HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2017 July 18.

Published in final edited form as:

J Chem Theory Comput. 2016 December 13; 12(12): 6201–6212. doi:10.1021/acs.jctc.6b00819.

Simultaneous optimization of biomolecular energy function on features from small molecules and macromolecules

Hahnbeom Park[†,‡], **Philip Bradley**[‡,§], **Per Greisen Jr.**[†,‡,‖], **Yuan Liu**[†,‡,∇], **Vikram Khipple Mulligan**[†,‡], **David E. Kim**[‡,⊥], **David Baker**[†,‡,⊥], and **Frank DiMaio**[*,†,‡]

[†]Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA

[‡]Institute for Protein Design, University of Washington, Seattle, Washington 98195, USA

[§]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, Washington 98019, USA

[⊥]Howard Hughes Medical Institute, University of Washington, Box 357370, Seattle, Washington 98195, USA

Abstract

Most biomolecular modeling energy functions for structure prediction, sequence design, and molecular docking, have been parameterized using existing macromolecular structural data; this contrasts molecular mechanics force fields which are largely optimized using small-molecule data. In this study, we describe an integrated method that enables optimization of a biomolecular modeling energy function simultaneously against small-molecule thermodynamic data and high-resolution macromolecular structural data. We use this approach to develop a next-generation Rosetta energy function that utilizes a new anisotropic implicit solvation model, and an improved electrostatics and Lennard-Jones model, illustrating how energy functions can be considerably improved in their ability to describe large-scale energy landscapes by incorporating both small-molecule and macromolecule data. The energy function improves performance in a wide range of protein structure prediction challenges, including monomeric structure prediction, protein-protein and protein-ligand docking, protein sequence design, and prediction of the free energy changes by mutation, while reasonably recapitulating small-molecule thermodynamic properties.

Keywords

structure prediction; biomolecular modeling; energy function; protein design; molecular docking

Corresponding Author dimaio@u.washington.edu.

[‖]**Present Address:** Global Research, Novo Nordisk A/S, DK-2760 Måløv, Denmark

[∇]**Present Address:** Center for Life Science, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

Notes

The authors declare no competing financial interest.

Author Contributions

H.P., P.B., D.B., and F.D. designed research; H.P. and F.D. performed research; P.B., P.G., Y.L., V.K.M., and D.K. contributed new analytic tools; H.P. and F.D. analyzed data; H.P., P.B., D.B., and F.D. wrote the paper.

INTRODUCTION

Accurate biomolecular energy functions are important for a wide range of challenges in computational structural biology, including protein structure prediction, protein structure determination from sparse data, protein design, protein small molecule docking and simulation of protein folding[1–6]. Energy functions play a central role in guiding conformational search and providing quantitative estimation of the likelihood of sampled conformations. Achieving high precision for biomolecular energy functions is one of the long-standing challenges in the area, as other challenges, particularly conformational search, are being addressed with advances in computing technology[7–9].

Full first principle quantum chemistry calculations on macromolecular systems are infeasible, and hence classical mechanics approximations are almost always used in which the energy is written as a sum of generally pairwise additive functions representing different interactions. Two quite different approaches have been taken to setting the values for the parameters in these functions. In the first approach, typically adopted in the development of molecular mechanics force fields, small molecule data is used to guide parameter optimization. The parameters in these molecular mechanics force fields are obtained by fitting thermodynamic and spectroscopic data on small molecules, and by attempting to match quantum chemistry results on small molecule systems[10–12]. The second approach uses the large number of experimentally determined ground states of biomolecules that have been determined by X-ray crystallography and NMR to guide parameter optimization[13–15]. Parameters are set such that the experimentally observed state (both structure and sequence) has lower energy than other states. This approach has been used to develop energy functions for biomolecular modeling studies such as protein structure prediction or docking.

Both approaches have limitations in applicability to macromolecule modeling studies. Molecular mechanics force fields derived by fitting to small molecule properties are limited by the accuracy of the approximations required for computational tractability[16–19]; additionally, it is not clear how transferable an approximate energy model derived from small model systems is to macromolecules[17,20]. However, structure-based approaches also have weaknesses: observed structural data only show relative preference for conformations, and do not give quantitative estimations of energy differences, and in parameter fitting, data sparseness — particularly when large numbers of parameters are optimized — may lead to incorrect estimation and over-fitting. Therefore, it is also obvious that macromolecule data alone is insufficient for deriving an energy function for macromolecular modeling.

In this study, we show that an approach integrating both sources of data in biomolecular energy function optimization can bring significant improvements to its performances on various protein structure prediction tasks. Our new optimization framework is capable of robust optimization of over hundred parameters with respect to a large set of computationally expensive tasks, and is used to derive a next-generation Rosetta[21,22] energy function for general protein structure prediction tasks, with improvements in the Lennard-Jones, implicit solvation, and electrostatic models. Similar ideas have been suggested to the development of modern molecular mechanics force fields[16,23,24] for molecular dynamics simulations. However, to our knowledge, it is first time this approach has been used in the

context of general structure prediction problems; furthermore, we show the applicability of this approach by testing on a variety of challenging structure prediction tasks including monomeric structure discrimination or conformational search, protein homology modeling, protein-protein and protein-ligand docking, sequence prediction at protein core, protein-protein, and protein-ligand interfaces, and prediction of free energy changes brought about by mutations ($\Delta\Delta$G), supported by massive amount of state-of-the-art structural modeling techniques. Finally, we show the robustness of this optimization, even when fitting >100 parameters, which allows us to assess the tightness of convergence of atomic-level parameters.

METHODS

Overview of the approach

We set out to parameterize an energy function based on experimental thermodynamic data of small molecules, and high-resolution structural data of macromolecules (shortly "structural data"), with the broader aim of better recapitulating the large-scale energy landscape of protein folding or complex formation, high-resolution structural features, and the balance between natural amino acid preferences. The experimental thermodynamic data consists of the liquid properties of small molecules containing functional groups from natural amino acids[12] and vapor-to-water transfer free energies of protein side-chain analogs[25]. The structural data consists of large numbers (> 1000 cluster centers) of alternative conformations (decoys) for protein structures and complexes of known structure, and high-resolution crystallographic data. The agreement of an energy function with these data is represented by a target function F_{total}:

$$\mathbf{F_{total}}\left[E\left(\Theta\right)\right]=w_{\text{thermodynamic}}\mathbf{F_{thermodynamic}}\left[E\left(\Theta\right)\right]+w_{\text{structural}}\mathbf{F_{structual}}\left[E\left(\Theta\right)\right] \quad [1]$$

where the target functions $F_{thermodynamic}$ and $F_{structural}$ are *functionals* of a biomolecular energy function $E(\Theta)$, which is a function of a set of parameters Θ (see the sections below for the details of $E(\Theta)$ in the study), and their relative contributions are adjusted by weights *w*. $F_{thermodynamic}[E(\Theta)]$ and $F_{structural}[E(\Theta)]$ are themselves a weighted linear sum of target functions evaluating performance on specific tasks; their exact composition varies depending on the aim of optimization, and is described in the following paragraphs and sections.

The energy parameters Θ subject to optimization consist of atom-type-dependent parameters, for example, the Lennard-Jones (LJ) radius and well-depth of each atom type. The total number of parameters simultaneously optimized in a single run is on the order of 100. A key advantage of the ability to simultaneously optimize a large number of parameters is that the introduction of significant changes to the physical models of energy terms (for example, an anisotropic solvation model, or change in LJ model of hydrogen atoms) may considerably shift the balance between the energy terms and require large-scale re-parameterization. Optimization of these large parameter sets, with respect to a wide range of thermodynamic and structural data, is performed by a newly developed parameter optimization protocol named *dualOptE* that uses Nelder-Mead simplex optimization[26] (Figure 1, details in following section).

We found several factors to be critical for energy function training to robustly transferrable to independent datasets. First, the training data need to be *diverse*; consequently, energy function performance is trained on a wide variety of sub-tasks, including recapitulation of sequence and side-chain rotamers, native monomeric structure discrimination, protein-protein docking, and the aforementioned thermodynamic recapitulation tasks. Second, the structure discrimination training sets must be *dynamic*; it is all too easy to train an energy function to consistently recognize the native structure in a sea of static decoys, but much more challenging when all structures are relaxed in the new energy function[27]. In *dualOptE*, all tests involve some reoptimization against the current energy function: for example, the test measuring the ability to discriminate near-native monomeric conformations or protein-protein interfaces first optimizes a pre-generated set of structures against the current parameterization before assessing discrimination quality. Third, each cycle of parameter optimization must be carried out in a limited amount of computer time. Since we need to assess hundred or thousands of parameterizations in the course of an optimization trajectory, each test has to run on the order of several minutes. For example, during parameter optimization, a full liquid MC simulation to estimate liquid phase properties of small molecules at each step is not computationally tractable; we instead use static sets of snapshots from MC simulations. Following completion of a given parameter optimization run we carried out full liquid MC simulations and found that the static approximation was fairly accurate as long as there were not large changes in the energy function.

We employed multiple iterations of this dual energy function optimization approach. The first iteration, yielding the energy function *opt-july15*, introduced a new anisotropic implicit solvent model into the Rosetta energy function. Rosetta has previously used the Lazaridis-Karplus (LK) isotropic occlusion-based implicit solvation model[28], where the occluded volume of each atom is proportional to the fractional desolvation energy. The new anisotropic solvation model combines the isotropic part from the original LK model with a ne wly introduced anisotropic polar term[29], which accounts for anisotropic interactions between polar heavy atoms and solvent: occlusion of water binding sites is made more energetically disfavorable than occlusion away from such sites. A second series of optimizations follows introduction of attractive dispersion forces to hydrogens (originally pseudo-united-atom) as well as a reworked electrostatic model, yielded the energy function *opt-nov15*. For both energy function “snapshots”, following optimization, the resulting energy functions were validated on a set of independent structure prediction tasks too computationally intensive to be used in optimization[30]. Details of energy functions (*opt-july15 and opt-nov15*) and energy terms, and a list of the tests used for optimization are described in following sections, a full list of atomic parameters determined by *DualOptE* appear in the Supplementary Tables S3–4, and details of the tests and datasets for optimization or independent validation in Supplementary Materials.

The resulting next-generation Rosetta energy function (*opt-nov15*) outperforms the previous energy function (*talaris2014*)[13] on a wide range of structure prediction tests independent of the training set data. In contrast to *opt-nov15*, *talaris2014* had been optimized solely relying on similar set of structural data we incorporate in the study without the use of small molecule data. We briefly summarize the energy function changes; full details are again provided below. First, there are changes in the physical models, notably the new anisotropic

solvation model, a sigmoidal dielectric model, and explicit modeling of the effects of proline on the backbone torsion angles of the preceding residue. Second, there are changes in the representation; in previous Rosetta energy functions hydrogens are purely repulsive to speed computation (much shorter range distance cutoffs were required), whereas in the new energy function hydrogens make attractive LJ interactions. Third, there are changes in the overall balance of forces: compared to *talaris2014*, both solvation and electrostatic forces are considerably stronger relative to other non-bonded interactions. Fourth, there are changes in many energy function parameters: the attractive interactions of sulfur and aliphatic carbons are stronger (which bring better agreement with small-molecule liquid phase data), and the partial charges of charged chemical groups are more evenly distributed (rather than being primarily on the tip atoms).

Parameter optimization using dualOptE

The aim of the parameter optimization method *dualOptE* is to explore a high (100+) dimensional parameter space, identifying a parameter setting that minimizes a target function (e.g. Eqn. 1). Parameters to be optimized generally consist of atomic parameters from multiple energy terms, resulting in a parameter space of dimension around 100. Given the relatively large number of parameters and the complexity of the target function, the most reasonable choice of optimization method is a derivative-free approach. In this study, Nelder-Mead simplex optimization[26] is applied. The method has several desirable features: it is derivative-free and generally requires relatively few function evaluations to converge.

The overall minimization process took around 5 days for *opt-july15* and *opt-nov15* using hundreds of computer cores in parallel (typically 160 cores in the study). Each minimization step took ~12 minutes, and the process was terminated when optimization converged; this generally occurred after ~600 iterations. To keep the optimization computationally tractable, each individual test needed to run in, at most, 4–6 minutes total. In doing so, this allowed the roughly 800 individual tests to run on a cluster in ~12 minutes.

While developing the *dualOptE* protocol, a key computational challenge related to the sequence recovery test was to synchronize amino-acid type reference weights with the current energy function parameterization at every single iteration. Reference weights in the energy model balances relative occurrences of amino acids, hence, are very specific for a particular parameterization of the energy function. Adding reference weights as additional parameters to optimization adds significant complexity to the parameter space. To address this issue, we developed a method that fits reference energies on-the-fly at each iteration, treating them as independent implicit parameters from the optimization. Given a set of target structures, the method considers each possible mutation at each position. For each mutation, all sidechain rotamers within 8Å of the mutated residue are re-optimized, and a reference-energy-free score is computed. Then the same Nelder-Mead simplex optimization is used to find the reference weights that maximize sequence recovery. Because all energies are pre-computed, this is very quick to evaluate, taking about 45 seconds total.

Optimization procedure

As an initial proof of concept, we applied our optimization scheme *dualOptE* to the current energy function, *talaris2014*. Optimization allowed per-atom-type solvation model and Lennard-Jones parameters to change, as well as a small set of parameters related to the electrostatic model. The total number of parameters for this initial experiment was about 80. The target optimization function (F_{total}), which is functional of energy function $E(\Theta)$, used took the form:

$$\begin{aligned}\mathbf{F_{total}}\left[E\left(\Theta\right)\right] = & w_{staticliquid}\mathbf{F_{staticliquid}}\left[E\left(\Theta\right)\right] + w_{\text{water- to- vapor}}\mathbf{F_{water-to-vapor}}\left[E\left(\Theta\right)\right] \\ & + w_{\text{sidechain- core}}\mathbf{F_{sidechain-core}}\left[E\left(\Theta\right)\right] + w_{\text{sidechain- interface}}\mathbf{F_{sidechain-interface}}\left[E\left(\Theta\right)\right] \\ & + w_{\text{sequence}}\mathbf{F_{sequence}}\left[E\left(\Theta\right)\right] \\ & + w_{monomeric,static}\mathbf{F_{monomeric-static}}\left[E\left(\Theta\right)\right] + w_{monomeric-\text{min}}\mathbf{F_{monomeric-min}}\left[E\left(\Theta\right)\right] \\ & + w_{\mathbf{protein-protein}}\mathbf{F_{protein-protein}}\left[E\left(\Theta\right)\right] + w_{\text{atompair}}\mathbf{F_{atompair}}\left[E\left(\Theta\right)\right]\end{aligned}$$

[2]

The first two terms evaluate the energy function against small molecule thermodynamic data, while the remainders are guided by high-resolution protein structural data. More specifically, *staticliquid* evaluates thermodynamic liquid properties[12] using static "snapshot" evaluation; *water-to-vapor* recapitulates the solvation free-energy change of side-chain analogs upon transferring from vapor to water[25]; *sidechain-core* and *sidechain-interface* predict the rotameric state of side-chains at the core and interfaces of native proteins, respectively; *sequence* predicts the amino acid identity of core residues of natural wild-type proteins; *monomer, static* and *monomer, min* evaluate monomeric structure discrimination on a pre-sampled conformation set, both before and after a short energy minimization, respectively; *protein-protein* evaluates the discrimination of pre-sampled protein-protein complex conformations; and *atompair* evaluates the atom-pair distance distribution observed from crystal structures. Further details of individual tests, including protocols for using small-molecule thermodynamic data for the parameterization, are described in Supplementary Methods.

This initial optimization revealed improvements on all tests, however, the improvements were quite subtle. Generally, the changes tended to slightly increase the contribution of solvation energy compared to other non-bonded energies. However, the small magnitude of improvement suggests that the current energy function, which has been heavily developed over the past dozen years, is close to optimally parameterized. Thus, it seems likely that the introduction of more dramatic changes to the energy model are necessary to see improvements in both structure prediction and recovery of small molecule thermodynamic properties.

Consequently, our next optimization experiment introduced several broad improvements to the underlying physical model, most notably, the introduction of anisotropy to the implicit solvation model[29], and a sigmoidal dielectric Coulombic model[31]. Additionally, we introduce a few more minor changes, outlined in the Supplementary Methods section. The

parameters optimized are the same as before: i) two LJ parameters (radius and well-depth) for all LJ atom types, ii) one parameter (ΔG_{free}) for all solvation atom types[28], iii) hydrogen bonding weights for individual donor or acceptor types, and iv) three parameters describing the shape of the dielectric as a function of distance. Prior to the *dualOptE* run, we manually adjusted the initial LJ parameters for certain atom types (CH3, CH2, CH1, and S) to improve the agreement with the liquid property data. Several rounds of optimization were carried out, first optimizing polar parameters (~50), than optimizing all parameters (~100). The resulting optimized energy function is referred to as *opt-july15*.

In the second round of optimization, we further extended the energy function by introducing several more large-scale changes and subsequently re-optimizing. First, we introduce E_{rama_prepro} as a replacement of E_{rama} to take into account of specific ϕ/ψ angle preferences of pre-Proline residues. Second, we incorporated a more reasonable description of LJ interactions by converting our previous pseudo-united atom representation (where hydrogens only have a short-ranged repulsive contribution) to an all-atom representation. The LJ parameters for affected atoms were initially borrowed from the OPLS force field[12] followed by manual adjustment (due to difference in LJ model) using liquid simulations prior to optimization. The parameter set subject to optimization consists of all the parameters optimized in the first round, plus a set of parameters describing the partial charges of all amino acids, for a total of ~115 parameters. While optimizing atomic partial charges, we applied a grouped optimization scheme: partial charges are grouped together based on their physical relations (similar to the grouped charge concept in CHARMM force fields[10]) and their magnitudes are scaled such that the overall strength of a group changes but its net charge does not vary. This scheme is advantageous not only in reducing the number of parameters, but also in keeping partial charges within a physically reasonable range (e.g. maintaining balance between atoms with a dipole). Nonpolar groups were scaled together, leading to a total of 17 group parameters describing partial charges. Drawing off expertise from our first experiments, the target function for second run was modified:

$$\begin{aligned}\mathbf{F_{total}}\left[E\left(\Theta\right)\right] = & w_{\text{water- to- vapor}}\mathbf{F_{water-to-vapor}}\left[E\left(\Theta\right)\right] \\ + w_{\text{sidechain- core}}\mathbf{F_{sidechain-core}}\left[E\left(\Theta\right)\right] + & w_{\text{sidechain- interface}}\mathbf{F_{sidechain-interface}}\left[E\left(\Theta\right)\right] \\ + w_{\text{sequence}}\mathbf{F_{sequence}}\left[E\left(\Theta\right)\right] + & w_{monomeric,static}\mathbf{F_{monomeric-static}}\left[E\left(\Theta\right)\right] \\ + w_{monomeric-\text{min}}\mathbf{F_{monomeric-min}}\left[E\left(\Theta\right)\right] + & w_{\text{protien- protien}}\mathbf{F_{protein-protein}}\left[E\left(\Theta\right)\right] \\ + w_{\text{xtal- grad}}\mathbf{F_{xtal-grad}}\left[E\left(\Theta\right)\right] + & w_{\text{atompair}}\mathbf{F_{atompair}}\left[E\left(\Theta\right)\right]\end{aligned} \quad [3]$$

Compared to the previous optimization (Eqn. 2), one test (*staticliquid*) was dropped while another (*xtal-grad*) was added. The reason for dropping the static evaluation test of liquid properties is that: a) unlike the previous optimization, the initial model shows relatively good agreement with these properties, b) the test was relatively time-consuming computationally, and c) the accuracy of the estimation using static snapshots was limited. We decided instead to use liquid properties as an independent validation measure. Meanwhile, a new test, *xtal-grad* was added, which assesses the magnitude of energy-function gradients following crystallographic refinement, enhancing the accuracy of the energy function at high resolution.

Following optimization with the non-bonded parameters above, optimization was also carried out on bonded parameters (spring constants for bond distance, angle, and improper torsion angles) while other parameters were fixed. Similar to partial charge optimization, parameters were grouped into about 29 sets and each set was scaled together. We applied a simpler target function here, as changes in bonded parameters are expected to give minimal effects to the excluded tests:

$$\begin{aligned}\mathbf{F_{total}}\left[E\left(\Theta\right)\right] = & w_{\text{sidechain- core}}\mathbf{F_{sidechain-core}}\left[E\left(\Theta\right)\right] + w_{\mathbf{sidechain-interface}}\mathbf{F_{sidechain-interface}}\left[E\left(\Theta\right)\right] \\ & + w_{\text{sequence}}F_{\text{sequence}}\left[E\left(\Theta\right)\right] + w_{monomeric,static}\mathbf{F_{monomeric-static}}\left[E\left(\Theta\right)\right] \\ & + w_{\text{monomeric- min}}\mathbf{F_{monomeric-min}}\left[E\left(\Theta\right)\right] \\ & + w_{\text{xtal- grad}}\mathbf{F_{xtal-grad}}\left[E\left(\Theta\right)\right] + w_{\text{atompair}}\mathbf{F_{atompair}}\left[E\left(\Theta\right)\right]\end{aligned}$$

[4]

The final optimized energy function *E(Θ)* using this target optimization function is referred to as *opt-nov15*.

Opt-nov15 energy model

The Rosetta energy model is specialized for macromolecular modeling studies, which means there are relatively strict requirements on functional representation and efficiency. An implicit description of solvent molecules (as opposed to explicit solvent representation) is necessary for instantaneous evaluation of the energetics of a single conformation. Additionally, solvation and electrostatics, which generally are the rate-limiting steps in total energy-function evaluation, are restricted to be pair-wise decomposable at the residue level, ensuring that rotamer and sequence optimization can be carried out efficiently; these terms must also be evaluated in a runtime that is of similar order-of-magnitude to Lennard-Jones (LJ) interactions.

The energy models used in this study (*opt-july15* and *opt-nov15*) are represented as weighted linear sum of multiple energy terms meeting the aforementioned restrictions:

$$\begin{aligned}E_{\text{total}} = & E_{LJ,\text{atr}} + W_{LJ,\text{rep}}E_{LJ,\text{rep}} + E_{\text{Coulomb}} + E_{\text{Hbond}} + E_{\text{solv,iso}} + E_{\text{solv,}an\text{iso}} \\ & + W_{\text{dun}}E_{\text{dun}} + W_{\text{rama}}E_{\text{rama}} + W_{\text{paapp}}E_{\text{paapp}} + W_{\text{bonded}}E_{\text{bonded}} \\ & + E_{\text{ref}}\end{aligned} \quad [5]$$

For each component, “w” and “E” represents weight and energy value of each component, respectively. In this study, non-bonded terms shown in first row of Eqn. 5 (*LJ_atr, Coulomb, Hbond, solv_iso, solv_aniso*) are weight-less except for LJ repulsion; its weight (w_{LJ_rep}) is fixed to 0.55 to address the overestimation of the exclusion effect by the 12th-order repulsive term. In Rosetta energy function torsion terms (*dun, rama, p_aa_pp*) are derived from statistics of macromolecule data hence have arbitrary units with respect to non-bonded terms having physical interpretations. Their weights are used by adjusting values from *talaris2014:* w_{dun}= 0.7, w_{rama}= 0.45, and $w_{p\_aa\_pp}$ = 0.6. Many of the energy components share the same functional form with *talaris2014* energy function[13], including LJ (*LJ_atr* and *LJ_rep*, except

that hydrogens only have repulsive part in *talaris2014*), isotropic Lazaridis-Karplus (LK) solvation model (*solv_iso*)[28], orientation-dependent hydrogen bonding (*Hbond*)[32], bonded terms (*bonded*), and amino acid reference weight term (*ref*). For these terms, a subset of parameters are optimized, while maintaining the original functional form. For backbone-dependent side-chain torsion preference (*dun*) and preference of amino acid given ϕ/ψ angles (*p_aa_pp*), we brought a more-recent parameterization[14]. Other terms, either newly introduced (*solv_aniso*) or with change in energy models (*Coulomb* and *rama*), are described in Supplementary Methods.

RESULTS

Improvements in monomeric structure prediction

We describe the improved performance of the new energy function on a battery of protein and small molecule tests evaluated on a set of proteins *distinct from those used in optimization*, as compared to the current Rosetta energy function, *talaris2014*. All results are summarized in Table 1, which shows energy function performance using two different metrics: a “weighted evaluation metric” that estimates the Boltzmann probability of the native-like conformations, and a “success-rate based” metric, which simply measures if the lowest-energy conformation is near-native (details of metrics are in Supplementary Methods). The most striking improvement was seen in monomeric structure prediction tasks, particularly in three sub-tests: near-native structure discrimination following reoptimization (Figure 2), structure prediction with *parallel loophash sampling* (PLS)[30], and homology modeling with *RosettaCM*[2].

The first of these tests, *decoy discrimination* in Table 1, evaluates the ability of the energy function to pick out near-native structures (up to 200 residues in length) from a set of pre-sampled compact structures broadly covering conformational space[33]. In order to address the dynamic aspect of the energy landscape (that is, local energy minima may vary greatly following parameter changes), a structural relaxation is carried out with a given energy function prior to evaluation. The test is further divided into structures for which the native conformation was used in optimization (*set1*) and structures for which the native was not seen by optimization (*set2*). Remarkable and consistent improvement is found as optimization proceeds from *talaris2014* to *opt-july15* to *opt-nov15* (Figure 2A), increasing the success rate of decoy discrimination by 20.8% (36.3% to 57.1%) and 14.1% (53.1% to 67.2%) on set1 and set2, respectively, when compared between *talaris2014* and *opt-nov15*; the fact that the improvement is consistent on *set2* (containing structures independent of those used in optimization) suggests that the energy function is not simply memorizing features from a small set of native structures. A few examples of energy landscapes and structures in Figure 3 show that improvements happen in various ways, discriminating structural variations from subtle local level (*1aaj, 1xmt*, blue in *1luz* and *1igd*) to secondary structure orientations (*2i4s, 1l8r, 1ifb*, yellow in *1igd*) to different fold-level (*2y4x*, orange in *1luz*). Also, improved secondary structure balance is observed in many cases (*1aaj, 1luz, 1igd*).

As relaxation of pre-sampled structures may still limit the conformational space considered, we ran a more extensive validation where the new energy function is directly used to guide

fold-level conformational sampling. The test is carried out using the *parallel loophash sampling* (PLS) protocol[30]. This highly parallelized and CPU-intensive protocol maintains a pool of low-energy structures, which it continually perturbs by replacing local portions of each protein with fragments randomly taken from alternate structures; new low-energy structures are continually added to the pool as they are discovered. Our previous studies have shown that the protocol serves as a powerful tool for identifying local minima of a given energy function. As shown in Figure 2B, the results of the PLS conformational sampling search follow a similar trend to that of decoy discrimination; the average Boltzmann-weighted discrimination over 36 targets improves from 0.639 to 0.752 and success rates from 30.6% to 63.9%, after running PLS with *talaris2014* and *opt-nov15*, respectively.

Finally, we show how the energy function performs on homology modeling, by using *RosettaCM*[2] to predict protein conformations from a set of known high-resolution homologous structures. A comparison of the energy functions on 69 homology modeling cases since August 2014 of the CAMEO continuous evaluation benchmark[34] is shown in Figure 2C. There is small but consistent improvement across this set with the new energy function; this is because structural sampling is strongly restricted by homologous information, hence improvements can be mostly found from high-resolution structural features like secondary structure packing.

Improvements in protein-protein and protein-ligand docking

The new energy function improves both protein-protein and protein-ligand docking (Figure 4) on independent tests. The improvement in protein-ligand docking is particularly notable, as no tests of this nature are used in optimization, again suggesting that our optimized energy function has fundamentally improved the underlying physical model. Various pre-sampled conformations using independent tools (Supplementary Materials) are relaxed in the same manner as in the monomeric tests prior to evaluation. Successful docking requires that: a) non-bonded interaction terms correctly estimate the magnitude of the attractive interactions upon complex formation, and b) they are properly balanced against the desolvation energy. Specific protein-protein (Figure 4A) and protein-ligand (Figure 4B) docking examples illustrate how *opt-nov15* recovers this delicate balance. In the protein-protein interaction case, *talaris2014* favors a non-native conformation, with larger buried surface area but fewer highly favorable interactions (blue), over the native conformation with smaller buried surface area but a greater number of highly favorable interactions (green, inset in Figure 4A). The protein-ligand interaction example shows another balancing issue: the native interface is larger but features more electrostatically favorable interactions. In this example, *talaris2014* fails because the strength of these favorable electrostatic interactions is not great enough to overcome the greater desolvation penalty incurred by the larger interface.

Improvements on protein design tasks

We tested the prediction ability of the energy function on independent protein design tasks by performing fixed-backbone sequence design[27] and measuring the agreement to native sequence profiles. Evaluating ability of an energy function on design tasks is a critical part of our assessment as Rosetta suite has been broadly used for computational protein sequence

designs. These results (Table 1) suggest *opt-nov15* also improves in balancing energetic preferences among different amino acids. This test is somewhat orthogonal to other structure prediction tasks in that every amino acid is considered at every context, so the penalty for burying for example an arginine has to be properly balanced against all other amino acids at that position. Two different types of test are considered: in the first (*Protein monomer, Protein-protein interface*, and *Protein-ligand interface*), each residue is designed independently while neighboring side-chains are only allowed to reorganize (not change identity); in the second (*full sequence design*), all sequence information is removed from the protein chain and full sequence optimization calculations are carried out. In both cases, both recovery rates and sequence-profile-weighted recovery metric values (see Supplementary Methods) see a consistent improvement of around 2%.

Free energy changes accompanying mutations

Computation of free energy change brought about by sequence mutations ($\Delta\Delta$G) is both a fundamental test of modeling accuracy and an increasingly relevant problem as high-throughput sequencing reveals sequence polymorphisms at an increasing rate[36]. This problem is related to design problems described above, but more challenging as it requires accurate prediction of changes in the structure resulting from single amino acid substitutions for successful estimation of the free energy changes. Despite efforts made in conformational sampling[37–39], it is clear there is significant room for improvement in both problems. We used a previously published benchmark set[37] containing 1211 mutations on monomeric proteins with high-resolution structural data available. We not only use the new energy function, also a new sampling protocol that takes advantage of recently developed Cartesian space sampling methods (more details in Supplementary Methods). Using this new protocol, the Pearson correlation between estimated versus experimentally measured $\Delta\Delta$G improves from 0.703 to 0.743; classification accuracy of stabilizing/destabilizing mutations also improves from 71.3% to 72.9%, compared to *talaris2014* (Figure S1). Compared to previous published protocols[37], the combination of the new energy function and protocol produces a non-trivial improvement in correlation with comparable accuracy in simple classification.

Improvements in recapitulation of high-resolution structural information

We also measure the ability of the energy function to recapitulate high-resolution structural data, on a set of structures not used in optimization. Here, we use two metrics: the recapitulation of atom-pair distance distributions from high-resolution structures, and the energy function gradients following crystallographic refinement with the target energy function (see Supplementary Methods). Results are shown in Table 1 and Figure S2. The most significant improvement found in atom-pair distance distributions are from aliphatic carbons; for instance, the error (as assessed by KL-divergence) is reduced by half for CH_2-CH_3 and CH_3-CH_3, the fourth and fifth most abundant atom pairs found in protein structures. Furthermore, the normalized energy function gradients following crystallographic refinement are reduced from 0.150 to 0.077, indicating that there are much weaker forces moving the structure away from a minimum consistent with high-resolution crystallographic data. These results are quite consistent; all 49 cases tested show a reduction in the normalized gradient.

Agreement of the energy function to thermodynamic data

Although thermodynamic data from small molecules was largely used as a regularizer in our optimization, restricting optimization to a physically realistic subspace of the high-dimensional parameter space, there were still significant improvements in the energy function's ability to recapitulate thermodynamic data of small molecules; this is not surprising since this is first time to train our energy function on small-molecule thermodynamic data. Figure 5 summarizes the results comparing liquid phase properties computed using full Monte Carlo simulation (simulation details in Supplementary Methods) over experimentally measured data on a set of small molecules (Figure S3). Optimization reduces the relative error from 19.5% to 6.3% for H_{vap} (heat of vaporization), and from 5.7% to 5.1% for density. Interestingly, the error in $C_{p,l}$ (latent heat capacity) — which was not used for optimization at any point — was also reduced from 14.6% to 11.3%. Errors found in H_{vap} estimation identify issues in the atomic parameters of *talaris2014* the were not identified from the macromolecular tests alone, such as underestimation of non-bonded interaction strength of aliphatic and sulfur-containing functional groups, which led to underestimation of hydrophobic interactions and a bias to aromatic side-chains at the protein core. The errors are still larger than what are observed from molecular mechanics force fields; however, it is notable that reasonable estimation of liquid phase small molecule properties is possible while improving performance of the energy function on protein structure prediction tasks. Parameters could have been more tightly constrained to better match liquid properties, however, it is questionable whether a perfect fit to a limited set of liquid properties accurately reflects the accuracy of an energy model[40,41]. Finally, water-to-vapor transfer energy of protein-sidechain analogs is in good agreement in all the energy functions (Figure S4). This is because in *talaris2014* all the solvation parameters were directly brought from original LK model, while in *opt-july15* and *opt-nov15* the data was used as regularizer of parameterization.

DISCUSSION

A subspace of equivalently good parameterizations

Important questions arising are i) how closely the parameters optimized on structural data would match to those of molecular mechanics force field, and ii) how confident we can be of parameter sets following optimization in *dualOptE* for structure prediction tasks. To address these questions, we set up two additional experiments: *opt-from-XX* (with *XX* either CHARMM or OPLS), which runs optimization starting with the LJ and partial charge parameters from other molecular mechanics force fields; and *opt-from-random*, where we start optimization by randomly perturbing *opt-nov15* parameters. Following optimization, all runs achieved similarly good scores on the target function, and showed similar performance on independent decoy discrimination tests (Supplementary Table S2), though the final parameter sets slightly differed from *opt-nov15* as well as from each other. As shown in Table 2, each *opt-from-MM* parameter set converged on a slightly different optimum than *opt-nov15* as well as from the values from CHARMM or OPLS; still, overall variation stays close to the variations observed among different molecular mechanics force fields. Regarding the second question, parameters obtained from multiple *opt-from-random* also show small deviations from *opt-nov15*; when 10 representative parameters are collected

from each of 3 independent runs (for a total of 30 equivalently good parameterizations), mean deviations from *opt-nov15* are 0.018 in partial charges, 0.040 Å in LJ radii, 0.013 kcal/mole in LJ well-depth, and 1.20 kcal/mole in solvation ΔG_{free} (per-parameter values and standard deviations reported in Supplementary Table S3-4). Measuring the correlation between parameters clarifies that deviations mainly originate from covariation of strongly coupled parameters rather than ambiguities within each of the values (Supplementary Table S5): for instance, LJ radii of aromatic hydrogen (Haro) and carbon (aroC) largely vary in a strongly coupled manner (Pearson correlation −0.866). The ability to systematically estimate the precision with which individual energy function parameters are determined is an advantage of our approach over conventional force field optimization approaches.

Combining two different data sources simultaneously helps parameter optimization in a complicated space

One of the most challenging parts in energy function optimization is maintaining the delicate balance between individual components of a composite energy function. We show that by training on a wide range of macromolecular data sets and challenges, together with small-molecule thermodynamic data, we can robustly optimize on the order of 100 parameters, allowing both discovery and maintenance of this balance. The importance of this balance is illustrated through the results on individual structure prediction tests: balancing non-bonded and torsional interactions is key in avoiding secondary structure biases in structure prediction (Figure 3); balancing solvation energies with other non-bonded energies is important in protein-protein docking tests; and balancing Lennard-Jones and electrostatic contributions important in matching small molecule thermodynamic properties.

The effectiveness of our dual optimization approach in optimizing many hundreds of parameters without over-fitting likely stems from several factors. First, each of the small-molecule tests constraints a different subset of parameters, and so, despite the large parameter space, each test conceptually subdivides parameter space into lower-dimensional subspaces, which leads to a tractable optimization process even though the overall dimensionality is quite high. Second, the relative sparsity of the small molecule features (compared to the number of parameters) leads to a large space of possible parameterizations satisfying the data[40–42]; this is additionally supported by variations found in atomic parameters among molecular mechanics force fields (Table 2). By optimizing in this space against the macromolecular structural data, we are less prone to get stuck in local minima of the structure-prediction target functions in physically unrealistic parts of parameter space. Indeed, optimized LJ parameters and partial charges stay similar to those of popular molecular mechanics force fields (Table 2). Finally, unlike previous approaches in energy function optimization[27], our framework readily handles structure prediction tests of a more dynamic nature, ensuring the tests give reasonable results even as we move far away from the starting point in parameter space.

Our approach is more generalizable than previous efforts combining small molecule and macromolecular data in optimization. In modern molecular mechanics force fields, adaptive corrections derived from macromolecule simulations significantly enhanced the recapitulation of local structural features, such as backbone or side-chain torsional

preferences[16,24], or folding energy landscapes of small proteins (less than 100 residues)[43]. Still, the impact of corrections to molecular mechanics force fields have been tested to a limited set of structure prediction tasks, not the global energy landscape properties over a broad set of protein structure prediction tasks considered in this manuscript. Our study from a structure prediction perspective is offering a complementary view to general biomolecular energy function optimization problem.

Finally, the presented optimization scheme is quite general, and allows for fast reparameterization following large-scale changes to the functional form of the energy function. One key weakness of the current energy function is the separation of hydrogen bonding and Coulomb electrostatics. While this offers some advantages, e.g. preventing the overestimation of LJ well-depth for hydrogen bonding atoms, its weakness arises from under-representation of electrostatics between non-hydrogen bonding partners such as electrostatic repulsion or multipole effects (e.g. **π-π** interaction), as supported by liquid simulation results (Figure 5). One future direction will be developing a unified electrostatic model that can address these issues. Another weakness concerns the solvation model; the current formulation has weakness describing the screening effect of electrostatic interactions by water or ions, in contrast to the more computationally expensive Poisson-Boltzmann (PB) or generalized Born (GB) solvation models. Incorporating such properties in an efficient pairwise decomposable manner is a difficult but potentially valuable future research effort[44]. Lastly, a more profound challenge in an implicit solvation model will be describing effects from well-ordered waters that form an essential part of biomolecular structures and interfaces.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Sources

This work was supported by US National Institutes of Health under award numbers R01GM092802 (H.P., D.E.K., and D.B.).

The authors declare no conflict of interest. We thank to Dr. Roland Dunbrack Jr. and Maxim Shapovalov at Fox Chaser Cancer Center for their work on development of torsional potential used in this study. We also thank to Dr. Andrew Leaver-Fay and Dr. Brian Kuhlman at University of North Carolina, and Dr. Shane O'Connor and Dr. Tanja Kortemme at University of California, San Fransisco for helpful discussions. An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

References

1. Wang RY-R, Kudryashev M, Li X, Egelman EH, Basler M, Cheng Y, Baker D, DiMaio F. Nat Methods. 2015; 12(4):335–338. [PubMed: 25707029]
2. Song Y, DiMaio F, Wang RY-R, Kim D, Miles C, Brunette T, Thompson J, Baker D. Structure. 2013; 21(10):1735–1742. [PubMed: 24035711]
3. Ovchinnikov S, Park H, Kim DE, Liu Y, Yu-Ruei Wang R, Baker D. Proteins. 2016; 84(S1):181–188. [PubMed: 26857542]

4. Kamisetty H, Ovchinnikov S, Baker D. Proc Natl Acad Soc. 2013; 110(39):15674–15679.
5. Lyskov S, Gray JJ. Nucleic Acids Res. 2008; 36:W233–W238. Web Server issue. [PubMed: 18442991]
6. Meiler J, Baker D. Proteins: Struct Funct Bioinf. 2006; 65(3):538–548.
7. Shaw, DE., Chao, JC., Eastwood, MP., Gagliardo, J., Grossman, JP., Ho, CR., Ierardi, DJ., Kolossváry, I., Klepeis, JL., Layman, T., McLeavey, C., Deneroff, MM., Moraes, MA., Mueller, R., Priest, EC., Shan, Y., Spengler, J., Theobald, M., Towles, B., Wang, SC., Dror, RO., Kuskin, JS., Larson, RH., Salmon, JK., Young, C., Batson, B., Bowers, KJ. Proceedings of the 34th annual international symposium on Computer architecture - ISCA '07. ACM Press; New York, New York, USA: 2007. p. 1
8. Shaw DE. J Comput Chem. 2005; 26(13):1318–1328. [PubMed: 16013057]
9. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. J Chem Theory Comput. 2013; 9(9):3878–3888. [PubMed: 26592383]
10. MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, et al. J Phys Chem B. 1998; 102(18):3586–3616. [PubMed: 24889800]
11. Ponder JW, Case DA. Adv Protein Chem. 2003; 66:27–85. [PubMed: 14631816]
12. Jorgensen WL, Maxwell DS, Tirado-Rives J. J Am Chem Soc. 1996; 118(45):11225–11236.
13. O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, DiMaio F, Bradley P, Kortemme T, Baker D, Snoeyink J, Kuhlman B. J Chem Theory Comput. 2015; 11(2):609–622. [PubMed: 25866491]
14. Shapovalov MV, Dunbrack RL Jr. Structure. 2011; 19(6):844–858. [PubMed: 21645855]
15. Xu D, Zhang Y. Proteins. 2012; 80(7):1715–1735. [PubMed: 22411565]
16. Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, Mackerell AD Jr. J Chem Theory Comput. 2012; 8(9):3257–3273. [PubMed: 23341755]
17. MacKerell AD, Feig M, Brooks CL. J Comput Chem. 2004; 25(11):1400–1415. [PubMed: 15185334]
18. Ren P, Wu C, Ponder JW. J Chem Theory Comput. 2011; 7(10):3143–3161. [PubMed: 22022236]
19. Lopes PEM, Huang J, Shim J, Luo Y, Li H, Roux B, Mackerell AD Jr. J Chem Theory Comput. 2013; 9(12):5430–5449. [PubMed: 24459460]
20. Raval A, Piana S, Eastwood MP, Dror RO, Shaw DE. Proteins: Struct Funct Bioinf. 2012; 80(8): 2071–2079.
21. Leaver-Fay, A., Tyka, M., Lewis, SM., Lange, OF., Thompson, J., Jacak, R., Kaufman, KW., Renfrew, PD., Smith, CA., Sheffler, W., Davis, IW., Cooper, S., Treuille, A., Mandell, DJ., Richter, F., Ban, YEA., Fleishman, SJ., Corn, JE., Kim, DE., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, JJ., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, JJ., Kuhlman, B., Baker, D., Bradley, P. Computer Methods, Part C; Methods in Enzymology. Vol. 487. Elsevier: 2011. p. 545-574.
22. Das R, Baker D. Annu Rev Biochem. 2008; 77:363–382. [PubMed: 18410248]
23. Nguyen H, Roe DR, Simmerling C. J Chem Theory Comput. 2013; 9(4):2020–2034. [PubMed: 25788871]
24. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. J Chem Theory Comput. 2015; 11(8):3696–3713. [PubMed: 26574453]
25. Radzicka A, Wolfenden R. Biochemistry. 1988; 27(5):1664–1670.
26. Nelder JA, Mead R. Comput J. 1965; 7(4):308–313.
27. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, Gray JJ, Kortemme T, Richardson JS, Havranek JJ, Snoeyink J, Baker D, Kuhlman B. Methods Enzymol. 2013; 523:109–143. [PubMed: 23422428]
28. Lazaridis T, Karplus M. Proteins. 1999; 35(2):133–152. [PubMed: 10223287]
29. Yanover C, Bradley P. Nucleic Acid Res. 2011; 39(11):4564–4576. [PubMed: 21343182]
30. Tyka MD, Jung K, Baker D. J Comput Chem. 2012; 33(31):2483–2491. [PubMed: 22847521]
31. Hingerty BE, Ritchie RH, Ferrell TL, Turner JE. Biopolymers. 1985; 24(3):427–439.
32. Kortemme T, Morozov AV, Baker D. J Mol Biol. 2003; 326(4):1239–1259. [PubMed: 12589766]

33. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D. Protein Sci. 2014; 23(1):47–55. [PubMed: 24265211]
34. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T. Database. 2013; 2013:bat031. [PubMed: 23624946]
35. Kryshtafovych A, Monastyrskyy B, Fidelis K. Proteins. 2014; 82(Suppl 2):7–13. [PubMed: 24038551]
36. Morozova O, Marra MA. Genomics. 2008; 92(5):255–264. [PubMed: 18703132]
37. Kellogg EH, Leaver-Fay A, Baker D. Proteins. 2011; 79(3):830–838. [PubMed: 21287615]
38. Yin S, Ding F, Dokholyan NV. Nat Methods. 2007; 4(6):466–467. [PubMed: 17538626]
39. Benedix A, Becker CM, de Groot BL, Caflisch A, Böckmann RA. Nat Methods. 2009; 6(1):3–4. [PubMed: 19116609]
40. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. J Chem Phys. 1983; 79(2): 926–935.
41. Wang L-P, Head-Gordon T, Ponder JW, Ren P, Chodera JD, Eastman PK, Martinez TJ, Pande VS. J Phys Chem B. 2013; 117(34):9956–9972. [PubMed: 23750713]
42. Mahoney MW, Jorgensen WL. J Chem Phys. 2000; 112(20):8910–8922.
43. Nguyen H, Maier J, Huang H, Perrone V, Simmerling C. J Am Chem Soc. 2014; 136(40):13959–13962. [PubMed: 25255057]
44. Gaillard T, Simonson T. J Comput Chem. 2014; 35(18):1371–1387. [PubMed: 24854675]

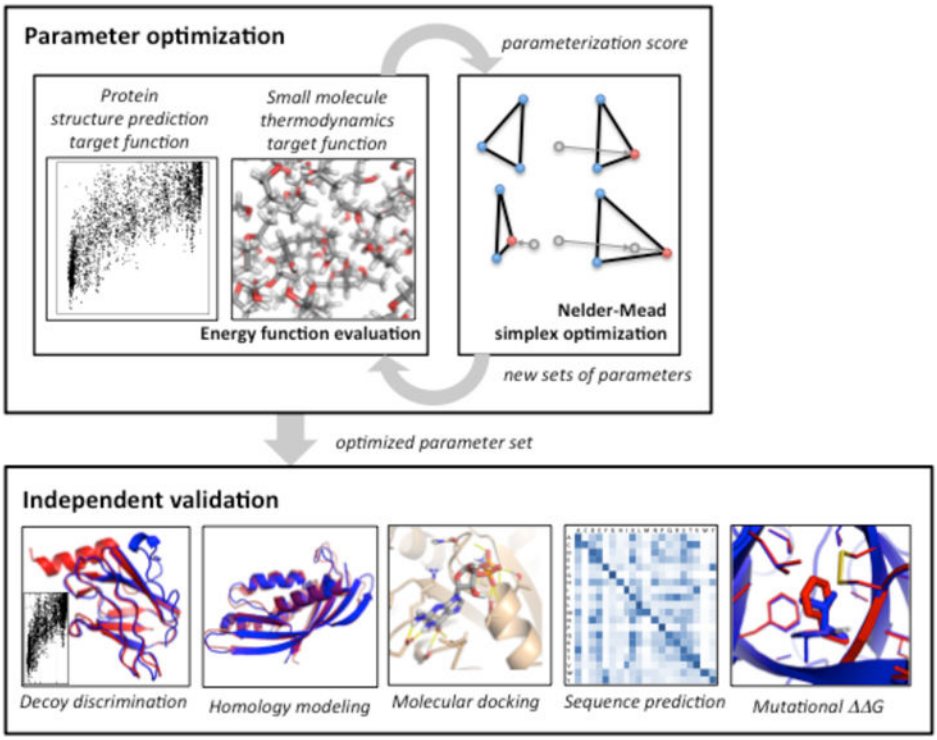

Figure 1.
A graphical overview of the parameter optimization procedure.

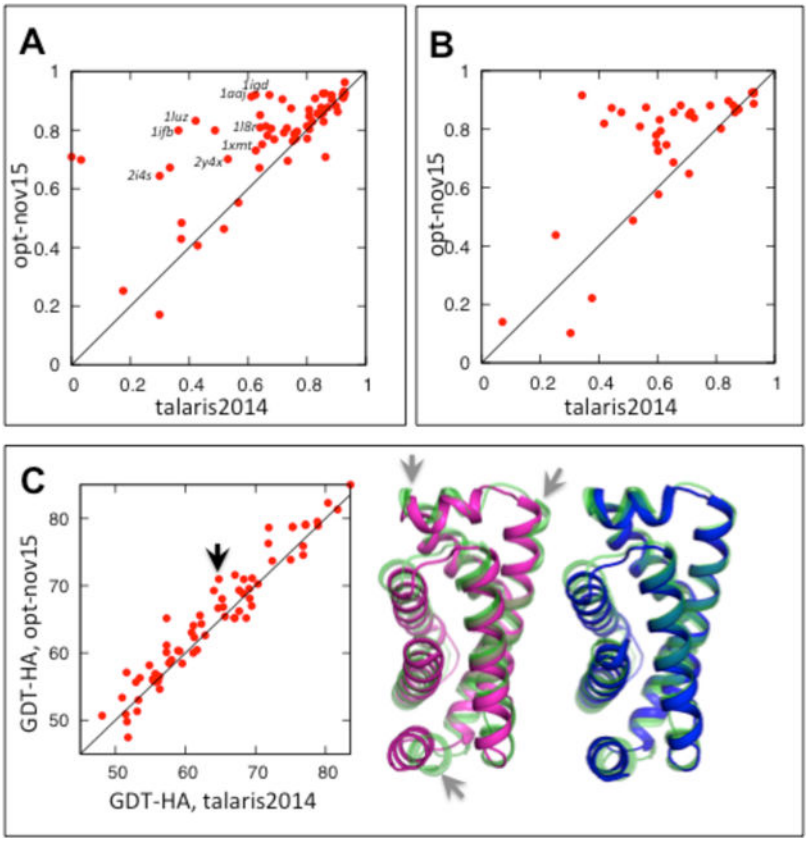

Figure 2.

Improvements in monomeric structure prediction from independent tests. In each scatter plot, a dot above diagonal line indicates improvement to a target. A) Decoy discrimination test. On the left, Boltzmann-weighted discrimination values are compared between *talaris2014* (X-axis) and *opt-nov15* (Y-axis) on 64 protein targets from validation set 2. B) *Parallel loophash sampling* (PLS) test. Boltzmann-weighted discrimination values are compared between *talaris2014* (X-axis) and *opt-nov15* (Y-axis) on 36 protein targets. C) Improved homology modeling on 69 CAMEO[34] targets using RosettaCM[2]. A comparison of homology model global distance test — high accuracy (GDT-HA)[35] from *talaris2014* (X-axis) and *opt-nov15* is shown on the left. GDT-HA is a measure of agreement (in %) of a model to its native structure in high-resolution. An example of highlight target is shown on the right (pointed out as black arrow): the structures generated and selected by *talaris2014* (magenta) and *opt-nov15* (blue) are overlaid on native (green) structure.

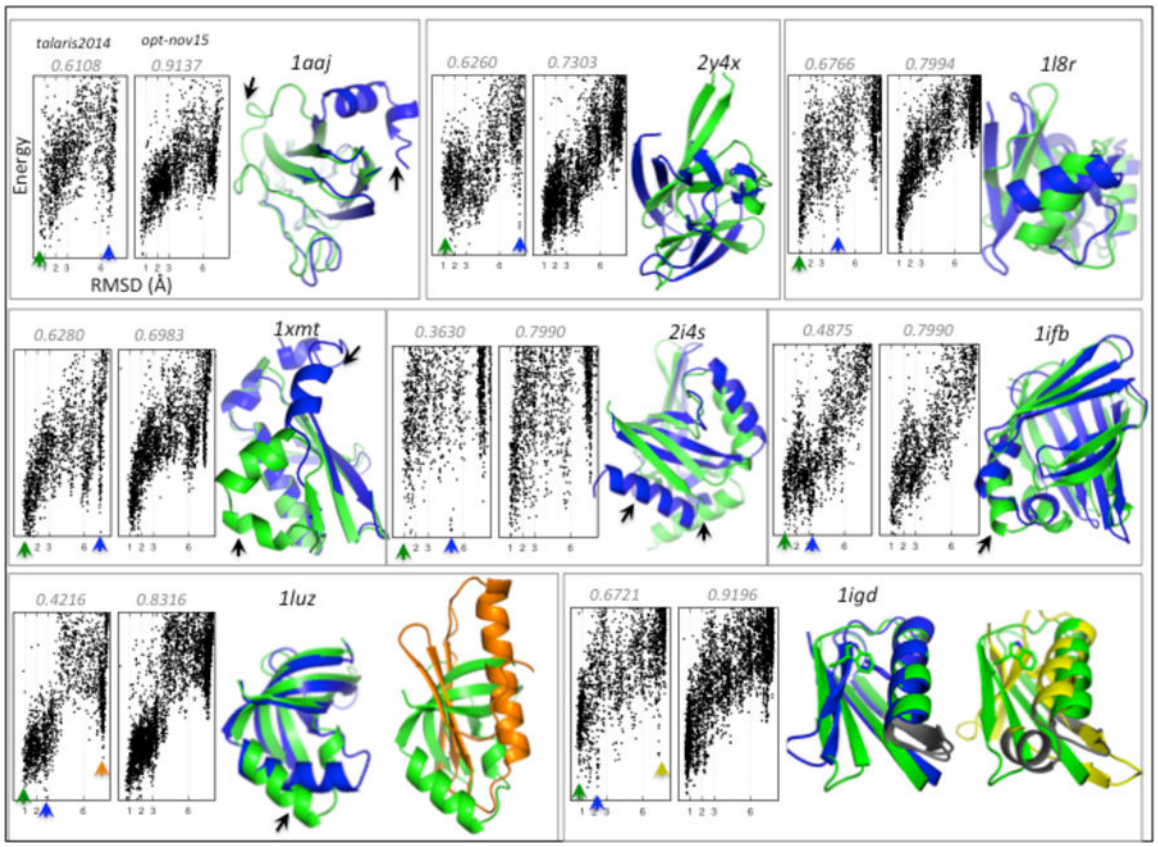

Figure 3.

Examples of proteins with successful recapitulation of energy landscapes. 8 cases from decoy discrimination set2 are shown (labeled in Figure 2A). For each case, energy landscapes by t*alaris2014* and *opt-nov15* are shown on the left, *talaris2014* on left and *opt-nov15* on right; at top of it Boltzmann-weighted discrimination values are shown with gray italic text. Each point in the energy landscape plots indicates a particular protein conformation, with the X-axis indicating the structural deviation from the native conformation, and the Y-axis indicating energy; in a good energy landscape the lowest energy conformations have lowest structural deviation. On the right, comparisons of conformations are shown with different colors: the near-native conformation in green, and low-energy false conformations in blue or orange. Energy values and RMSD for these conformations are shown as arrows with corresponding colors on the energy landscape.

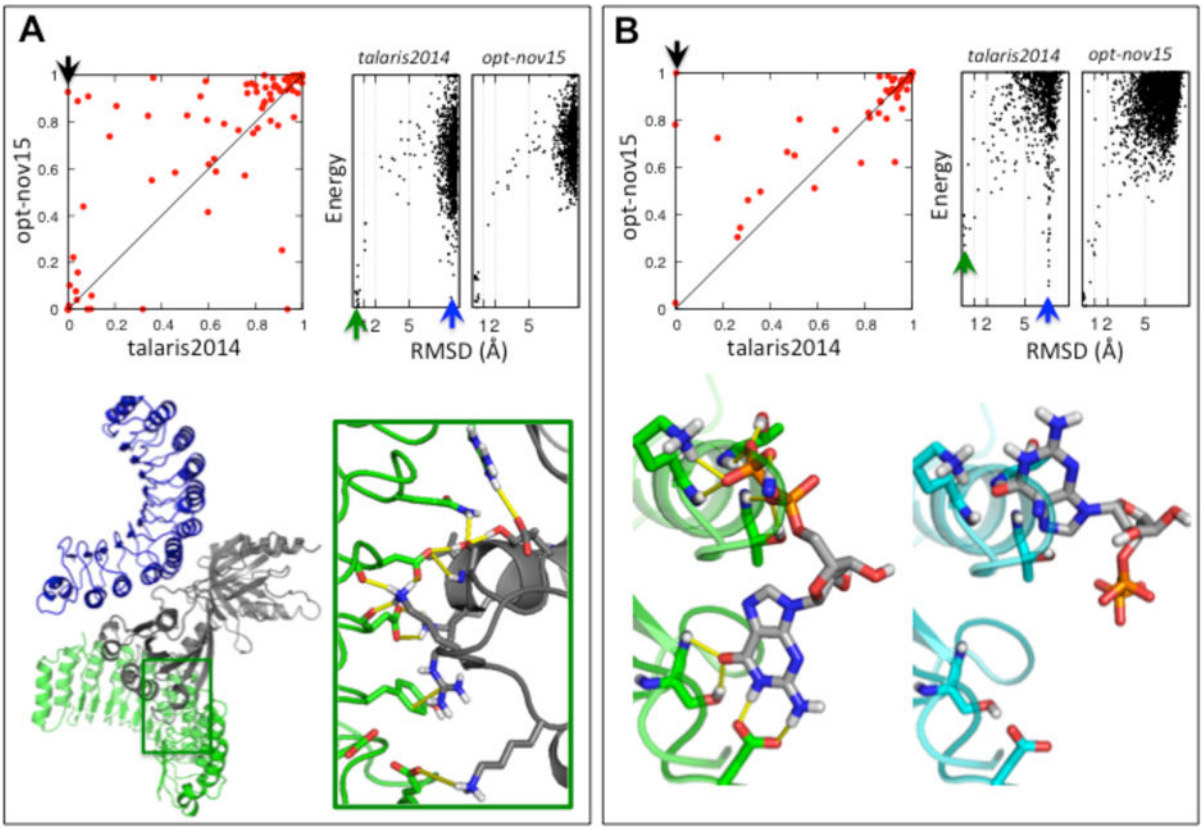

Figure 4.

Improvements in docking energy landscape recovery from independent tests. In each scatter plot, a dot above diagonal line means improvement to a target. A) Protein-protein docking and B) protein-ligand docking. In each panel, a scatter plot of the Boltzmann-weighted discrimination values for targets are plotted on the top left; an example energy landscape is shown in the top right, *talaris2014* on left and *opt-nov15* on right; and the corresponding structure pointed out as black arrow is on the bottom. A) Protein receptor is colored in gray, and alternative conformations of the partners are colored in green (near-native) or blue (false conformer). Favorable native interactions are highlighted in the inset. B) Ligands are colored in gray, and protein in green or cyan.

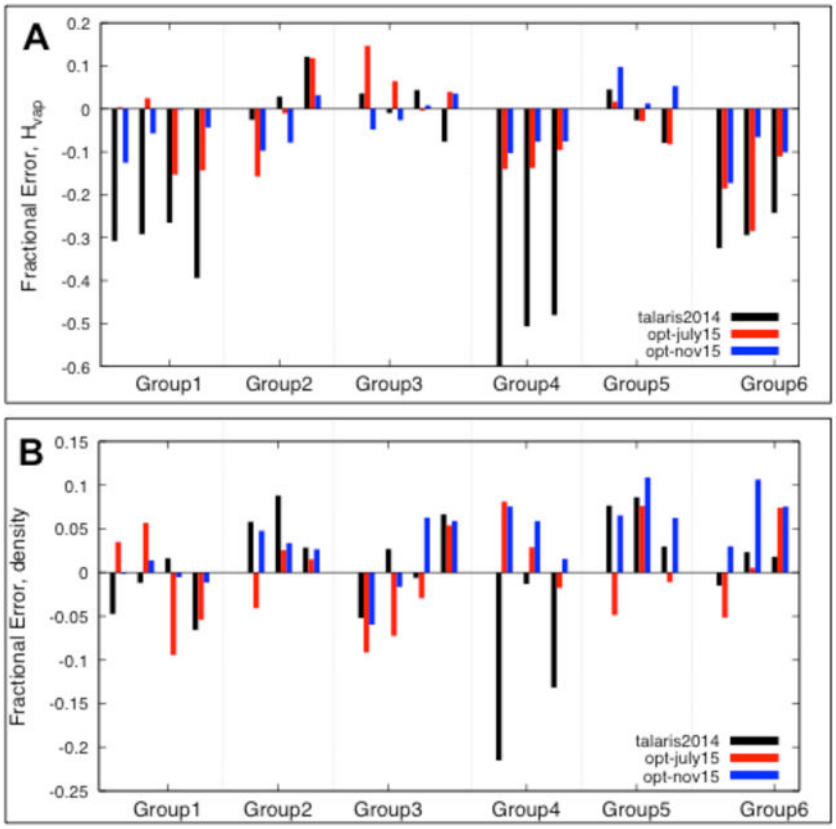

Figure 5.

Our optimized energy function reasonably recapitulates thermodynamic liquid properties of small molecules. A, B) Fractional errors in heat of vaporization (A) and density (B). Negative values indicate underestimation of reference experimental value, and positive overestimation. Colors indicate the energy function used: *talaris2014* (black), *opt-july15* (red), and *opt-nov15* (blue). Each bar corresponds to a small molecule in Figures S3. Most of the molecules represent functional groups in natural amino acids: aliphatic (Group 1), aromatic (Group 2), alcohol (Group 3), sulfide or thiol (Group 4), and amide (Group 5); several polar molecules contain functional groups not found in natural amino acids (Group 6).

Table 1

Performance of the energy functions on various tasks independent from training. Shown are weighted metrics, ranging from 0 to 1 and greater the better, with success rates (in percent) in parentheses, unless specified. Best value in each of the tasks is shown in bold. Details of the independent tests are described in Supplementary Materials.

Tasks		**talaris2014**	**opt-july15**	**opt-nov15**
Structure prediction	Decoy discrimination, set1 *1)*	0.580 (36.3)	0.648 (45.5)	**0.705 (57.1)**
	Decoy discrimination, set2 *1)*	0.686 (53.1)	0.725 (64.1)	**0.781 (67.2)**
	Parallel loophash sampling *1)*	0.639 (30.6)	0.666 (47.2)	**0.752 (63.9)**
	Homology modeling *2)*	63.9	—	**65.1**
Molecular docking	Protein-protein docking *1)*	0.717 (73.0)	0.762 (76.0)	**0.794 (81.0)**
	Protein-ligand docking *1)*	0.863 (82.1)	0.865 (88.1)	**0.941 (92.5)**
Sequence design	Protein monomer *3)*	0.258 (45.1)	0.270 (46.3)	**0.282 (47.0)**
	Protein-protein interface *3)*	0.283 (49.0)	0.304 **(51.3)**	**0.316** (51.0)
	Protein-ligand interface *3)*	0.390 (58.4)	0.411 (59.6)	**0.425 (60.3)**
	Full sequence design *4)*	38.9	39.6	**40.6**
	Mutational $\Delta\Delta G$ *5)*	0.704 (71.3)	**0.750** (72.6)	0.743 **(72.9)**
High-resolution geometry	Atom-pair distribution *6)*	0.00991	0.00972	**0.00796**
	Xtal gradient *7)*	0.230	0.174	**0.128**

1) Values are reported in average “Boltzmann-weighted discrimination” (see Supplementary Methods) over tested targets. Boltzmann-weighted discrimination measures the extent of discrimination of native-like conformation against non-native conformations, ranging from 0 (no discrimination) to 1 (complete discrimination).

2) Values are reported in average GDT-HA (global distance test — high accuracy)[35] of 69 tested targets.

3) Values are reported in entropy-weighted profile recovery (see Supplementary Methods). Higher values indicate better performance.

4) Success rate is shown only.

5) Values are reported in Pearson correlation coefficients, with classification rates in parenthesis.

6) KL-divergence of the atom-pair distance distribution after relaxation of structures against that from crystal structures. Weighted average for 93 most frequent atom pairs are reported. Lower values indicate better agreement.

7) Values are reported in normalized gradient (arbitrary unit). Lower values indicate better agreement. Data directly brought from optimization run.

Table 2

Differences in partial charges, LJ radii, and well-depths between various force fields. Mean absolute differences are reported.

Partial charges					
	CHARMM[1)]	**opt-from-CHARMM[2)]**	**OPLS[3)]**	**opt-from-OPLS[4)]**	**AMBER[5)]**
opt-nov15	0.049	0.038	0.088	0.049	0.093
CHARMM	—	0.020	0.056	0.080	0.078
opt-from-CHARMM	—	—	0.065	0.057	0.073
OPLS	—	—	—	0.044	0.118
opt-from-OPLS	—	—	—	—	0.092
LJ radii					
	CHARMM	opt-from-CHARMM	OPLS	opt-from-OPLS	AMBER
opt-nov15	0.062	0.046	0.061	0.082	0.071
CHARMM	—	0.027	0.066	0.065	0.096
opt-from-CHARMM	—	—	0.047	0.049	0.077
OPLS	—	—	—	0.081	0.081
opt-from-OPLS	—	—	—	—	0.122
LJ well-depths					
	CHARMM	opt-from-CHARMM	OPLS	opt-from-OPLS	AMBER
opt-nov15	0.015	0.009	0.014	0.013	0.024
CHARMM	—	0.018	0.018	0.025	0.035
opt-from-CHARMM	—	—	0.016	0.011	0.038
OPLS	—	—	—	0.011	0.024
opt-from-OPLS	—	—	—	—	0.019

1) CHARMM36[16]

2) Optimized parameters, started from CHARMM36 partial charges and LJ parameters.

3) OPLS-allatom[12]

4) Optimized parameters, started from OPLS-allatom partial charges and LJ parameters.

5) AMBER14ffSB[24]