., Genet.* **S3**, 171 (1999).
- <sup>28</sup>R. Samudrala, Y. Xia, E. Huang, and M. Levitt, *Proteins: Struct., Funct., Genet.* **S3**, 194 (1999).
- <sup>29</sup>Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala, *J. Mol. Biol.* **300**, 171 (2000).
- <sup>30</sup>J. Mourik, C. Clementi, A. Maritan, F. Seno, and J. R. Banavar, *J. Chem. Phys.* **110**, 10123 (1999).