# Fold Recognition using the OPLS All-Atom Potential and the Surface Generalized Born Solvent Model

Anthony K. Felts[1], Anders Wallqvist[1*,**], Emilio Gallicchio[1], Donna Bassolino[2], Stanley R. Krystek[2], and Ronald M. Levy[1**]

[1]Department of Chemistry, Rutgers University, Wright-Rieman Laboratories, 610 Taylor Rd, Piscataway, New Jersey 08854-8087
[2]Bristol-Myers Squibb, Pharmaceutical Research Institute, Route 206 & Provinceline Road, Princeton, NJ 08543-4000

**Abstract.** Protein decoy data sets provide a benchmark for testing scoring functions designed for fold recognition and protein homology modeling problems. It is commonly believed that statistical potentials based on reduced atomic models are better able to discriminate native-like from misfolded decoys than scoring functions based on more detailed molecular mechanics models. Recent benchmark tests, however, suggest otherwise. Further analysis of the effectiveness of all atom molecular mechanics scoring functions for detecting misfolded decoys and direct comparison with results obtained using a statistical potential derived for a reduced atomic model are presented in this report. The OPLS all-atom force field is used as a scoring function to detect native protein folds among the Park & Levitt large decoy sets. Solvent electrostatic effects are included through the Surface Generalized Born (SGB) model. The OPLS potential with SGB solvation (OPLS-AA/SGB) provides good discrimination between native-like structures and non-native decoys. From an analysis of the individual energy components of the OPLS-AA/SGB potential for the native and the best–ranked decoy, it is determined that a roughly even balance of the terms of the potential is responsible for distinguishing the native from the misfolded conformations. Different combinations of individual energy terms provide less discrimination than the total energy. The effects of scoring decoys using several dielectric models are compared also. With the SGB solvation model, close to 100% of the structures with energies within 100 kcal/mol of the native state minimum are native-like. In contrast, only 20% of the low energy structures are found to be native-like when a distance dependent dielectric is used instead of SGB to model solvent electrostatic effects. The results are consistent with observations that all-atom molecular potentials coupled with intermediate level solvent dielectric models are competitive with knowledge-based potentials for decoy detection and protein modeling problems such as fold recognition.

**Keywords: protein decoys, homology modeling, protein solvation,fold recognition**

* Current address: NCI-Frederick/SAIC Bldg.430, P.O.Box B, Frederick MD 21702
** Corresponding authors.

## 1 Introduction

An essential requirement for protein structure prediction methods is the ability to discriminate native and native-like conformations from significantly misfolded ones. Several methods have been proposed which can fit roughly into three categories: knowlege-based, physics-based, or a combination of these [1]. Several varieties of knowledge-based empirical scoring functions for ranking protein conformations have been proposed [2–6,1]. Some of these implement statistical potentials which are "trained" to recognize native conformations. Knowledge-based potentials are well suited for fold recognition applications where the best conformation of a protein is selected from a database of known protein conformations. Scoring functions applicable to ab initio folding studies, which require differentiable potentials and the inclusion of excluded volume terms, have also been developed. These are based on combinations of knowledge-based potentials and reduced atomic models sometimes augmented by simplified solvation models based on hydrophobic or hydrophilic exposure [7].

Physics-based all-atom molecular mechanics force fields have not been generally considered practical for fold detection because they are parameterized on small molecule data rather the on proteins directly: the level of atomic detail contained in these models is considered poorly matched to the fold detection problem with respect to both accuracy and computational cost. Recent studies have shown, however, that a scoring function based on the potential energy from an all-atom molecular mechanics force field can recognize native protein conformations among a set of decoys as well as the best available knowledge-based scoring functions [1].

The use of an all-atom force-field minimizes the assumptions which are inherent in an empirical scoring function and, as will be shown, the inclusion of more refined solvation models enhances our ability to discriminate native folds. An additional value of the all-atom potential lies in its suitability for modeling proteins at higher resolution. This is an important feature for applications in studies concerned with the relationship between structure and function such as homology modeling, structure-based drug-design and protein-protein recognition.

Although all-atom force fields allow for explicit simulations of a solvent, the cost required to appropriately sample solvent configurations rapidly becomes prohibitive. Simplified solvation models are more computationally efficient while preserving a reasonably accurate representation of the interactions between the protein and its aqueous environment. Although no continuum model can wholly account for the explicit inclusion of solvation [8,9], free energies of solvation of small molecules have been obtained accurately with these methods to within a fraction of a kcal/mole relative to experiments [10–15].

Solvation effects have been included using a variety of simple models [16–23]. These models have been based on exposed surface area, dielectric continuum methods, and screened or modified Coulomb interactions. The validity

of a continuum representation of the solvent based on the Poisson-Boltzmann equation has been studied extensively for small and large molecules [24–30]. Continuum solvation models that treat solute and solvent as two dielectric regions with different dielectric constants have been used successfully to account for solute free energies of hydration [31,32,11,33,34]. Dielectric models based on the Born model [35] have been developed for which free energies of hydration are comparable to the predictions of Poisson-Boltzmann and explicit solvent models [36–42].

The inclusion of solvation effects with an all-atom molecular mechanics force field has been shown to be important for the recognition of the native state [43,16,44,45,17]. Scheraga and co-workers [46,47] used explicit all-atom protein models in conjunction with solvation models based on the molecular exposed surface area. A similar approach by Wang et al [48,49] showed that the inclusion of solvation effects can successfully discriminate the native from non-native structures. Vieth and coworkers [50] generated structures of the small 33 residue GCN4 leucine zipper proteins using a simplified lattice model; promising structures were then converted to all-atom models and evaluated using a molecular mechanics force field. A hierarchical method of generating large numbers of protein folds was also employed by Monge et al. [20]. to select and evaluate structures using the AMBER all-atom force field model [51] with the generalized Born continuum solvent model of Still and co-workers [37] representing the aqueous environment. For decoy sets of three different proteins the protocol performed reasonably well in distinguishing the native structure. All-atom models with continuum solvent were used also as the basis for discriminating non-native states for a small set of twelve deliberately misfolded proteins studied by Vorobjev et al. [52]. In their protocol conformations for each protein are first sampled from a molecular dynamics trajectory in order to capture micro-states of the protein; this is followed by an evaluation using a dielectric continuum model. Lazaridis and Karplus [22] used the CHARMM19 protein force field together with a Gaussian solvation shell model for the solvation free energy to distinguish deliberately misfolded from native conformations considered on a pairwise basis, and in large decoy sets. Dominy and Brooks [53] also used the CHARMM19 force field but with the addition of the generalized Born solvation term [37] to distinguish misfolded conformations in the EMBL [17] and Park & Levitt [54] decoy sets.

Given the complexity of the protein potential surface it is virtually impossible to consistently find the global minimum starting from an arbitrary point on the surface. Instead, tests have been designed whereby the scoring function is "challenged" to find the native conformation among an ensemble of conformations, most of which are compact but non-native. Many empirical energy functions have been used to identify the correct native structure among a collection of known protein structures using fold recognition techniques [55,2,56–60]. Scoring functions are also used to identify native-like conformations from a large set containing native and decoy non-native

conformers [61,54,62,63,22,64]. Due to the large ensemble of conformations available, the use of large decoy sets to evaluate scoring functions is a more demanding test than fold recognition and is well suited for the evaluation of scoring functions based on an all-atom force field.

In this work we show that the all atom (OPLS-AA) force field for proteins [65] together with a surface integral formulation of the generalized Born model (SGB) [40,42] is capable of discriminating between native and non-native folds among large sets of compact decoy structures. Validation of the scoring protocol is performed on a large database of well-packed misfolded and near-native protein conformations generated by an algorithm designed to cover exhaustively the relevant parts of conformational space [66,54,67]. The inclusion of near-native decoys in these sets is important in determining whether the scoring function is well behaved in the vicinity of an idealized native conformation, since it is unlikely that any ab initio method of generating conformations will generate the native state exactly. In any case, the native state actually represents an ensemble of closely related conformations.

Individual components of the energy perform worse than the total energy, e.g. for the bulk of the well-packed decoys the van der Waals energy provides very little information about structural similarity between a well-packed non-native structure and the native state. It is also shown that some aspects of the SGB model results can be mimicked by a screened electrostatic energy, although the SGB approximation provides a better discriminatory measure between non-native and native states.

## 2 Methods

### 2.1 Details of the Calculations using IMPACT

The energy of each protein structure investigated was calculated using the OPLS-AA/SGB force field implemented in the IMPACT modeling program (Schrödinger, Inc.) [68]. Initial structures were first minimized in order to remove any artifacts that result from the coordinates being generated with a different energy function; only minimized energies are reported here. All non-native coordinates were taken from the Park and Levitt decoy sets [54] as described below; native protein coordinates were obtained from the Protein Data Bank (PDB) [69]. The force field employed in the calculation of the atomic interactions was the OPLS all-atom force field [65], including parameters for all intramolecular degrees of freedom. The surface formulation of the generalized Born model [37,39] (SGB) as coded in IMPACT was used to estimate the solvation energy [40,70].

The total energy for a protein in vacuum is given by,

$$U_{\mathrm{tot}}^{\mathrm{vac}} = U_{\mathrm{bond}} + U_{\mathrm{angle}} + U_{\mathrm{torsion}} + U_{\mathrm{Coulomb}} + U_{\mathrm{vdW}}, \tag{1}$$

where the first three terms refers to intramolecular interactions arising from the connectivity of the molecule and the last terms reflect nonlocal interactions within the protein. The van der Waals energy, $U_{\mathrm{vdW}}$, is modeled by the

standard 6-12 Lennard–Jones interaction. The energy of the protein in water calculated according to the SGB continuum solvent model is

$$U_{\rm tot}^{\rm con} = U_{\rm tot}^{\rm vac} + U_{\rm SGB} + U_{\rm cav}\,, \tag{2}$$

where $U_{\rm SGB}$ denotes the electrostatic contribution to the solvation energy calculated using the SGB method, and the cavity term $U_{\rm cav}$ is taken as $\gamma A$ where $A$ is the accessible surface area of the molecule and $\gamma = 5$ cal/(Å$^2$ mole) [40].

The SGB model is the surface implementation [40,42] of the generalized Born model [37]. The generalized Born equation

$$U_{\rm SGB} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\rm in}} - \frac{1}{\epsilon_{\rm w}}\right)\sum_{ij}\frac{q_i q_j}{f_{ij}(r_{ij})}\,, \tag{3}$$

where $q_i$ is the charge of atom $i$ and $r_{ij}$ is the distance between atoms $i$ and $j$, gives the electrostatic component of the free energy of transfer of a molecule with interior dielectric $\epsilon_{\rm in}$ from vacuum to a continuum medium of dielectric constant $\epsilon_{\rm w}$, by interpolating between the two extreme cases that can be solved analytically: the one in which the atoms are infinitely separated and the other in which the atoms are completely overlapped. The interpolation function $f_{ij}$ in Eq. (3) is defined as

$$f_{ij} = \left[r_{ij}^2 + \alpha_i\alpha_j \exp(-r_{ij}^2/4\alpha_i\alpha_j)\right]^{\frac{1}{2}}\,, \tag{4}$$

where $\alpha_i$ is the Born radius of atom $i$ defined as the effective radius that reproduces through the Born equation

$$U_{\rm single}^i = -\frac{1}{2}\left(\frac{1}{\epsilon_{\rm in}} - \frac{1}{\epsilon_{\rm w}}\right)\frac{q_i^2}{\alpha_i}\,, \tag{5}$$

the electrostatic free energy, $U_{\rm single}^i$, of the molecule when only the charge of atom $i$ is turned on. The SGB method estimates $U_{\rm single}^i$ by integrating the interaction between atom $i$ and the charge induced on the molecular surface by the Coulomb field of this atom

$$U_{\rm single}^i = -\frac{1}{8\pi}\left(\frac{1}{\epsilon_{\rm in}} - \frac{1}{\epsilon_{\rm w}}\right)\int_S \frac{q_i^2}{|\mathbf{r}-\mathbf{r}_i|^4}(\mathbf{r}-\mathbf{r}_i)\cdot\mathbf{n}(\mathbf{r})d^2\,\mathbf{r}\,. \tag{6}$$

The SGB method has been shown to compare well with the exact solution of the Poisson-Boltzmann (PB) equation. The SGB implementation used in this work includes further correction terms that bring the SGB reaction field energy even closer in agreement with exact PB results [40].

To help assess the ability of the energy function to discriminate between non-native and native protein conformations, the energy gaps between the decoy conformations and the native are evaluated,

$$\Delta U = U_{\rm tot}^{\rm decoy} - U_{\rm tot}^{\rm native}\,, \tag{7}$$

Energy gaps of individual energy terms have also been examined (see Eqs. 1 and 2). Unless explicitly noted all results presented below were performed without energy cut-offs, i.e. all possible non-bonded interactions are included in the total energy. The structural similarity between two protein conformations is expressed as a root mean square deviation (rmsd) between the best overlap of the alpha-carbon ($C_\alpha$) atoms of the two conformations.

### 2.2 Knowledge-based Calculations using ProCeryon

Fold recognition as implemented in ProCeryon (ProCeryon Biosciences Inc.) [71] was used for the protein structures contained in the decoy sets (described below) to discern the native fold from an ensemble of folds. The energy of each protein in the decoy sets was evaluated using the knowledge-based potentials as implemented in the ProCeryon software. Three scores formed the basis for our evaluation. The inter-atom pairwise potentials, a surface potential and a combination of the pairwise and surface scores. An 8 Å cutoff was used for the pairwise potentials and a 12 Å cutoff was used for the surface potentials. All the default values were used for the gap restrictions because they are irrelevant when the sequences are identical as in the case of the decoys.

### 2.3 Dataset of Decoys

Although we are probing various energy functions for their ability to differentiate between native and non-native structures, none of the coordinate sets are originally generated by these functions. The vastness of the conformational space and the complexity of an all-atom potential energy function effectively hinders the full sampling of the appropriate degrees of freedom. Scoring conformations with the OPLS/SGB potential may be considered as a last step in the process of generating protein folds, i.e. only at the end would it be appropriate to spend the time and effort to evaluate a complex all-atom potential energy function. For this study we focus on an existing decoy dataset as our conformational space. This dataset, the Park and Levitt database of structure decoys [54], has proven to be highly non-trivial to score correctly.

The Park and Levitt database contains structure decoys for 7 small proteins [54]. The protein structures were generated by exhaustively enumerating the backbone rotamers states of ten selected residues in each protein using an off-lattice model with four discrete dihedral angle states per rotatable bond. From this dataset, containing hundreds of thousand of conformations, the authors selected for further evaluation only compact structures that scored well using a variety of scoring functions as well as those having a reasonable rmsd from the native [54]. The coordinates, available on the internet (http://dd.stanford.edu), are all-atom models built from the $C_\alpha$ atoms with the program SEGMOD [72]. No further refinement of these coordinates was done except for minimizing the structures using our energy function (see

Eqs. 1-2). The decoy datasets are summarized in Table 1 and encompass a range of small proteins from 54-75 residues with varying topological folds. The number of decoys in these sets range from 630 for 1ctf (the carboxy-terminal domain of L7/L12 50s ribosomal protein from *Escherichia coli*) to 687 for 4pti (bovine pancreatic trypsin inhibitor).

| PDB name | $N_{\mathrm{res}}$ | $N_{\mathrm{decoy}}$ | q ($e$) |
|---|---|---|---|
| 1ctf | 68 | 630 | -2 |
| 1r69 | 63 | 675 | +4 |
| 1sn3 | 65 | 660 | +1 |
| 2cro | 65 | 674 | +6 |
| 3icb | 75 | 653 | -7 |
| 4pti | 58 | 687 | +6 |
| 4rxn | 54 | 677 | -12 |

**Table 1.** The sequence length, $N_{\mathrm{res}}$, the number of decoys, $N_{\mathrm{decoy}}$, and total charge of the seven proteins of the Park & Levitt set [54]

The energy of each native and model structure is minimized using the full atomic model with and without the SGB dielectric continuum solvation energy term. Also, the energy of each structure, native and decoy, was evaluated using the ProCeryon program [71].

## 3 Results and Discussion

The problem of differentiating non-native states from native-like states can be expressed as the ability of a scoring function, depending only on the coordinates of each structure, to score the native states better than any other structures. If such a scoring function were used also to generate structures, a further desirable property would be that in the vicinity of the native state the structural similarity to the native state would be a monotonically increasing function of improved scores.

### 3.1 OPLS-AA/SGB Calculations on the Park & Levitt decoys

Examination of minimized energies for the 7 extensive datasets of protein decoys (see Figure 1) shows that using the OPLS-AA/SGB potential, protocol no decoy scores better than the X-ray structure. The correlation between structural similarity and score is strong only for structures with low rmsd. For rmsd $>$ 4 Å this correlation breaks down. Native-like states appear around 2 Å at low energies, with the bulk of the decoys being in non-native like conformations with rmsd above 4 Å.

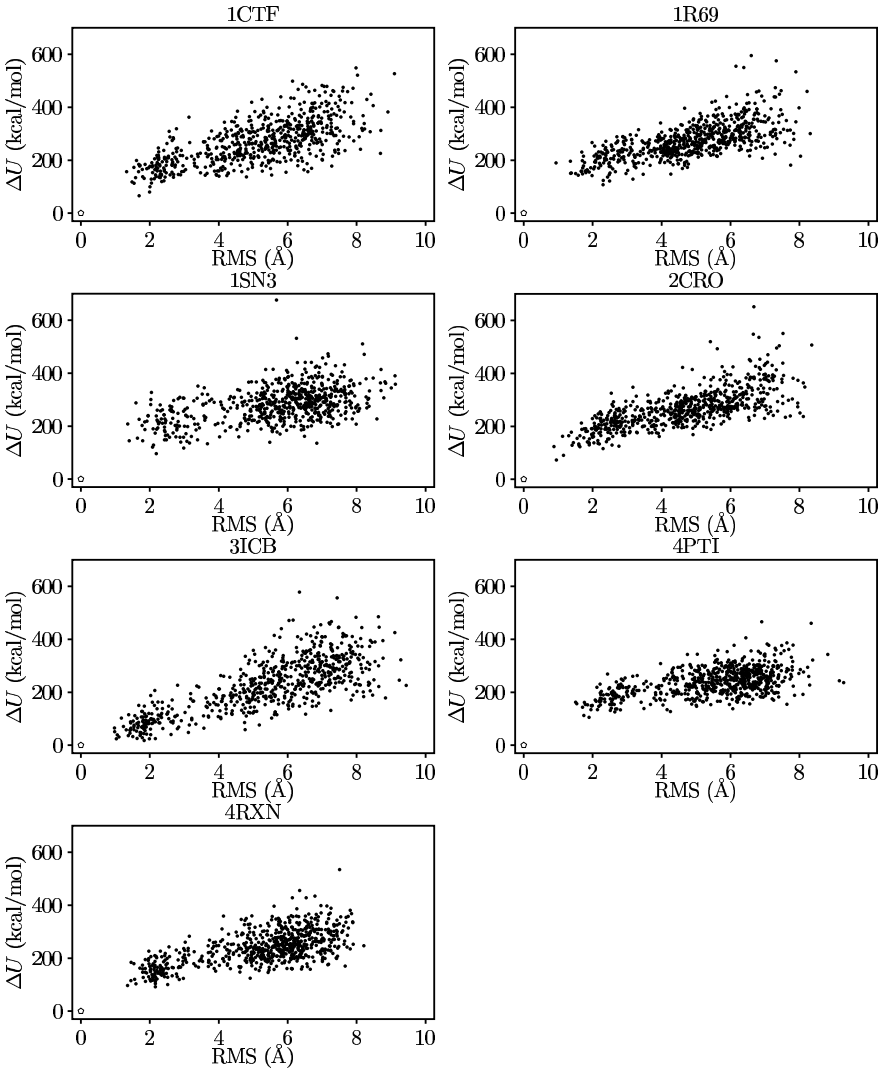


**Fig. 1.** OPLS-AA/SGB: energy gap/RMS correlation plots for the Park & Levitt decoy sets.

In Table 2 we report the statistical indicators of the quality of the scoring function. Some of the indicators depend on defining the reference structure as the native X-ray structure. It has been verified that similar results are obtained by selecting any native-like decoy as the reference structure. A global view of the results for the Park & Levitt sets is given in Fig. 2. The fraction, $P(\Delta U)$, of native-like decoys with an energy gap from the native less than $\Delta U$ is shown. A decoy conformation with an rms less than 3 Å is considered native-like. Fig. 2 indicates, for example, that structures with an energy gap from the native less than 100 kcal/mole have a ∼90% chance of being native-like, whereas a decoy with a +200 kcal/mole energy gap from the native has only a 20% chance of being native-like. For these datasets there are no decoy structures with a total energy, $U_{\rm tot}^{\rm con}$, below that of the native state (i.e. energy-minimized X-ray coordinates; see Fig. 1). This suggests that if a fold prediction program can generate protein structures within 100 kcal/mole of the native state there should be a high (> 90%) chance of finding native-like states in this dataset.

| PDB name | $U_{\rm native}$ | min($\Delta U$) | rmsd | $Z_{\rm nat}$ | $\overline{Z}_{\rm nat-like}$ | $r_S$ |
|---|---|---|---|---|---|---|
| 1ctf | -4213.92 | +65.55 | 1.69 | -3.24 | -1.08 | 0.66 |
| 1r69 | -3499.46 | +107.16 | 2.30 | -4.03 | -1.01 | 0.70 |
| 1sn3 | -3467.53 | +96.08 | 2.19 | -4.22 | -1.04 | 0.45 |
| 2cro | -3628.30 | +72.55 | 0.94 | -3.69 | -0.95 | 0.73 |
| 3icb | -4694.45 | +18.08 | 1.84 | -2.18 | -1.34 | 0.76 |
| 4pti | -3055.04 | +105.07 | 1.89 | -4.53 | -1.15 | 0.47 |
| 4rxn | -3363.51 | +92.06 | 2.16 | -3.76 | -1.29 | 0.58 |

**Table 2.** OPLS-AA/SGB results: the minimized energy, $U_{\rm native}$, of the native conformation; the energy gap, min($\Delta U$) and the rms deviation between the best scoring decoy and the native conformation; the native Z-score, $Z_{\rm nat}$, the average Z-score, $\overline{Z}_{\rm nat-like}$, of the native-like conformations in the Park & Levitt decoy sets [54] and the Spearman rank–order correlation coefficient, $r_S$, for the energies and rms's of all structures in the decoy sets.

Another measure of the fitness of the scoring functions is to evaluate the rmsd of the lowest energy structure in each decoy set. The results are summarized in Table 2. The rmsd of the lowest energy decoy range from 0.94–2.20 Å with an average rmsd of 1.9 Å. These decoys fall within the native-like designation. The average energy deviation from the native energy is +79.5 kcal/mole, which represents an average deviation of +2 % from the native total energy values. As we shall see below, not all scoring functions examined yield decoy energies consistently higher than the native energy.

In Table 2 we also report the native Z-score, $Z_{\rm nat}$, and the average Z-score of the native-like decoys, $\overline{Z}_{\rm nat-like}$. The Z-score of conformation $i$ is defined

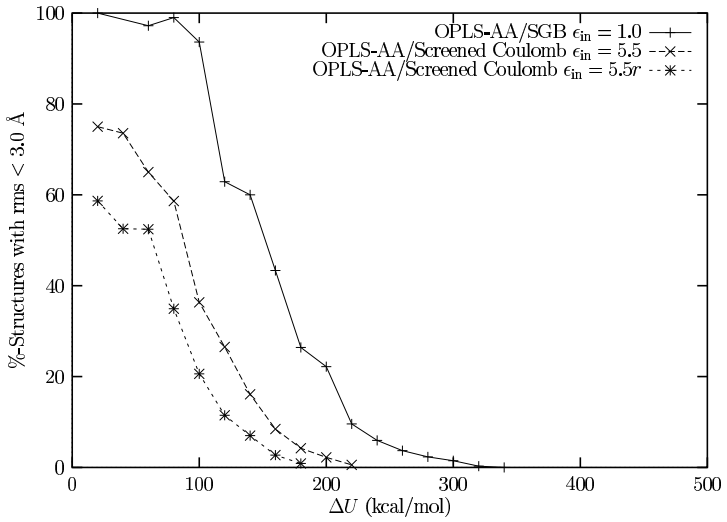


**Fig. 2.** Fraction of the Park & Levitt decoys with energy gap from the native less than $\Delta U$ which are native-like (rmsd from native $< 3$ Å), using the OPLS-AA/SGB potential function and the vacuum OPLS-AA potential with screened Coulomb interactions.

as

$$Z_i = \frac{E_i - \overline{E}}{\sigma}, \tag{8}$$

where $E_i$ is the energy of the particular conformation, $\bar{E}$ is the average score and $\sigma$ is the standard deviation of the distribution of scores in the set. The average Z-score, $\overline{Z}_{\mathrm{nat-like}}$ is obtained by averaging the Z-scores of the native-like decoys. The Z-score measures the ability of the scoring function to recognize native conformations. Assuming the distribution of scores is approximately Gaussian, a native Z-score of, say, $-2$ indicates that the native structure is ranked in the best 1% in the decoy set. In general, the more negative the Z-score the better. The values of the native Z-scores range from $-3.2$ to $-4.5$ indicating that the scoring function is extremely successful in finding the native structure among the decoys. The native-like average Z-score represents the ability of the scoring function to discriminate the native-like conformations from the non-native conformations. The more negative the average native-like Z-score the larger the probability that a low-energy conformation is a conformation structurally similar to the native. The calculated values of the Z-scores ranging from $-0.95$ to $-1.34$ indicate that, although on average the native-like conformations have lower energies than the non-native conformations, a significant number of native-like structures have a favorably low Z-score. This can also be seen from Figure 1 by looking at the

vertical position of the low-rmsd structures with respect to the bulk of the decoys. This does not necessarily indicate a deficiency of the energy function but rather that for native-like conformations, i.e. those with the correct fold, the energy is also sensitive to the position and orientation of the amino acid side-chains. An incorrect placement of a side-chain may be enough to increase the energy of a native-like fold to the level of the misfolded conformations. A native-like energy is achieved only when all of the structural elements of the protein are placed correctly [22].

Park and Levitt [54] have evaluated six simple empirical scoring functions using the same decoy sets examined in this work. A comparison between the native and native-like Z-scores calculated here with those obtained by Park and Levitt shows that the OPLS-AA/SGB energy model clearly outperforms the six empirical scoring functions examined in the Park and Levitt work. Moreover, none of the empirical scoring functions examined by Park and Levitt was able to consistently rank first the native conformation, whereas the OPLS-AA/SGB model does.

While the native-like Z-scores measure to some extent the correlation between the calculated energies and the structural similarity measured by the rms between the misfolded and native conformations, a more thorough indicator of this is given by the Spearman rank–order correlation coefficient [73]. Instead of finding the correlation coefficient of a linear regression fit through the raw data, a correlation is determined between the ranks of the energies and the rms's. This non-parameterized fit avoids calculating any non-significant correlations which might be due to scaling effects alone. In Table 2, we present the Spearman rank–order coefficients for each decoy data set. For all seven proteins, these coefficients are 0.45 or better; and for five of the seven, they are around 0.6 or better showing a significant correlation between the energy ordering of the conformations and their structural similarity to the native.

A brief comparison to another all-atom force field with solvation can be made with the work by Dominy and Brooks [53] who performed calculations using the CHARMM19 force field in conjunction with the generalized Born solvation model on three (1r69, 2cro, and 3icb) of the seven Park & Levitt decoy sets. They were capable of completely discriminating the native from the misfolded conformations for two of the proteins (1r69 and 2cro) and were almost as successful with the third. Our Z-scores (see Table 2) were similar to theirs (their Z-scores were 4.2, 3.3, and 2.2 for 1r69, 2cro, and 3icb respectively) [53], but our Spearman rank–order coefficients (also in Table 2) were significantly better (their coefficients were 0.53, 0.59, and 0.73 respectively) [53]. The latter indicates that the OPLS-AA/SGB potential produces a better correlation between the conformations' energies and their structural similarity to the native. This provides an advantage in distinguishing native-like folds from those which are clearly non-native.

### 3.2 Comparison between OPLS-AA/SGB and ProCeryon

A further comparison is conducted with the emperical knowledge-based scoring function developed by Sippl et al. [74,75] incorporated into the ProCeryon software package [71]. This scoring function is comprised of two terms: a pairwise scoring function and a surface term representing the interaction between the solvent and the protein. The performance of each individual term along with the combination of the two is examined. Tables 3, 4, and 5 show how well the native is discriminated from misfolded structures based on these three scoring schemes. (Note that for the scores from the ProCeryon program, the higher the score is, the better the fold. This is opposite from the calculations using OPLS-AA/SGB where the lower the energy is, the more stable the fold. Likewise, the more positive a resulting Z-score for a given ProCeryon score, the more significantly that score lies on the outside of the distribution of scores.) The surface score alone does not distinguish the native from the non-natives well. Only for 3icb does the surface score rank the native the best. This is also reflected in the native Z-scores which are only around 2.2 or lower. Either using the pairwise score or the combined score is moderately successful in properly discriminating the native from the rest of the decoys. The pairwise scoring scheme ranks the native the best for six out of the seven proteins, but for 4rxn, the one failure, it ranks 24 non-native conformations better. While the combined score only ranks the native best for five of the seven proteins, it does not rank the native so drastically low for the two it failed. Those two are 2cro for which the native ranked third and, again, 4rxn where the native ranked fifth. In contrast, with the OPLS-AA/SGB model, the native structure ranks first for all seven proteins in the Park & Levitt set. Also, the native Z-scores from the OPLS-AA/SGB calculations are significantly better for most of the cases than the results obtained with the knowledge-based potential. (Compare Tables 2, 3, 4, and 5.) For the native-like Z-scores, however, the differences are not really significant. This suggests that the pairwise and combined scoring functions can separate native-like structures from the non-native-like ones about as well as the OPLS-AA/SGB model. This is supported by the Spearman rank–order coefficients for the pairwise and combined scores which do not significantly differ from those for the OPLS-AA/SGB energies. The surface scores, however, do not correlate well overall with structural similarity.

| PDB name | $U_{\mathrm{native}}$ | $\min(\Delta U)$ | rmsd | $Z_{\mathrm{nat}}$ | $\overline{Z}_{\mathrm{nat-like}}$ | $r_S$ |
|---|---|---|---|---|---|---|
| 1ctf | 3.88 | -0.64 | 1.69 | 2.64 | 1.07 | 0.67 |
| 1r69 | 5.70 | -0.72 | 3.33 | 3.08 | 0.99 | 0.67 |
| 1sn3 | 7.49 | -1.34 | 5.40 | 3.78 | 1.07 | 0.41 |
| 2cro | 5.48 | -0.77 | 1.34 | 3.17 | 0.86 | 0.66 |
| 3icb | 4.95 | -0.04 | 1.83 | 2.09 | 1.31 | 0.79 |
| 4pti | 7.88 | -2.14 | 2.53 | 4.56 | 0.96 | 0.47 |
| 4rxn | 4.20 | 0.93 | 5.61 | 1.56 | 1.11 | 0.47 |

**Table 3.** ProCeryon pairwise scoring function results: the minimized energy, $U_{\mathrm{native}}$, of the native conformation; the energy gap, $\min(\Delta U)$ and the rms deviation between the best scoring decoy and the native conformation; the native Z-score, $Z_{\mathrm{nat}}$, the average Z-score, $\overline{Z}_{\mathrm{nat-like}}$, of the native-like conformations in the Park & Levitt decoy sets [54] and the Spearman rank–order correlation coefficient, $r_S$, for the energies and rms's of all structures in the decoy sets.

| PDB name | $U_{\mathrm{native}}$ | $\min(\Delta U)$ | rmsd | $Z_{\mathrm{nat}}$ | $\overline{Z}_{\mathrm{nat-like}}$ | $r_S$ |
|---|---|---|---|---|---|---|
| 1ctf | 7.62 | 0.63 | 5.38 | 2.22 | 1.20 | 0.60 |
| 1r69 | 6.66 | 1.05 | 2.45 | 1.97 | 1.02 | 0.53 |
| 1sn3 | 4.89 | 2.00 | 5.45 | 1.77 | 0.73 | 0.33 |
| 2cro | 4.50 | 1.90 | 1.88 | 1.31 | 0.63 | 0.52 |
| 3icb | 7.89 | -0.07 | 2.15 | 2.20 | 1.16 | 0.73 |
| 4pti | 4.10 | 1.21 | 5.44 | 1.72 | 0.08 | 0.20 |
| 4rxn | 6.76 | 0.63 | 1.74 | 2.25 | 0.92 | 0.42 |

**Table 4.** ProCeryon surface scoring function results: the minimized energy, $U_{\mathrm{native}}$, of the native conformation; the energy gap, $\min(\Delta U)$ and the rms deviation between the best scoring decoy and the native conformation; the native Z-score, $Z_{\mathrm{nat}}$, the average Z-score, $\overline{Z}_{\mathrm{nat-like}}$, of the native-like conformations of the Park & Levitt decoy sets [54] and the Spearman rank–order correlation coefficient, $r_S$, for the energies and rms's of all structures in the decoy sets.

| PDB name | $U_{\mathrm{native}}$ | $\min(\Delta U)$ | rmsd | $Z_{\mathrm{nat}}$ | $\overline{Z}_{\mathrm{nat-like}}$ | $r_S$ |
|---|---|---|---|---|---|---|
| 1ctf | 7.39 | -0.37 | 2.34 | 2.56 | 1.25 | 0.67 |
| 1r69 | 8.74 | -0.21 | 2.45 | 2.65 | 1.09 | 0.64 |
| 1sn3 | 7.64 | -0.49 | 5.40 | 2.95 | 0.98 | 0.40 |
| 2cro | 7.32 | 0.52 | 2.05 | 2.63 | 0.86 | 0.69 |
| 3icb | 7.93 | -0.55 | 1.87 | 2.30 | 1.31 | 0.81 |
| 4pti | 7.59 | -1.08 | 3.47 | 3.63 | 0.59 | 0.40 |
| 4rxn | 6.98 | 0.56 | 6.29 | 2.24 | 1.11 | 0.48 |

**Table 5.** ProCeryon combined scoring function results: the minimized energy, $U_{\mathrm{native}}$, of the native conformation; the energy gap, $\min(\Delta U)$ and the rms deviation between the best scoring decoy and the native conformation; the native Z-score, $Z_{\mathrm{nat}}$, the average Z-score, $\overline{Z}_{\mathrm{nat-like}}$, of the native-like conformations of the Park & Levitt decoy sets [54] and the Spearman rank–order correlation coefficient, $r_S$, for the energies and rms's of all structures in the decoy sets.

| PDB name | $U_{\mathrm{native}}$ | $\min(\Delta U)$ | rmsd | $Z_{\mathrm{nat}}$ | $\overline{Z}_{\mathrm{nat-like}}$ | $r_S$ |
|---|---|---|---|---|---|---|
| 1ctf | -2795.74 | +43.68 | 6.49 | -2.62 | -0.51 | 0.40 |
| 1r69 | -2489.72 | +76.49 | 1.65 | -3.03 | -0.42 | 0.36 |
| 1sn3 | -2495,10 | +0.04 | 1.42 | -3.10 | -0.59 | 0.29 |
| 2cro | -1122.06 | -35.12 | 0.93 | -2.37 | -0.68 | 0.60 |
| 3icb | -2795.74 | -282.69 | 1.19 | -0.63 | -0.84 | 0.54 |
| 4pti | -1324.06 | +37.53 | 6.21 | -2.97 | -0.71 | 0.25 |
| 4rxn | -3581.88 | -8.95 | 1.60 | -2.47 | -1.13 | 0.60 |

**Table 6.** Vacuum OPLS-AA results: the minimized energy, $U_{\mathrm{native}}$, of the native conformation; the energy gap, $\min(\Delta U)$ and the rms deviation between the best scoring decoy and the native conformation; the native Z-score, $Z_{\mathrm{nat}}$, the average Z-score, $\overline{Z}_{\mathrm{nat-like}}$, of the native-like conformations of the Park & Levitt decoy sets [54]. and the Spearman rank–order correlation coefficient, $r_S$, for the energies and rms's of all structures in the decoy sets.

### 3.3 OPLS-AA Vacuum Calculations

It is instructive to evaluate the importance of each component of the OPLS-AA/SGB energy function in recognizing native conformations. Because all the decoys are well packed, there is very little discrimination based on packing, as measured by the van der Waals energies, of the non-native states from the near-native conformations. In order to establish the role of intramolecular and solvent electrostatic interactions, we have calculated the energy scores in vacuum, $U_{\mathrm{tot}}^{\mathrm{vac}}$, using the same protocol used for the calculations in continuum solvent. The results are summarized in Table 6. For several proteins the native conformation does not correspond to the minimum energy and decoys with large rmsd from the native have very favorable scores. The native Z-score and the near-native average Z-scores have also significantly degraded (compare Tables 2 and 6). This can be clearly seen in Figure 3 showing the energy rmsd correlation plots for the 7 proteins studied. And this is indicated by the large drop in the Spearman rank–order coefficients for the vacuum energies as compared to those from calculations with a continuum solvent. The gain achieved by including the solvation term is particularly noticeable for the 3icb dataset. Figure 4 shows the distribution of energy gaps from the native for the 3icb decoys using either the vacuum OPLS-AA energy or the OPLS-AA/SGB energy. A shift of the distribution to positive values indicates that no decoy structures have energies lower than the native structure. Vacuum energies are scattered above and below the native state energy with little correlation between energy and structural similarity as indicated by the low Spearman rank–order coefficients shown in Table 6. The OPLS-AA/SGB energies produce a sharper distribution than the vacuum energies. It is clear that for this decoy set the vacuum energy is significantly poorer than the energy in solution in discriminating native folds.

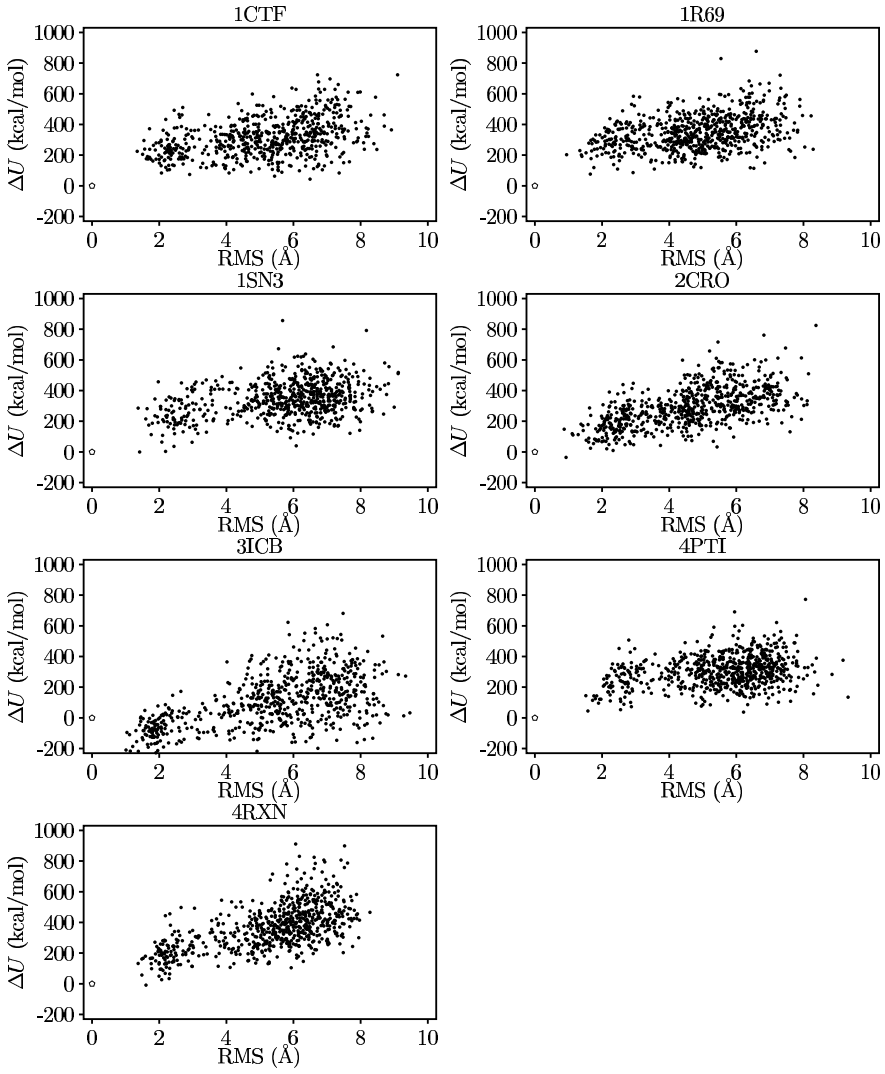


**Fig. 3.** Vacuum OPLS-AA : energy gap/RMS correlation plots for the Park & Levitt decoy sets.

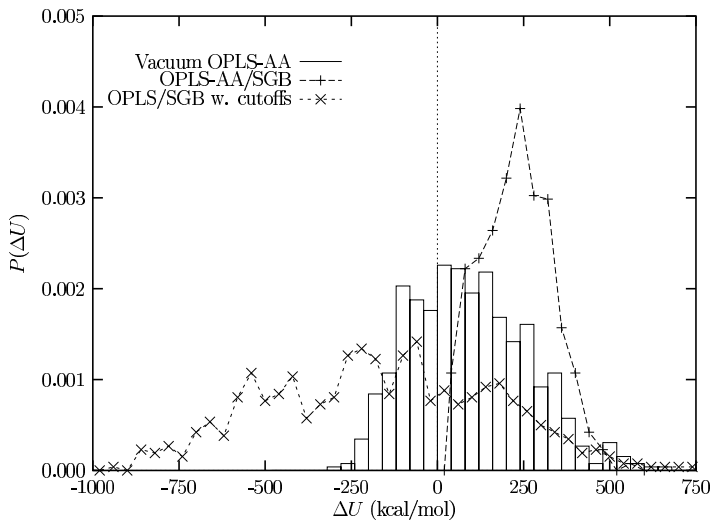


**Fig. 4.** The distribution of energy gaps from the native for the 3icb dataset of the Park & Levitt decoys using various energy functions.

An important contribution to protein stability arises from the tendency for packing non-polar side-chains in the interior of the proteins and placing polar residues on the solvent exposed surface of the protein [43,44,76,77]. These tendencies are not represented well by the intramolecular potential in vacuum which in general ranks equally the strength of interaction between two non-polar residues and between a non-polar residue and polar residue, and does not particularly favor the placement of a polar residue on the protein surface. The solvation energy calculated using the SGB model, however, reproduces hydrophobic interactions and favors the placement of polar residues on the protein surface where they can interact strongly with the solvent. The presence of a hydrophobic core and a polar surface is a key feature of the native protein conformation in solution. Several empirical scoring function have been designed to recognize these features [20,54,66,67,63]. A model that does not take into account solvation effects is likely to perform poorly in native fold recognition among large numbers of compact decoys.

Another important function of dielectric continuum models is to dampen the strength of the electrostatic interactions between polar and charged residues. Conformations having salt bridges and intramolecular hydrogen bonds are strongly favored in vacuum but much less so in solution. The SGB implicit solvent model provides a mechanism to filter out non-native conformations with artificially low intramolecular electrostatic energies that would be otherwise given a favorable score.

In these calculations all charged interactions are included in the total energy; employing a cut-off for atom-atom interactions destroys the correlation between low energy values and native-like structures. Figure 4 shows that the proper evaluation of the long-range Coulomb interactions is crucial in selecting native conformations. If the electrostatic interactions are spatially truncated many non-native structures assume lower total energies than the native structure. As shown in Fig. 4, the correlation between energy and structural similarity using the OPLS-AA/SGB force field with a non-bonded cut-off of 9 Å is poor. This is a direct consequence of neglecting the long-range part of Coulomb interactions, and is aggravated by the highly charged nature of some of the proteins examined (see Table 1).

### 3.4 Energy Components

The ability of a scoring function to discriminate between native and non-native conformations depends on the delicate balance between the components of the scoring function [2,20,54,67,63]. As described in this section, we find that, although some combinations of energy components show improvement over each individual component, the total OPLS-AA/SGB energy is the best scoring function overall.

An analysis of the energy components of Eqs. 1 and 2 presented in Figure 5 shows that for the Park & Levitt dataset (Table 1), containing only well packed structures, the van der Waals energy difference with respect to the native is positive for most of the decoys. The van der Waals energy, however, does not strongly correlate with structural similarity to the native. This point is illustrated in Figure 6 that shows the distribution of energy gaps from the native of both the native-like (rmsd $< 3$ Å) and misfolded (rmsd $> 3$ Å) 3icb decoys. In contrast, the discriminating power of the total OPLS-AA/SGB energy is indicated by the relatively small overlap between the native-like and misfolded distributions of energy gaps (see Fig. 6). A similar separation is not achieved with the van der Waals energy, indicating that the van der Waals energy alone does not provide good discrimination when used as a scoring function.

The electrostatic energy components, the intramolecular Coulomb energy and the solvation energy, taken individually, are not effective scoring functions; the sum of the two, however, is significantly better as indicated in Figs. 7 and 8 ($\epsilon_w = 1$ distribution). As shown in Fig. 7, the solvation energy is strongly anti-correlated with the electrostatic energy. A positive intramolecular electrostatic energy gap from the native is counteracted by a negative solvation energy gap, and vice versa. Because the solvation energy does not completely offset the intramolecular electrostatic energy, decoys having an intramolecular electrostatic energy less favorable than the native will generally continue to have a less favorable total electrostatic energy (intramolecular + solvation) with respect to the native. The contribution of the solvation energy term, however, is large enough to reverse the sign of the energy gap for those

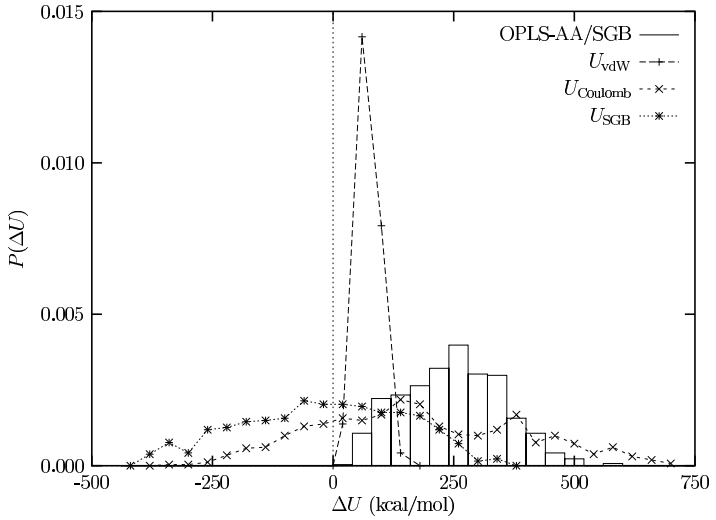


**Fig. 5.** Distribution of energy gaps from the native of the 3icb Park & Levitt decoys for the total OPLS-AA/SGB energy and for the van der Waals, $U_{\mathrm{vdW}}$, intramolecular Coulomb, $U_{\mathrm{Coulomb}}$, and solvation, $U_{SGB}$, energy components.

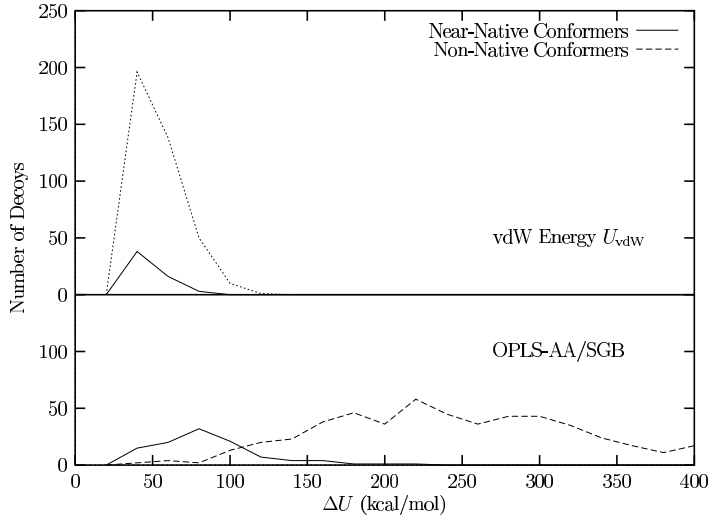


**Fig. 6.** Near-native and non-native distributions of the OPLS-AA/SGB and van der Waals energy gaps from the native for the Park & Levitt 3icb decoys.

decoys having an intramolecular energy more favorable than the native, for which there are many examples in the Park & Levitt set (see Fig. 8). The native state corresponds to a balance between optimizing the intramolecular Coulomb interactions and the intermolecular protein-solvent interactions.

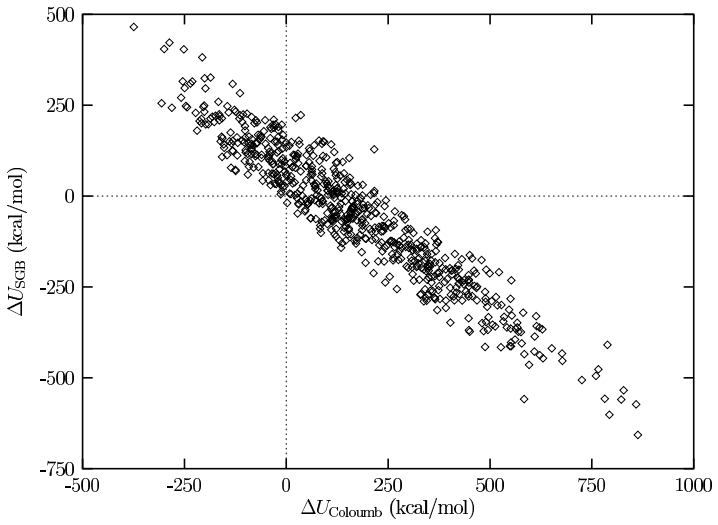


**Fig. 7.** Correlation plot between the intramolecular Coulomb energy gap $\Delta U_{\mathrm{Coulomb}}$ and the solvation energy gap $\Delta U_{\mathrm{SGB}}$ for the 3icb decoys.

Monge et al. [20] have also studied various energy decompositions of an all-atom force field supplemented by a continuum solvation model. They analyzed a decoy dataset generated by a simplified model employing fixed known secondary structure. The authors observe that the relative differences of both van der Waals and Coulomb energies are about 1–2% above the native values, but the total electrostatic component is the dominant factor in distinguishing non-native states from the native ones. They found that a fraction of the decoys had vdW energies lower than the native. Their model performed reasonably well though some non-native conformations had better scores than the native state. This was not observed in the datasets we studied using the OPLS-AA/SGB scoring function.

### 3.5 Differences in Energy Components between the Native and Best-ranked Decoy

A striking result of the OPLS-AA/SGB potential as seen in Table 2 and Figure 1 is how well separated in energy the native conformation is from the decoys. The obvious question is what terms of the potential are most

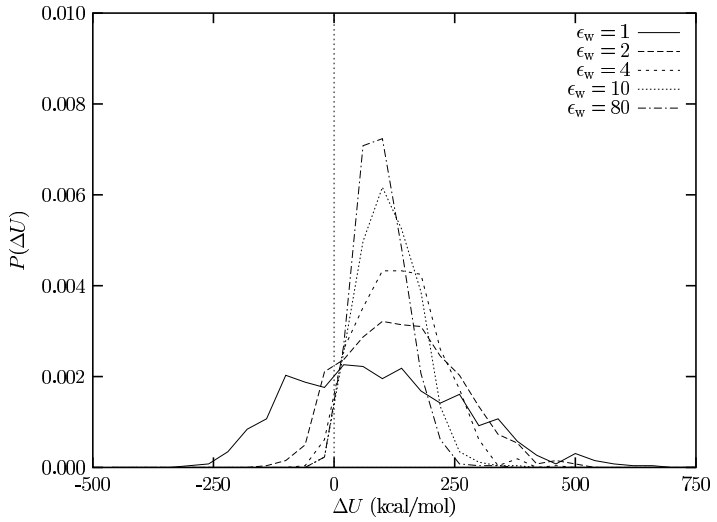


**Fig. 8.** The distributions of the screened Coulomb OPLS-AA energy gaps from the native for the 3icb decoys as a function of dielectric constant.

responsible for stabilizing the native conformation. This can have implications for ab initio protein folding since it may indicate what aspect of the decoy conformation would need to be corrected for it to adopt a more native-like fold. In Table 7, the differences in the minimized energies for each term of the OPLS-AA/SGB potential between the best ranking decoys and the native are shown. Three of the potential energy terms did not significantly contribute to the total difference in energy between the native and the decoy: the bond stretch energy, the 1–4 Lennard–Jones energy, and the SGB cavity energy. Examples where the other terms contribute significantly are spread throughout. Most notably are the non-bonded electrostatic and the solvation reaction field terms. These two terms, however, appear to cancel each other. Because of this observation, various terms were combined and compared. The two following combinations were made: the total "bonded" energy comprised of those terms which involve atoms directly bonded together or connected to form an angle or torsion angle and the total electrostatic energy comprised of the direct Coulomb and solvation terms. The non-bonded Lennard–Jones term was left as is. These combinations are shown in Table 8. For five (1ctf, 1r69, 1sn3, 4pti, and 4rxn) of the seven proteins, these three terms contribute roughly evenly to the total difference. The two combinations account each for at least 31% on average of the total difference while the Lennard–Jones term accounts for about 38%. For these five cases, a balance of the potential components is apparently the key to the native's stability relative to the decoys. For the other two cases, the decoy of 2cro has a smaller difference

in the "bonded" energy term to the native; therefore, the Lennard–Jones and the combined electrostatic energies each account for more of the difference but in roughly equal amounts. For the best–ranked decoy of 3icb, which has the smallest difference to the native for any protein, the two combined components (the "bonded" and total electrostatic) stabilize it more than the native; the Lennard–Jones term, however, destabilizes this structure considerably. The potential terms of 3icb balance each other in a way such that, again, the native is more stable than even the best–ranked decoy.

The importance of balancing potential terms can be seen by examining those energy terms from the minimizations using the OPLS-AA vacuum potential. In Table 9, the combined "bonded" energy (as described above), the non-bonded Lennard–Jones, and the direct Coulomb energy for the best ranking decoys of the seven proteins based on the vacuum potential are compared. Three of the seven proteins have decoys with significantly lower energies than the native. From the table, the term which favors the best–ranked decoy over the native consistently is the direct Coulomb. In the OPLS-AA/SGB potential, the opposite is observed where the direct Coulomb term generally favors the native while the solvation term favors the decoy. For the above structures with a more favorable direct Coulomb contribution, an added SGB solvation energy is significantly greater than the native's. A balance of energy terms is again acheived (though the trends are reversed for the direct Coulomb and solvation energies) such that the total energy favors the native.

| PDB | $\Delta U$ | Differences in Components of the OPLS-AA/SGB Potential | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| name | | Bond | Angle | Torsion | LJ-1,4 | El-1,4 | LJ | Coul. | Rxn-Fld | Cav. |
| 1ctf | 65.55 | -1.24 | -7.85 | 6.49 | -4.33 | 31.45 | 16.61 | 247.25 | -222.86 | 0.03 |
| 1r69 | 107.16 | 1.27 | 0.38 | 1.31 | -1.76 | 23.51 | 49.81 | 225.11 | -192.33 | -0.12 |
| 1sn3 | 96.08 | 1.31 | 17.34 | -9.73 | -4.03 | 29.65 | 41.89 | 246.46 | -227.41 | 0.60 |
| 2cro | 72.55 | -2.19 | -1.20 | 9.01 | -0.28 | 2.91 | 30.12 | 29.52 | 4.79 | -0.13 |
| 3icb | 18.08 | -1.37 | -13.01 | -7.01 | -4.79 | 1.76 | 59.06 | 2.41 | -19.11 | 0.13 |
| 4pti | 105.07 | 3.11 | 28.33 | -13.11 | -4.47 | 11.34 | 41.41 | 168.78 | -130.43 | 0.11 |
| 4rxn | 92.06 | 1.18 | 4.43 | 36.09 | -4.67 | -5.22 | 31.50 | -20.90 | 49.75 | -0.12 |

**Table 7.** Differences in the OPLS-AA/SGB Potential Components: the total difference, $\Delta U$, and the differences between the various terms of the OPLS-AA/SGB potential are given between the best ranked decoy and the native. Those terms are the bond stretch energy (labelled "Bond" in the table), the angle bending energy ("Angle"), the torsion angle energy ("Torsion"), the Lennard–Jones energy between atoms separated by three bonds ("LJ-1,4"), the electrostatic energy between atoms separated by three bonds ("El-1,4"), the Lennard–Jones energy between non-bonded atoms ("LJ"), the the direct Coulomb energy ("Coul."), the SGB reaction-field energy ("Rxn-Fld") and the energy required to form the cavity in the solvent that the protein will fill ("Cav.").

| PDB name | $\Delta U$ | Combinations of OPLS-AA/SGB Components | | | rmsd |
|---|---|---|---|---|---|
| | | $\Delta U_{\text{bonded}}$ | $\Delta U_{\text{LJ}}$ | $\Delta U_{\text{el.}}$ | |
| 1ctf | 65.55 | 24.52(37.4) | 16.61(25.3) | 24.42(37.3) | 1.69 |
| 1r69 | 107.16 | 24.70(23.0) | 49.80(46.5) | 32.66(30.5) | 2.30 |
| 1sn3 | 96.08 | 34.54(35.9) | 41.89(43.6) | 19.65(20.5) | 2.19 |
| 2cro | 72.55 | 8.25(11.4) | 30.12(41.5) | 34.18(47.1) | 0.94 |
| 3icb | 18.08 | -24.41(-135.0) | 59.06(326.7) | -16.57(-91.6) | 1.84 |
| 4pti | 105.07 | 25.20(24.0) | 41.41(39.4) | 38.46(36.6) | 1.89 |
| 4rxn | 92.06 | 31.83(34.6) | 31.50(34.2) | 28.73(31.2) | 2.16 |

**Table 8.** Differences in Combinations of the OPLS-AA/SGB Potential Components: the total difference, $\Delta U$, and the differences between the combinations of various terms of the OPLS-AA/SGB potential are given between the best ranked decoy and the native along with the rms deviation between the decoy and the native. The first combination, $\Delta U_{\text{bonded}}$, is composed of the "bonded" energy terms: "Bond", "Angle", "Torsion", "LJ-1,4", and "El-1,4" described in Table 7. The non-bonded Lennard–Jones energy, $\Delta$LJ, is presented by itself. And the second combination, $\Delta U_{\text{el.}}$, is composed of the direct Coulomb and solvation energies: "Coul", "Rxn-Fld", and "Cav." in Table 7. The percentage that each term contributes to the total difference is given to the left of each value in parantheses.

| PDB name | $\Delta U$ | Combinations of OPLS-AA Components | | | rmsd |
|---|---|---|---|---|---|
| | | $\Delta U_{\text{bonded}}$ | $\Delta U_{\text{LJ}}$ | $\Delta U_{\text{Coul.}}$ | |
| 1ctf | 43.68 | 34.35 | 85.76 | -76.43 | 6.49 |
| 1r69 | 76.49 | 28.20 | 53.23 | -4.94 | 1.65 |
| 1sn3 | 0.04 | 178.57 | 78.52 | -257.05 | 1.42 |
| 2cro | -35.12 | 8.67 | 55.13 | -98.92 | 0.93 |
| 3icb | -282.69 | 23.05 | 54.20 | -359.94 | 1.19 |
| 4pti | 37.53 | -14.30 | 63.53 | -11.70 | 6.21 |
| 4rxn | -8.95 | 14.89 | 58.64 | -82.48 | 1.60 |

**Table 9.** Differences in Combinations of the OPLS-AA Vacuum Potential Components: the total difference, $\Delta U$, and the differences between the combinations of "bonded" terms, the non-bonded Lennard–Jones term and the direct Coulomb term of the OPLS-AA potential are given between the best ranked decoy and the native along with the rms deviation between the decoy and the native. The combination, $\Delta U_{\text{bonded}}$, is composed of the "bonded" energy terms: "Bond", "Angle", "Torsion", "LJ-1,4", and "El-1,4" described in Table 7. The non-bonded Lennard–Jones energy, $\Delta$LJ, and the non-bonded electrostatic energy, $\Delta U_{\text{Coul.}}$, are presented as is.

### 3.6 Approximate Effective Dielectric Models

**Screened Coulomb Approximation** As shown in Fig. 7 the solvation energy gaps with respect to the native are strongly correlated with the intramolecular Coulomb energy gaps. The equation

$$\Delta U_{\text{SGB}} = \alpha + \beta \Delta U_{\text{Coulomb}} \tag{9}$$

can be fitted obtaining $\beta = -0.82$ with a regression coefficient of 0.94. If we collate the total electrostatic interaction energy $\Delta U_{\rm ele}$ as the sum of the Coulomb and solvation energies we find

$$\Delta U_{\rm ele} \equiv \Delta U_{\rm Coulomb} + \Delta U_{\rm SGB} \cong 0.18 \Delta U_{\rm Coulomb}. \tag{10}$$

This suggests that it might be possible to employ a screened Coulomb model to account for solvation effects.

The screened Coulomb effective electrostatic interaction between two charges $q$ a distance $r$ apart is

$$U_{\rm Coulomb}(r)/\epsilon_{\rm w} = \frac{q^2}{\epsilon_{\rm w} r}. \tag{11}$$

The effect of the surrounding medium is accounted for by the value of $\epsilon_{\rm w}$, usually taken as 80 for water. Figure 8 shows the energy distributions for the 3icb decoy set relative to the native state for the vacuum case and for various values of the effective dielectric constant. A good energy function should only produce energy gaps values in the positive range. It is clear that for this decoy set, a simple electrostatic energy evaluation in vacuum ($\epsilon_{\rm w} = 1$) results in many decoy structures with energies substantially below the native values. Moreover, no correlation between the rmsd from the native and the energy is observed. Increasing the value of the effective dielectric constant removes some of the negative energy gaps and increases the propensity for the low energy decoy structures to have low rmsd (not shown). None of the effective dielectric constants used, however, was able to differentiate all of the decoys from the native structure. This point is also illustrated in Figure 2 depicting the fraction of native-like structures with energy gaps from the native less than $\Delta U$ using $\epsilon_{\rm w} = 5.5$ as suggested by the relation in Eq. 10. It is clear that the screened Coulomb scoring function provides less discrimination between decoys and native structures than the SGB solvation model.

If a simple relationship between the reaction field energy calculated via the SGB model and the Coulomb energy as in Eq. 11 could be found, there would be no need to employ more complicated continuum models. Although the bulk of the correlation between these two terms can be explained by a screened Coulomb interaction, the discrimination between native and non-native states is degraded by such an approximation. The dispersion in the reaction field energy versus the Coulomb energy, which is not contained in the screened Coulomb model, provides a more detailed description of solvation effects which aids the discrimination of native-like conformations from misfolded ones.

Although the SGB solvation energy is correlated with the intramolecular Coulomb energy, it is not clear that the best values to use for an effective dielectric constant is given by Eq. 10. The fraction of native-like structures with energy gap less than a given energy difference calculated over all the data sets in Table 1, reported in Fig. 9, shows the efficiency achieved using

different values of $\epsilon_w$. None of the effective dielectric models achieves 100% discrimination for energy values within 20 kcal/mol of the native state energy. Using $\epsilon_w = 1$ yields a broad range of energies for both native-like and non-native states as discussed above. In comparison, using either $\epsilon_w = 5.5$ or 80.0, yields distributions of energies that are like those given in Figure 8 for the calbindin dataset. The fraction of native-like structures with energies similar to the native state is around 60% for an effective dielectric constant of 80.0. This fraction increases to about 75% for $\epsilon_w = 5.5$.

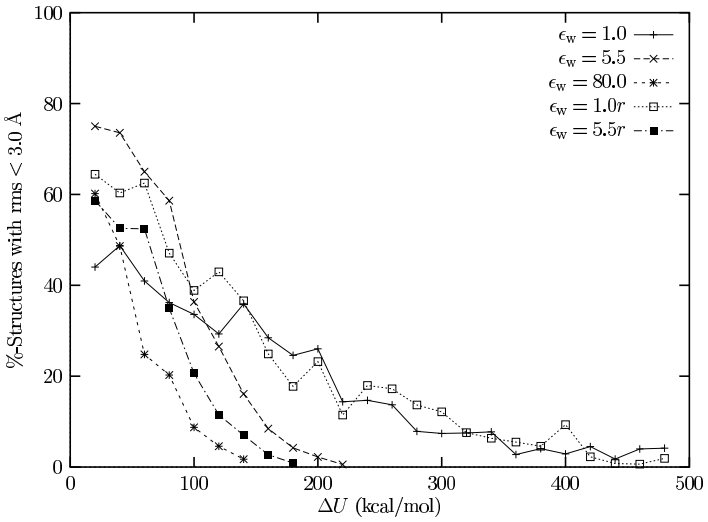


**Fig. 9.** Fraction of the Park & Levitt decoys with energy gap from the native less than $\Delta U$ which are native-like (rmsd from native < 3 Å), using the vacuum OPLS-AA potential with screened Coulomb interactions.

**Distance Dependent Dielectric Approximation** An alternative to the simple screened Coulomb interaction in protein modeling is the distance dependent dielectric function [51]. In this approximation the effective electrostatic interaction between two partial charges $q$ at distance $r$ is written as,

$$U_{\text{Coulomb}}(r)/\epsilon_w r = \frac{q^2}{\epsilon_w r^2}. \tag{12}$$

Although unphysical in nature, it has been suggested that the extra screening afforded by the $1/r^2$ function can capture some of the additional polarization effects contained in higher level implicit solvent models [51]. By calculating the energies of the decoy conformers in Table 1 using the distance dependent

dielectric approximation, we obtain energy distributions similar to those obtained using the simple screened Coulomb model. Moreover, as shown in Fig. 9, both effective dielectric models produce qualitatively similar results. For both values of $\epsilon_w$ studied, 1.0 and 5.5, the fraction of native-like structures with energy similar to the native energy is significantly less than 100%. Comparison between the distance dependent dielectric and the non distance dependent dielectric function in Figure 9 and Figure 2 demonstrate that the distance dependent function is less discriminatory for the decoy datasets studied here. While the distance dependent dielectric constant has been successfully employed in some cases[78], we find that, though it is better than the vacuum Coulomb potential, a simple non-distance dependent screened Coulomb model is more effective (Fig. 9). None of the screened Coulomb models are as effective as the SGB solvation potential for the protein decoy data sets investigated here.

| PDB | $\epsilon_{in}$ | $U_{total}^{native}$ | $U_{vdW}^{native}$ | $U_{Coulomb}^{native}$ | $U_{SGB}^{native}$ | $U_{cav}^{native}$ |
|---|---|---|---|---|---|---|
| | | kcal/mol | kcal/mol | kcal/mol | kcal/mol | kcal/mol |
| 1ctf | 1.0 | -4213.9 | -475.5 | -5340.3 | -1367.6 | +37.9 |
| | 2.0 | -2065.9 | -519.7 | -2595.2 | -688.3 | +38.4 |
| | 5.5 | -730.6 | -532.8 | -925.5 | -244.0 | +38.7 |
| 1r69 | 1.0 | -3499.5 | -497.2 | -3722.9 | -1168.9 | +37.2 |
| | 2.0 | -1709.9 | -539.0 | -1781.7 | -593.3 | +37.7 |
| | 5.5 | -599.5 | -554.3 | -627.9 | -210.8 | +38.1 |
| 1sn3 | 1.0 | -3467.5 | -465.1 | -4784.2 | -972.8 | +36.3 |
| | 2.0 | -1688.1 | -499.8 | -2315.2 | -500.3 | +36.8 |
| | 5.5 | -585.3 | -511.8 | -821.5 | -180.1 | +37.1 |
| 2cro | 1.0 | -3628.3 | -522.4 | -3514.8 | -1462.2 | +40.4 |
| | 2.0 | -1763.1 | -567.2 | -1662.8 | -749.7 | +41.0 |
| | 5.5 | -604.8 | -578.9 | -585.2 | -264.8 | +41.4 |
| 3icb | 1.0 | -4694.5 | -587.3 | -5163.5 | -2350.6 | +45.4 |
| | 2.0 | -2271.4 | -641.0 | -2466.5 | -1195.6 | +46.1 |
| | 5.5 | -766.8 | -656.8 | -865.7 | -427.2 | +46.4 |
| 4pti | 1.0 | -3055.0 | -423.9 | -2542.0 | -1366.9 | +34.1 |
| | 2.0 | -1464.2 | -448.4 | -1208.6 | -686.6 | +34.6 |
| | 5.5 | -473.2 | -455.1 | -425.0 | -240.9 | +34.8 |
| 4rxn | 1.0 | -3363.5 | -373.6 | -2496.6 | -2791.5 | +31.3 |
| | 2.0 | -1598.8 | -399.3 | -1190.1 | -1389.9 | +31.6 |
| | 5.5 | -498.1 | -407.6 | -410.9 | -489.1 | +31.8 |

**Table 10.** Selected energy components from Eqs. 1- 2 for the native state using the continuum model ($\epsilon_w = 80.0$) as a function of interior dielectric constant, $\epsilon_{in}$

### 3.7 Dependence on the Interior Dielectric Constant

The SGB solvent model requires the separation of space into an exterior region containing the solvent medium and an interior region containing the protein charge distribution. In the current implementation of the SGB model the van der Waals surface of the protein is used to define the dividing surface. The default value for the dielectric constant of the solvent is 80, corresponding to pure water at room temperature. The dielectric constant of the interior region, $\epsilon_{\text{in}}$, has been up to this point set at the value of 1, corresponding to the vacuum dielectric constant. We have also examined the cases $\epsilon_{\text{in}} = 2$ and 5.5 to see whether the OPLS-AA/SGB results can be further improved. The energy components obtained for the native conformations contained in the Park & Levitt set are given in Table 10. A larger interior dielectric constant results in a lower total energy of the system due to the increase of the dielectric shielding inside the protein. The Coulomb energy and the reaction field contributions are both reduced in an amount roughly proportional to the interior dielectric constant. The van der Waals energy partly compensates for the reduction in electrostatic energy, but the variation in $U_{\text{vdW}}^{\text{native}}$ is relatively small.

The fraction of native-like decoys of the Park & Levitt set as a function of energy gap is shown in Fig. 10 for the values of $\epsilon_{\text{in}}$ examined. The number of native-like conformations (rmsd $< 3$ Å) with an energy score similar to the native increases as we decrease the dielectric constant of the interior region. It is only with an interior dielectric of 1.0 that all misfolded conformations can be eliminated based on energy alone. The discriminatory power of the OPLS-AA/SGB energy model in this fold recognition test is optimal for this choice of the internal dielectric, though it may not be optimal in other modeling contexts.

## 4 Conclusions

The OPLS-AA molecular mechanics energy function coupled with the Surface Generalized Born solvation model is found to be able to discriminate the native structures of several proteins from their decoys. The results show that for a number of cleverly constructed decoys the OPLS-AA/SGB scoring function correctly singles out native-like states from the bulk of the non-native conformations. Not all of the native-like structures were clearly separated in the datasets, indeed some distant non-native conformations score better than some native-like (rmsd $< 3$ Å) conformations. This suggests that if the current scoring method is to be applied to a set of ab initio generated structures it is critical that the algorithm for constructing native-like structures be such that a broad range of the relevant parts of the native-like conformational space are sampled.

The ability of the OPLS-AA/SGB model to recognize native conformations is found to be comparable and in many cases superior to the best

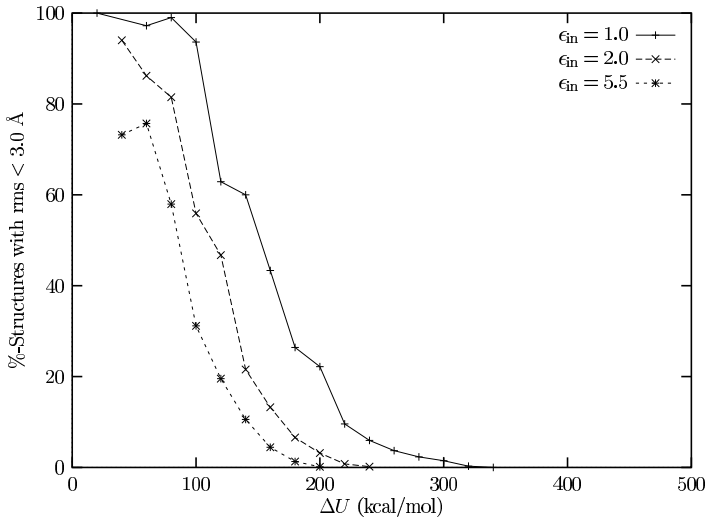


**Fig. 10.** Fraction of the Park & Levitt decoys with energy gap from the native less than $\Delta U$ which are native-like (rmsd from native $< 3$ Å), using the OPLS-AA/SGB potential with various values of the interior dielectric constant.

knowledge-based scoring functions. Other studies have shown the usefulness of molecular mechanics force fields augmented by implicit solvation models in this area[1]. Lazaridis and Karplus [22] have shown that the CHARMM protein force field combined with their EEF1 effective solvation free energy model [79] is able to achieve 100% discrimination of the native conformations in a large decoy dataset and in the single decoy dataset they examined. They also observe, in agreement with our findings, that significantly poorer results are obtained by omitting the solvation free energy term. They obtain these results despite the use of a computationally fast solvation model which has the form of an effective pair potential and is simpler than the SGB solvation model. Recently, Petrey and Honig [80] have applied the CHARMM protein force field together with a dielectric continuum model based on the Poisson-Boltzmann equation, to the problem of native fold recognition in the single decoy dataset [17] achieving a discrimination level close to 100%. They also applied a simplified solvation model containing only the intramolecular electrostatic energy and a hydrophobic residue burial estimator to evaluate the Park & Levitt decoy sets. In two cases (3icb and 4rxn) their method does not clearly rank the X-ray conformation favorably. Petrey and Honig observe that the solvation energy often favors the misfolded conformation in the single decoy sets concluding that the solvation energy is not useful in recognizing the native conformation.

Their findings are in contrast to the analysis and comparison of the OPLS-AA/SGB potential terms of the natives and the best–ranked decoys. Even though the solvation energy generally favors the misfolded conformations, these structures are disfavored relative to the native conformation when the total electrostatic energy (the sum of the direct Coulomb and solvation terms) is considered. Plus, the other terms of the potential, the Lennard–Jones non-bonded term and the "bonded" terms, contribute further towards stabilizing the native fold relative to the decoys. It is a balance of these potential terms leading to total energies that distinguish the natives from the rest. For instance, in the absence of the SGB solvation term during the minimizations, a large number of the Park & Levitt decoys obtain more favorable direct Coulomb energies relative to the native. Several of the best–ranked decoys, based on the vacuum OPLS-AA potential, end up with a total energy less than the native's due to these Coulomb energies overwhelming the other terms which favor the native fold. But after minimizing with the SGB solvation term, the Coulomb energies of the decoys all become disfavorable relative to the native. This leads to the total energy of the native being the lowest in all cases.

The OPLS-AA/SGB scoring function was also compared with the screened Coulomb OPLS-AA scoring function. Whereas a significant fraction of the decoys with scores within 100 kcal/mol from the native are misfolded using a screened Coulomb potential, essentially all of the decoys within 100 kcal/mol from the native are native-like using the OPLS-AA/SGB scoring function.

The ability to discriminate native-like protein conformations from non-native conformations is one of the fundamental problems in theoretical protein structure prediction. The use of knowledge-based scoring potentials, derived from a combination of structural and thermodynamic data, are currently the most widely used methods. It is often assumed that such models are inherently better than all-atom force-fields. This work shows the importance of correctly modeling the physical forces underlying protein folding. Thanks to their simplicity, knowledge-based scoring schemes are less costly to evaluate compared to all-atom models. In the future it should be possible to combine the best features of the two approaches to rapidly generate plausible protein conformations using knowledge-based potentials more reliably, and then discriminate between conformers using all-atom scoring functions.

## 5 Acknowledgments

This project has been supported by the National Institutes of Health Grant GM-30580, the Center for Biomolecular Simulations at Columbia University, and by the High Performance Computing Project at Rutgers University. The authors thank Dr. Lynne Reed Murphy for help with some of the calculations.

## References


1. Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000,** *10,* 139–145.
2. Wodak, S. J.; Rooman, M. J. *Curr. Opin. Struct. Biol.* **1993,** *3,* 247–259.
3. Jones, D. T.; Thornton, J. M. *Curr. Opin. Struct. Biol.* **1996,** *6,* 210–216.
4. Plaxco, K. W.; Riddle, D. S.; Grantcharova, V.; Baker, D. *Curr. Opin. Struct. Biol.* **1998,** *8,* 80–85.
5. Hao, M.; Scheraga, H. A. *Curr. Opin. Struct. Biol.* **1999,** *9,* 184–188.
6. Osguthorpe, D. J. *Curr. Opin. Struct. Biol.* **2000,** *10,* 146–152.
7. Eyrich, V.; Standley, D.; Felts, A.; Friesner, R. *Proteins: Struct. Funct. Genet.* **1999,** *35,* 41–57.
8. Rick, S. W.; Berne, B. J. *J. Am. Chem. Soc.* **1994,** *116,* 3949–3954.
9. Levy, R. M.; Gallicchio, E. *Annu. Rev. Phys. Chem.* **1998,** *49,* 531–567.
10. Rashin, A. A.; Bukatin, M. A. *J. Phys. Chem.* **1994,** *98,* 386–389.
11. Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994,** *98,* 1978–1988.
12. Tannor, D. J.; Marten, B.; Murphy, R.; Friesner, R. A.; Sitkoff, D.; Nicholls, A.; Ringnalda, M.; Goddard, III, W. A.; Honig, B. *J. Am. Chem. Soc.* **1994,** *116,* 11875–11882.
13. Sitkoff, D.; Ben-Tal, N.; Honig, B. *J. Phys. Chem.* **1996,** *100,* 2744–2752.
14. Hawkins, G.; Cramer, C.; Truhlar, D. *J. Phys. Chem.* **1996,** *100,* 19824–19839.
15. Gallicchio, E.; Zhang, L.; Levy, R. M. **2001,** In preparation.
16. Eisenberg, D.; McLachlan, A. D. *Nature* **1986,** *319,* 199–203.
17. Holm, L.; Sander, C. *J. Mol. Biol.* **1992,** *225,* 93–105.
18. van Gunsteren, W. F.; Luque, F. J.; Timms, D.; Torda, A. E. *Annu. Rev. Biophys. Biomol.Struct.* **1994,** *23,* 847–863.
19. Smith, P. E.; Pettitt, B. M. *J. Phys. Chem.* **1994,** *98,* 9700–9711.
20. Monge, A.; Lathrop, E. J. P.; Gunn, J. R.; Shenkin, P. S.; Friesner, R. A. *J. Mol. Biol.* **1995,** *247,* 995–1012.
21. Schaefer, M.; van Vlijmen, H. W.; Karplus, M. *Advances in Protein Chemistry* **1998,** *51,* 1–57.
22. Lazaridis, T.; Karplus, M. *J. Mol. Biol.* **1999,** *288,* 477–487.
23. Vorobjev, Y. N.; Hermans, J. *Biophys. Chem.* **1999,** *78,* 195–205.
24. Gilson, M. K.; Honig, B. *Proteins: Struct. Funct. Genet.* **1988,** *4,* 7–18.
25. Bashford, D.; Karplus, M. *Biochemistry* **1990,** *29,* 10219–10225.
26. Rashin, A. A. *J. Phys. Chem.* **1990,** *94,* 1725–1733.
27. Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Chem.* **1990,** *19,* 301–332.
28. Warshel, A.; Åqvist, J. *Annu. Rev. Biophys. Chem.* **1991,** *20,* 267–298.
29. Gilson, M. K.; Davis, M. E.; Luty, B. A.; McCammon, J. A. *J. Phys. Chem.* **1993,** *97,* 3591–3600.
30. Honig, B.; Sharp, K.; Yang, A.-S. *J. Phys. Chem.* **1993,** *97,* 1101–1109.
31. Mohan, V.; Davis, M. E.; McCammon, J. A.; Pettitt, B. M. *J. Phys. Chem.* **1992,** *96,* 6428–6431.
32. Simonson, T.; Brünger, A. T. *J. Phys. Chem.* **1994,** *98,* 4683–4694.
33. Ösapay, K.; Young, W. S.; Bashford, D.; Brooks III, C. L.; Case, D. A. *J. Phys. Chem.* **1996,** *100,* 2698–2705.
34. Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. *J. Phys. Chem. B* **1997,** *101,* 1190–1197.
35. Born, M. *Z. Physik* **1920,** *1,* 45–48.
36. Hirata, F.; Rejfern, P.; Levy, R. *J. Quantum Chem.* **1988,** *15,* 179–188.

37. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990,** *112,* 6127–6129.
38. Jean-Charles, A.; Nichols, A.; Sharp, K.; Honing, B.; Tempczyk, A.; Hendrickson, T. F.; Still, W. C. *J. Am. Chem. Soc.* **1991,** *113,* 1454–1455.
39. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997,** *101,* 3005–3014.
40. Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998,** *102,* 10983–10990.
41. Roux, B.; Simonson, T. *Biophysical Chemistry* **1999,** *78,* 1–20.
42. Zhang, L.; Gallicchio, E.; Friesner, R.; Levy, R. M. *J. Comp. Chem.* **2001,** *22,* 591–607.
43. Novotny, J.; Bruccoleri, R.; Karplus, M. *J. Mol. Biol.* **1984,** *177,* 787–818.
44. Novotny, J.; Rashin, A. A.; Bruccoleri, R. *Proteins: Struct. Funct. Genet.* **1988,** *4,* 19–30.
45. Chiche, L.; Gregoret, L. M.; Cohen, F. E.; Kollman, P. A. *Proc. Natl. Acad. Sci. USA* **1990,** *87,* 3240–3243.
46. Vila, J.; Williams, R. L.; Vasquez, M.; Scheraga, H. A. *Proteins: Struct. Funct. Genet.* **1991,** *10,* 199–218.
47. Williams, R. L.; Vila, J.; Perrot, G.; Scheraga, H. A. *Proteins: Struct. Funct. Genet.* **1992,** *14,* 110–119.
48. Wang, Y.; Zhang, H.; Li, W.; Scott, R. A. *Proc. Nat. Acad. Sci. (USA)* **1995,** *92,* 709–713.
49. Wang, Y.; Zhang, H.; Scott, R. A. *Protein Sci.* **1995,** *4,* 1402–1411.
50. Vieth, M.; Kolinski, A.; III, C. L. B.; Skolnick, J. *J. Mol. Biol.* **1994,** *237,* 361–367.
51. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagone, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984,** *106,* 765–784.
52. Vorobjev, Y. N.; Almagro, J. C.; Hermans, J. *Proteins: Struct. Funct. Genet.* **1998,** *32,* 399–413.
53. Dominy, B.; Brooks III, C. *manuscript,* **2001**.
54. Park, B.; Levitt, M. *J. Mol. Biol.* **1996,** *258,* 367–392.
55. Hendlich, M.; Lackner, P.; Weitckus, S.; Floeckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G.; Sippl, M. J. *J. Mol. Biol.* **1990,** *216,* 167–180.
56. Sippl, M. J. *Curr. Opin. Struct. Biol.* **1995,** *5,* 229–235.
57. Jernigan, R. L.; Bahar, I. *Curr. Opin. Struct. Biol.* **1996,** *6,* 195–209.
58. Miyazawa, S.; Jernigan, R. J. *J. Mol. Biol.* **1996,** *256,* 623–644.
59. Wallqvist, A.; Smythers, G. W.; Covell, D. G. *Protein Sci.* **1997,** *6,* 1627–1642.
60. Miyazawa, S.; Jernigan, R. L. *Proteins: Struct. Funct. Genet.* **1999,** *36,* 357–369.
61. Covell, D.; Jernigan, R. *Biochemistry* **1990,** *29,* 3287–3294.
62. Ozkan, B.; Bahar, I. *Proteins: Struct. Funct. Genet.* **1998,** *32,* 211–222.
63. Samudrala, R.; Moult, J. *J. Mol. Biol.* **1998,** *275,* 895–916.
64. Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D. *Proteins: Struct. Funct. Genet.* **1999,** *34,* 82–95.
65. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996,** *118,* 11225–11236.
66. Huang, E. S.; Subbiah, S.; Tsai, J.; Levitt, M. *J. Mol. Biol.* **1996,** *257,* 716–725.
67. Park, B. H.; Huang, E. S.; Levitt, M. *J. Mol. Biol.* **1997,** *266,* 831–846.

68. Kitchen, D. B.; Hirata, F.; Westbrook, J. D.; Levy, R.; Kofke, D.; Yarmush, M. *J. Comp. Chem.* **1990,** *11,* 1169–1180.
69. Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. Protein Data Bank. In *Crystallographic Databases - Information Content, Software Systems, Scientific Applications*; Allen, F. H.; Bergerhoff, G.; Sievers, R., Eds.; Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester: 1987.
70. Zhang, L.; Gallicchio, E.; Levy, R. M. Implicit solvent models for protein-ligand binding: Insights based on explicit solvent simulations. In *Simulation and theory of electrostatic interactions in solution. AIP conference proceedings 492*; Pratt, L. R.; Hummer, G., Eds.; American Institute of Physics: 1999.
71. *ProCeryon – A software package for fold recognition and protein structure analysis, ProCeryon Biosciences;* 1999-2000.
72. Levitt, M. *J. Mol. Biol.* **1992,** *226,* 507–533.
73. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in FORTRAN: The Art of Scientific Computing;* Press Syndicate of the University of Cambridge: Cambridge, second ed.; 1992.
74. Sippl, M. *J. Mol. Biol.* **1990,** *213,* 859–883.
75. Sippl, M. *J. Comput. Aided Mol. Design* **1993,** *7,* 473–501.
76. Dill, K. A. *Biochemistry* **1990,** *29,* 7133–7155.
77. Dill, K. A. *Curr. Opin. Struct. Biol.* **1993,** *3,* 99–103.
78. Schaefer, M.; Bartels, C.; Karplus, M. *Theoretical Chemistry Accounts* **1998,** *101,* 194–204.
79. Lazarids, T.; Karplus, M. *Proteins: Struct. Funct. Genet.* **1999,** *35,* 133–152.
80. Petrey, D.; Honig, B. *Protein Sci.* **2000,** *9,* 2181–2191.