The Minimized Dead-End Elimination Criterion and Its Application to Protein Redesign in a Hybrid Scoring and Search Algorithm for Computing Partition Functions over Molecular Ensembles

IVELIN GEORGIEV,[1] RYAN H. LILIEN,[2] BRUCE R. DONALD[1,3]
[1] *Department of Computer Science, Duke University, Durham, NC, USA*

[2] *Department of Computer Science and Banting–Best Department of Medical Research, University of Toronto, Canada*

[3] *Department of Biochemistry, Duke University Medical Center, Durham, NC, USA*

Received 6 June 2007; Revised 3 November 2007; Accepted 26 November 2007
DOI 10.1002/jcc.20909
Published online 21 February 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: One of the main challenges for protein redesign is the efficient evaluation of a combinatorial number of candidate structures. The modeling of protein flexibility, typically by using a rotamer library of commonly-observed low-energy side-chain conformations, further increases the complexity of the redesign problem. A dominant algorithm for protein redesign is dead-end elimination (DEE), which prunes the majority of candidate conformations by eliminating rigid rotamers that provably are not part of the global minimum energy conformation (GMEC). The identified GMEC consists of rigid rotamers (i.e., rotamers that have not been energy-minimized) and is thus referred to as the rigid-GMEC. As a postprocessing step, the conformations that survive DEE may be energy-minimized. When energy minimization is performed after pruning with DEE, the combined protein design process becomes heuristic, and is no longer provably accurate: a conformation that is pruned using rigid-rotamer energies may subsequently minimize to a lower energy than the rigid-GMEC. That is, the rigid-GMEC and the conformation with the lowest energy among all energy-minimized conformations (the minimized-GMEC) are likely to be different. While the traditional DEE algorithm succeeds in not pruning rotamers that are part of the rigid-GMEC, it makes no guarantees regarding the identification of the minimized-GMEC. In this paper we derive a novel, provable, and efficient DEE-like algorithm, called minimized-DEE (MinDEE), that guarantees that rotamers belonging to the minimized-GMEC will not be pruned, while still pruning a combinatorial number of conformations. We show that MinDEE is useful not only in identifying the minimized-GMEC, but also as a filter in an ensemble-based scoring and search algorithm for protein redesign that exploits energy-minimized conformations. We compare our results both to our previous computational predictions of protein designs and to biological activity assays of predicted protein mutants. Our provable and efficient minimized-DEE algorithm is applicable in protein redesign, protein-ligand binding prediction, and computer-aided drug design.

© 2008 Wiley Periodicals, Inc.   J Comput Chem 29: 1527–1542, 2008

Key words: dead-end elimination; rotameric energy minimization; protein design; conformational ensembles; partition functions; provably-accurate algorithms

Introduction

Computational Protein Design

The ability to engineer proteins has many biomedical applications. A number of computational approaches to the protein redesign problem have been reported. To improve the accuracy of the redesign, protein flexibility has been incorporated into most previous structure-based algorithms for protein redesign.[1–7] A study of bound and unbound structures found that most structural changes involve only a small number of residues and that these changes

Correspondence to: B. R. Donald; e-mail: brd+jcc07@cs.duke.edu

Contract/grant sponsor: The National Institutes of Health; contract/number: R01 GM-65982

Contract/grant sponsor: The National Science Foundation; contract/grant number: EIA-0305444

are primarily side-chains, and not backbone.[8] Hence, many protein redesign algorithms use a rigid backbone and model side-chain flexibility with a rotamer library that consists of a discrete set of low-energy commonly-observed side-chain conformations.[9,10] The major challenge for redesign algorithms is the efficient evaluation of the exponential number of candidate conformations, resulting not only from mutating residues along the peptide chain, but also by employing rotamer libraries. The development of pruning conditions capable of eliminating the majority of mutation sequences and conformations in the early, and less costly, redesign stages has been crucial.

GMEC-based algorithms for protein redesign are based on the assumption that protein folding and binding can be accurately predicted by examining the global minimum energy conformation (GMEC). Since identifying the GMEC using a model with a rigid backbone, a rotamer library, and a pairwise energy function is known to be NP-hard,[11,12] different heuristic approaches have been proposed.[1–4,13–15] A provable and efficient deterministic algorithm, which has become the dominant choice for GMEC-based protein design, is dead-end elimination (DEE).[16] DEE reduces the size of the conformational search space by eliminating rigid rotamers that provably are not part of the GMEC. Most important, since no protein conformation containing a dead-ending rotamer is generated, DEE provides a combinatorial factor reduction in computational complexity.

When energy minimization is performed after pruning with DEE, the process becomes heuristic, and is no longer provably accurate: a conformation that is pruned using rigid-rotamer energies may subsequently minimize to a structure with lower energy than the rigid-GMEC. Therefore, the traditional DEE conditions are not valid for pruning rotamers when searching for the lowest-energy conformation among all energy-minimized rotameric conformations (the minimized-GMEC, or minGMEC).

NRPS Redesign and K^*

Traditional ribosomal peptide synthesis is complemented by nonribosomal peptide synthetase (NRPS) enzymes in some bacteria and fungi. NRPS enzymes consist of several domains, each of which has a separate function. Substrate specificity is generally determined by the adenylation (A) domain.[17–19] Among the products of NRPS enzymes are natural antibiotics (penicillin, vancomycin), antifungals, antivirals, immunosuppressants, and antineoplastics. The main techniques for NRPS enzyme redesign are domain-swapping,[20–23] signature sequences,[17,18,24] and active site manipulation from a structure-based mutation search utilizing ensemble docking (the $\boldsymbol{K^*}$ method).[25]

The NRPS system discussed in this article is the phenylalanine adenylation domain of Gramicidin Synthetase A (GrsA-PheA), which, together with Gramicidin Synthetase B (GrsB), produces the natural antibiotic gramicidin S. The $\boldsymbol{K^*}$ algorithm has recently been used to gain new insights into the enzyme's mechanism and selectivity.[26] Redesigning GrsA-PheA to switch its specificity from the wildtype phenylalanine to a different substrate (e.g., Leu or Tyr) may produce a modified version of gramicidin. Thus, structure-based computational protein redesign can play a role in engineering combinatorial biosynthesis for small-molecule diversity. The redesign of NRPS enzymes can lead to the synthesis of novel NRPS products, such as new libraries of antibiotics.[27] More generally, novel molecular function can be achieved by redesigning an enzyme's active site so that it will perform its chemical reaction on a novel substrate.

The $\boldsymbol{K^*}$ algorithm[25] has been demonstrated for NRPS redesign, but is a general algorithm that is, in principle, capable of redesigning any protein. $\boldsymbol{K^*}$ is an ensemble-based scoring technique that uses a Boltzmann distribution to compute partition functions for the bound and unbound states of a protein. The ratio of the bound to the unbound partition function is used to compute a provably-good approximation ($\boldsymbol{K^*}$) to the binding constant for a given design sequence. A volume and a steric filter are applied in the initial stages of a redesign search to prune the majority of the conformations from more expensive evaluation. The number of evaluated conformations is further reduced by a provable ε-approximation algorithm. Protein flexibility is modeled for both the protein and the ligand using energy-minimization and rotamers.[25] In a recent study by Stevens et al.,[26] the $\boldsymbol{K^*}$ software was successfully applied in a redesign of GrsA-PheA: *in vitro* experiments showed that the top $\boldsymbol{K^*}$-predicted mutations improved the enzyme's specificity for a novel substrate.

Contributions of the Paper

Boltzmann probability implies that low-energy conformations are more likely to be assumed than high-energy conformations. The motivation behind energy minimization is therefore well-established and algorithms that incorporate energy minimization often lead to more accurate results. However, if energy minimization is performed after pruning with DEE, then the combined protein design process is heuristic, and not provable. We show that a conformation pruned using rigid-rotamer energies may subsequently minimize to surpass the putative rigid-GMEC.

We derive a novel, provable, and efficient DEE-like algorithm, called minimized-*DEE* (MinDEE), that guarantees that no rotamers belonging to the minGMEC will be pruned. We show that our method is useful not only in (a) identifying the minGMEC (a GMEC-based method), but also (b) as a filter in an ensemble-based scoring and search algorithm for protein redesign that exploits energy-minimized conformations. We achieve (a) by implementing a MinDEE/$\boldsymbol{A^*}$ algorithm in a search to switch the binding affinity of the Phe-specific adenylation domain of the NRPS Gramicidin Synthetase A (GrsA-PheA) towards Leu. The latter goal (b) is achieved by implementing MinDEE as a combinatorial filter in a hybrid algorithm,[†] combining $\boldsymbol{A^*}$ search and our previous work on $\boldsymbol{K^*}$.[25] The experimental results, based on a 2-point mutation search on the 9-residue active site of the GrsA-PheA enzyme, confirm that the new Hybrid MinDEE-$\boldsymbol{K^*}$ algorithm has a much higher pruning efficiency than the original $\boldsymbol{K^*}$ algorithm. Moreover, it takes only 30 s for MinDEE to determine which rotamers can be provably pruned. We make the following contributions in this paper:

1. Derivation of MinDEE, a novel, provable, and efficient DEE-like algorithm that incorporates energy minimization, with applications in both GMEC- and ensemble-based protein design.

[†] For brevity, we will henceforth refer to this algorithm as the *Hybrid MinDEE-$\boldsymbol{K^*}$* algorithm.

2. Introduction of a MinDEE/$\boldsymbol{A}^*$ algorithm that identifies the minGMEC and returns a set of low-energy conformations;
3. Introduction of a Hybrid MinDEE-$\boldsymbol{K}^*$ ensemble-based scoring and search algorithm, improving on our previous work on $\boldsymbol{K}^*$[25] by replacing a constant-factor with a combinatorial-factor provable pruning condition; and
4. The use of our novel algorithms in a redesign mutation search for switching the substrate specificity of the NRPS enzyme GrsA-PheA; we compare our results to previous computational predictions of protein designs and to biological activity assays of predicted protein mutants.

A preliminary version of this work was presented at a conference.[28] In ref. 29, nonoverlapping improvements to the current work and other algorithmic DEE enhancements are presented.

Derivation of the Minimized-DEE Criterion

The Original DEE Criterion

In this section we briefly review the traditional-DEE theorem.[16,30–32] Traditional-DEE refers to the original DEE, which is not provably correct when used in a search for the minimized-GMEC. Our notation is chosen to remain consistent with previous work. The total energy, E_{T}, of a given rotameric-based conformation can be written as $E_{\mathrm{T}} = E_{t'} + \sum_i E(i_r) + \sum_i \sum_{j>i} E(i_r, j_s)$, where $E_{t'}$ is the template self-energy (i.e., backbone energies or energies of rigid regions of the protein not subject to rotamer-based modeling), i_r denotes rotamer r at position i, $E(i_r)$ is the self energy of rotamer i_r (the intra-residue and residue-to-template energies), and $E(i_r, j_s)$ is the non-bonded pairwise interaction energy between rotamers i_r and j_s. The rotamers assumed in the rigid-GMEC are written with a subscript g. Therefore i_g is the rotamer assumed in the rigid-GMEC at position i. The following two bounds are then noted: for all $i, j (i \neq j)$, $\max_{s \in R_j} E(i_t, j_s) \geq E(i_t, j_g)$, and $\min_{s \in R_j} E(i_g, j_s) \leq E(i_g, j_g)$, where R_j is the set of allowed rotamers for residue j. For clarity, we will not include R_j in the limits of the max and min terms, since it will be clear from the notation from which set s must be drawn. The DEE criterion for rotamer i_r is defined as:

$$E(i_r) + \sum_{j \neq i} \min_s E(i_r, j_s) > E(i_t) + \sum_{j \neq i} \max_s E(i_t, j_s). \qquad (1)$$

Any rotamer i_r satisfying the DEE criterion (eq. 1) is provably not part of the rigid-GMEC ($i_r \neq i_g$), and is considered "dead-ending" (Fig. 1). Extensions to this initial DEE criterion allow for additional pruning while maintaining correctness with respect to identifying the rigid-GMEC.[16,30–33]

DEE with Energy Minimization: MinDEE

We now derive generalized DEE pruning conditions which can be used when searching for the minimized-GMEC. The fundamental difference between traditional-DEE and MinDEE is that the former enjoys significant independence among multiple energy terms during a rotamer swap. For example, when conformations are not energy-minimized, changing rotamer i_r to i_t does not affect the energy term $E(j_s)$; however, when energy minimization is allowed, the value of this energy term may change as the rotameric conformations i_r and j_s minimize from their initial rotameric conformations (Fig. 2). Therefore, to be provably correct, one must account for a range of possible energies. The conformation of a residue may change during energy minimization, however we constrain this movement to a region of conformation space called a voxel[34,35] to keep one rotamer from minimizing into another. In this framework, the voxel $\mathcal{V}(i_r)$ for rotamer i_r is simply all conformations of residue i within a $\pm\theta$ range around each rotamer dihedral when starting from the rotamer[‡] i_r. We similarly define the voxel $\mathcal{V}(i_r, j_s)$ for the pair of rotamers i_r and j_s to be the region of conformation space $\mathcal{V}(i_r) \times \mathcal{V}(j_s)$. Next, we can define the maximum, minimum, and range of voxel energies:

$$E_{\oplus}(i_r) = \max_{z \in \mathcal{V}(i_r)} E(z), \quad E_{\ominus}(i_r) = \min_{z \in \mathcal{V}(i_r)} E(z),$$
$$E_{\oslash}(i_r) = E_{\oplus}(i_r) - E_{\ominus}(i_r).$$

Analogous definitions exist for pairwise terms:

$$E_{\oplus}(i_r, j_s) = \max_{z \in \mathcal{V}(i_r, j_s)} E(z), \quad E_{\ominus}(i_r, j_s) = \min_{z \in \mathcal{V}(i_r, j_s)} E(z),$$
$$E_{\oslash}(i_r, j_s) = E_{\oplus}(i_r, j_s) - E_{\ominus}(i_r, j_s).$$

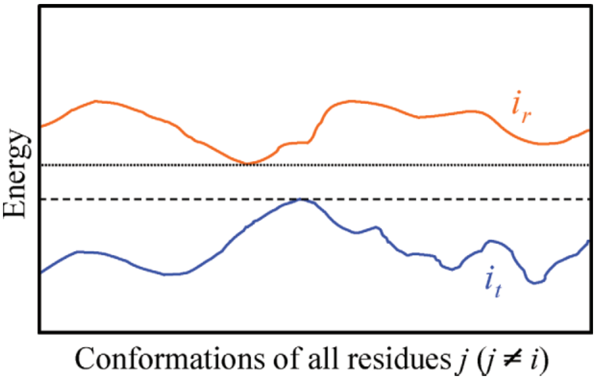


Figure 1. Rotamer pruning by dead-end elimination. A cartoon of the protein's conformational energy for all conformations of residues j ($j \neq i$) assuming the presence of rotamer r (orange, top) and rotamer t (blue, bottom) at position i. In this example, the lowest (best) conformational energy achievable with rotamer i_r is indicated by the dotted line and the highest (worst) conformational energy achievable with rotamer i_t is indicated by the dashed line. Since the energy of all conformations is reduced in switching from i_r to i_t, rotamer i_r can be pruned as dead-ending. In practice, the use of eq. (1) avoids the requirement of having to enumerate the exponential number of possible conformations for all residues j ($j \neq i$). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

[‡] The voxel space for each rotamer can be multi-dimensional, depending on the number of dihedrals. The largest number of dihedrals for a single rotamer is 4 (Arg and Lys).

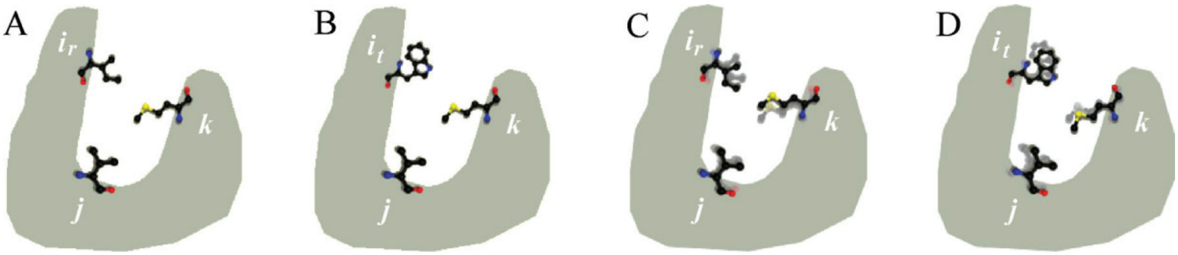


Figure 2. Energy-minimized DEE. Without energy minimization the swapping of rotamer i_r for i_t (Panel A to Panel B) leaves unchanged the conformations and self and pairwise energies of residues j and k. When energy minimization is allowed, the swapping of rotamer i_r for rotamer i_t (Panel C to Panel D) may cause the conformations of residues j and k to minimize (i.e., move) to form more energetically favorable interactions (from the faded to the solid conformations in Panels C and D).

We now define the MinDEE criterion for rotamer i_r to be:

$$E_{\ominus}(i_r) + \sum_{j \neq i} \min_{s} E_{\ominus}(i_r, j_s) - \sum_{j \neq i} \max_{s} E_{\oslash}(j_s)$$
$$- \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_{\oslash}(j_s, k_u) > E_{\oplus}(i_t) + \sum_{j \neq i} \max_{s} E_{\oplus}(i_t, j_s). \qquad (2)$$

Proposition 1. *When eq. (2) holds, rotamer i_r is provably not part of the minimized-GMEC.*

The proof of Proposition 1 is given in Appendix A.

The most significant difference between traditional-DEE and MinDEE is the accounting for possible energy changes during minimization, which are incorporated through the introduction of the terms $\sum_j \max_s E_{\oslash}(j_s)$ and $\sum_j \sum_k \max_{s,u} E_{\oslash}(j_s, k_u)$. Using precomputed energy bounds, the MinDEE pruning condition [eq. (2)] can be computed as efficiently as the traditional-DEE pruning condition [eq. (1)]. The complexity of deciding eq. (2) is $O(q^2 n^2)$, where n is the number of residue positions and q is the maximum number of rotamers per residue position. The MinDEE framework can be used whenever a bound on a pairwise energy function can be obtained and is therefore not critically dependent upon the particular energy function or type of minimization employed.

In this section, we presented a generalization of traditional-DEE, to obtain an initial pruning criterion for MinDEE. Extensions to the traditional-DEE pruning conditions have made them more efficient.[16,30–33] An excellent review of these advanced pruning techniques appears in ref. 30. These methods allow more individual rotamers to be pruned during DEE and extend the DEE criterion to identify dead-ending rotamer pairs. Analogously to section "DEE with energy minimization: MinDEE", we have derived MinDEE equivalents to four extensions to traditional DEE for increased pruning efficiency.[29]

Two Applications of MinDEE

The MinDEE criterion can efficiently reduce the search space for a given protein design problem by pruning rotamers that are provably not part of the minimized-GMEC. We have applied MinDEE as a pruning filter in two different protein design algorithms. The details of these algorithms are described in the following two sections. In section "MinDEE/A^* search algorithm (GMEC-based Redesign)", we use MinDEE as a part of MinDEE/$\boldsymbol{A}^*$, a provably-accurate GMEC-based protein design algorithm. In the MinDEE/$\boldsymbol{A}^*$ algorithm, MinDEE is first used to provably prune the majority of the candidate conformations; the minimized-GMEC and all low-energy conformations (and thus sequences) within a specified threshold from the minimized-GMEC energy are then generated and energy-minimized using the $\boldsymbol{A}^*$ search technique.[36] MinDEE/$\boldsymbol{A}^*$ ranks mutation sequences[¶] based on the energy of the single best conformation for each mutation sequence (hence, the term GMEC-based algorithm). In section "MinDEE/A^* search algorithm (GMEC-based redesign)", we first review the traditional-DEE/$\boldsymbol{A}^*$ GMEC-based algorithm proposed in ref. 36 (section "Traditional-DEE with A^*"); we then derive the MinDEE/$\boldsymbol{A}^*$ algorithm that, in contrast to traditional-DEE/$\boldsymbol{A}^*$, is provably-accurate with rotameric energy minimization (section "MinDEE with A^*").

In section "Hybrid MinDEE-K^* algorithm (ensemble-based redesign)", we describe how MinDEE can be used as a pruning filter in Hybrid MinDEE-$\boldsymbol{K}^*$, an *ensemble-based* protein design algorithm. For a given protein-ligand complex, Hybrid MinDEE-$\boldsymbol{K}^*$ computes a provably-accurate approximation, $\boldsymbol{K}^*$, to the association binding constant by computing Boltzmann-weighted partition functions over rotameric *ensembles* of conformations. Given a set of candidate mutation sequences and a target ligand, Hybrid MinDEE-$\boldsymbol{K}^*$ computes the $\boldsymbol{K}^*$ scores for each sequence and ranks sequences in order of their computed scores (higher scores imply better binding). In the beginning of section "Hybrid MinDEE-K^* algorithm (ensemble-based redesign)", we discuss the general motivation behind the Hybrid MinDEE-$\boldsymbol{K}^*$ algorithm. In section "Efficient partition function computation using A^* search", we derive a provably-accurate algorithm for partition function computation over conformational ensembles that also exploits MinDEE pruning and the $\boldsymbol{A}^*$ search; in Appendix B, we present an improvement to the partition function computation algorithm of section "Efficient partition function computation using A^* search". In particular, the efficient partition function computation is generalized to prune rotamers and sequences, so that in protein redesign the optimal sequences (in terms of $\boldsymbol{K}^*$ score) are computed. Finally, in section "Algorithm", we describe the application of the partition

[¶] A mutation sequence specifies an assignment of amino-acid type to each residue position in a protein.

function computation algorithms in Hybrid MinDEE-$\boldsymbol{K}^*$, as well as the complete sequence of Hybrid MinDEE-$\boldsymbol{K}^*$ algorithmic steps.

MinDEE/A^* Search Algorithm (GMEC-Based Redesign)

*Traditional-DEE with A**

In ref. 36, an $\boldsymbol{A}^*$ branch-and-bound algorithm was developed to compute a number of low-energy conformations for a single mutation sequence (i.e., a single protein). In this algorithm, traditional-DEE was first used to reduce the number of side-chain conformations, and then surviving conformations were enumerated in order of conformation energy by expanding sorted nodes of a conformation tree (Fig. 3).

The following derivation of the DEE/$\boldsymbol{A}^*$ combined search closely follows.[36] The $\boldsymbol{A}^*$ algorithm scores each node in a conformation tree using a scoring function $f = g + h$, where g is the cost of the path from the root to that node (the energy of all self and pairwise terms assigned through depth d) and h is an estimate (lower bound) of the path cost to a leaf node (a lower bound on the sum of energy terms involving unassigned residues). The value of g (at depth d) can be expressed as $g = \sum_{i=1}^{d}(E(i_r) + \sum_{j=i+1}^{d} E(i_r,j_s))$. The lower bound h can be written as $h = \sum_{j=d+1}^{n} E_j$, where n is the total number of flexible residues and $E_j = \min_s(E(j_s) + \sum_{i=1}^{d} E(i_r,j_s) + \sum_{k>j}^{n} \min_u E(j_s,k_u))$. The $\boldsymbol{A}^*$ algorithm maintains a list of nodes (sorted by f) and in each iteration replaces the node with the smallest f value by an expansion of the children of that node. This process of expansion is continued until the node with the smallest f value is a leaf node. This leaf node corresponds to a fully-assigned conformation and is returned by the algorithm. To reduce the branching factor of the conformation tree, the DEE algorithm is used to preprocess the set of allowed rotamers. If more than one low-energy conformation is to be extracted from the A* search, the DEE criterion must be modified. If low-energy conformations within E_w of the GMEC are to be returned by the DEE/A* search, then the DEE criterion must be modified to only eliminate rotamers that are provably not part of any conformation within E_w of the GMEC. The original DEE criterion [eq. (1)] is thus changed to: $E(i_r) - E(i_t) + \sum_{j\neq i} \min_s E(i_r,j_s) - \sum_{j\neq i} \max_s E(i_t,j_s) > E_w$.

*MinDEE with A**

The traditional-DEE/$\boldsymbol{A}^*$ algorithm[36] can be extended to include energy minimization by substituting our newly derived MinDEE (section "DEE with energy minimization: MinDEE") for traditional-DEE. So that no conformations within E_w of the energy-minimized GMEC are pruned, the MinDEE equation [eq. (2)] becomes:

$$E_\ominus(i_r) + \sum_{j\neq i} \min_s E_\ominus(i_r,j_s) - \sum_{j\neq i} \max_s E_\oslash(j_s)$$
$$-\sum_{j\neq i} \sum_{k\neq i,k>j} \max_{s,u} E_\oslash(j_s,k_u) - E_\oplus(i_t) - \sum_{j\neq i} \max_s E_\oplus(i_t,j_s) > E_w. \tag{3}$$

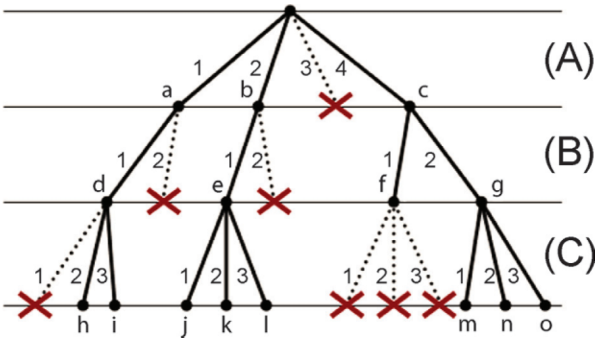


Figure 3. An example conformation tree. In a conformation tree, the rotamers of flexible residue i are represented by the branches at depth i. Internal nodes of a conformation tree represent partially-assigned conformations and each leaf node represents a fully-assigned conformation. Nodes marked with ×s have been pruned from further consideration. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

We modify the definition of the $\boldsymbol{A}^*$ functions g and h to use the minimum energy terms $E_\ominus(i_r)$ and $E_\ominus(i_r,j_s)$ in place of $E(i_r)$ and $E(i_r,j_s)$. Thus, we have:

$$g = \sum_{i=1}^{d}\left(E_\ominus(i_r) + \sum_{j=i+1}^{d} E_\ominus(i_r,j_s)\right), \quad h = \sum_{j=d+1}^{n} E_j, \tag{4}$$

where

$$E_j = \min_s\left(E_\ominus(j_s) + \sum_{i=1}^{d} E_\ominus(i_r,j_s) + \sum_{k=j+1}^{n} \min_u E_\ominus(j_s,k_u)\right). \tag{5}$$

A lower bound on the minimized energy of the partially-assigned conformation is given by g, while a lower bound on the minimized energy for the unassigned portion of the conformation is given by h. Thus, the MinDEE/$\boldsymbol{A}^*$ search generates conformations in order of increasing lower bounds on the conformation's minimized energy.

We combine our modified MinDEE criterion [eq. (3)] with the modified $\boldsymbol{A}^*$ functions [eqs. (4) and (5)] in a provable search algorithm for identifying the minimized-GMEC and obtaining a set of low-energy conformations. First, MinDEE prunes the majority of the conformations by eliminating rotamers that are provably not within E_w of the minimized-GMEC. The remaining conformations are then generated in order of increasing lower bounds on their minimized energies. The generated conformations are energy-minimized and ranked in terms of increasing actual minimized energies. The single best conformation for each unique mutation sequence is then used to rank the mutation sequence predictions.

The MinDEE/$\boldsymbol{A}^*$ search must guarantee that upon completion, all conformations within E_w of the minimized-GMEC are returned. Since in the $\boldsymbol{A}^*$ algorithm conformations are returned in order of increasing lower bounds on the minimized energies, the minimized-GMEC may not be among the top conformations if the lower bound on its energy does not rank high. We therefore derive the following condition for halting the MinDEE/$\boldsymbol{A}^*$ search. Let $B(s)$ be the lower bound on the energy of conformation s (see Appendix C, which describes how lower energy bounds are precomputed for all rotamer pairs) and let E_{m} be the current minimum energy among the minimized conformations returned so far in the $\boldsymbol{A}^*$ search.

Proposition 2. *The MinDEE/$\boldsymbol{A}^*$ search can be halted once the lower bound $B(c)$ on the energy of the next conformation c returned by $\boldsymbol{A}^*$, satisfies $B(c) > E_{\mathrm{m}} + E_{\mathrm{w}}$. The set of returned conformations is guaranteed to contain every conformation whose energy is within E_{w} of the energy of the minimized-GMEC. Moreover, at that point in the search, the conformation with energy E_{m} is the minimized-GMEC.*

Proof. Let $E(s)$ be the actual energy of a minimized conformation s. Let Y be the set containing conformation c (the next conformation returned by $\boldsymbol{A}^*$) and all conformations not yet returned. Since $\boldsymbol{A}^*$ returns conformations *in order* of increasing lower bounds on the energy, we know that $E(s) \geq B(s) \geq B(c)$ for any conformation $s \in Y$. Thus, if $B(c) > E_{\mathrm{m}} + E_{\mathrm{w}}$ holds, then $E(s) > E_{\mathrm{m}}+E_{\mathrm{w}}$. Hence, no conformations in Y have energies within E_{w} of the energy of the minimized-GMEC, proving that all conformations within E_{w} of the minimized-GMEC energy have already been returned. Moreover, note that at that point in the search, the conformation with energy E_{m} is actually the minimized-GMEC. ■

Using both MinDEE and $\boldsymbol{A}^*$ search together, our algorithm obtains a combinatorial pruning factor by eliminating the majority of the conformations, which makes the search for the minimized-GMEC computationally feasible. The MinDEE/$\boldsymbol{A}^*$ algorithm incorporates energy minimization with provable guarantees, and is thus more capable of returning conformations with lower energy states than traditional-DEE.

Hybrid MinDEE-$\boldsymbol{K}^*$ Algorithm (Ensemble-Based Redesign)

We now present an extension and improvement to the original $\boldsymbol{K}^*$ protein design algorithm[25] by using a version of the MinDEE criterion plus $\boldsymbol{A}^*$ branch-and-bound search. The $\boldsymbol{K}^*$ ensemble-based scoring function approximates the association binding constant for a given protein-ligand complex with the following quotient: $K^* = \frac{q_{\mathrm{PL}}}{q_{\mathrm{P}} q_{\mathrm{L}}}$, where q_{PL}, q_{P}, and q_{L} are the partition functions for the protein-ligand complex, the free (unbound) protein, and the free ligand, respectively. For a given protein design problem, partition functions and $\boldsymbol{K}^*$ scores are efficiently computed for all candidate mutation sequences with the target ligand; sequences are then ranked in order of their computed $\boldsymbol{K}^*$ scores (higher scores imply better binding). In this section, we describe how our MinDEE pruning criterion and the $\boldsymbol{A}^*$ search can be exploited for the partition function and $\boldsymbol{K}^*$ computation.

A partition function q over a set (ensemble) of conformations S is defined as $q = \sum_{s \in S} \exp(-E_s/RT)$, where E_s is the energy of conformation s, T is the temperature in Kelvin, and R is the gas constant. In a naive $\boldsymbol{K}^*$ implementation, each partition function would be computed by a computationally-expensive energy minimization of all rotamer-based conformations. However, because the contribution to the partition function of each conformation is exponential in its energy, only a subset of the conformations significantly contribute to the partition function value. By identifying and energy-minimizing *only* the significantly-contributing conformations, a provably-accurate ε-approximation algorithm substantially improved the algorithm's efficiency[25]. In this section we illustrate how the newly-derived MinDEE and $\boldsymbol{A}^*$ algorithms (section "MinDEE with A^*") can be used to generate and minimize only those conformations that contribute significantly to the partition function, and hence, for which energy minimization is required. The MinDEE criterion must be used in this algorithm because the $\boldsymbol{K}^*$ scoring function is based on energy-minimized conformations. Since pruned conformations never have to be examined, the Hybrid MinDEE-$\boldsymbol{K}^*$ algorithm provides a combinatorial improvement in runtime over the previously-described constant-factor ε-approximation algorithm[25] (where a lower-bound on *each* conformation's minimum energy was quickly examined to determine if full energy minimization was required).

Efficient Partition Function Computation Using A^* Search

Here, we present an efficient algorithm for computing the q_{PL}, q_{P}, and q_{L} partition functions used to compute a $\boldsymbol{K}^*$ approximation score for a given mutation sequence. Using the $\boldsymbol{A}^*$ algorithm with MinDEE, we can generate the conformations of a rotamerically-based ensemble in order of increasing lower bounds on the conformation's minimized energy. We can efficiently compute the lower bound on a conformation's energy as a sum of precomputed pairwise minimum energy terms (see Appendix C). As each conformation c is generated from the conformation tree, we compare its lower bound $B(c)$ on the conformational energy to a moving *stop-threshold* and halt the $\boldsymbol{A}^*$ search once $B(c)$ becomes greater than the threshold. The $\boldsymbol{A}^*$ algorithm guarantees that all remaining conformations will have minimized energies above the stop-threshold. We now prove that a partial partition function q^* computed using only those conformations with energies below (i.e., better than) the stop-threshold will lie within a factor of ε of the true partition function q. Note that, by definition, $q \geq q^*$. Thus, q^* is an ε-approximation to q, i.e., $q^* \geq (1-\varepsilon)q$.

Since the application of the MinDEE criterion (Eq. 2) for each rotamer i_r requires that the corresponding minimum energy terms be accessed, we can easily piggyback the computation of a lower bound B_{i_r} on the energy of all conformations that contain a pruned rotamer i_r:

$$B_{i_r} = E_{t'} + E_{\ominus}(i_r) + \sum_{j \neq i} \min_{s} E_{\ominus}(j_s) + \sum_{j \neq i} \min_{s} E_{\ominus}(i_r, j_s) + \sum_{j \neq i} \sum_{k \neq i, k > j} \min_{s,u} E_{\ominus}(j_s, k_u).$$

Let E_0 be the minimum lower energy bound among all conformations containing at least one pruned rotamer, $E_0 = \min_{i_r \in S} B_{i_r}$, where S is the set of pruned *rotamers*. E_0 can be precomputed during the MinDEE stage and prior to the $\boldsymbol{A}^*$ search. Let p^* be the partition function computed over the set P of pruned *conformations*, so that $p^* \leq k\exp(-E_0/RT)$, where $|P| = k$. Also, let X be the set of conformations not pruned by MinDEE and let q^* be the partition function for the top m conformations already returned by $\boldsymbol{A}^*$; let q' be the partition function for the n conformations that have not yet been generated, all of which have energies above E_t, so that $q' \leq n\exp(-E_t/RT)$; note that $|X| = m + n$. Finally, let $\rho = \frac{\varepsilon}{1-\varepsilon}$. We can then guarantee an ε-approximation to the full partition function q using:

Proposition 3. *If the lower bound $B(c)$ on the minimized energy of the $(m + 1)^{\text{st}}$ conformation returned by $\boldsymbol{A}^*$ satisfies $B(c) \geq -RT(\ln(q^*\rho - k\exp(-E_0/RT)) - \ln n)$, then the partition function computation can be halted, with q^* guaranteed to be an ε-approximation to the true partition function q, that is, $q^* \geq (1 - \varepsilon)q$.*

Proof. The full partition function q is computed using all conformations in both P and X:

$$q = q^* + q' + p^*. \tag{6}$$

Thus,

$$q \leq q^* + n\exp(-E_t/RT) + k\exp(-E_0/RT). \tag{7}$$

Hence, if

$$q^* \geq (1 - \varepsilon)(q^* + n\exp(-E_t/RT) + k\exp(-E_0/RT)), \tag{8}$$

then $q^* \geq (1 - \varepsilon)q$. Solving Eq. (8) for E_t, we obtain the desired stop-threshold:

$$-RT(\ln(q^*\rho - k\exp(-E_0/RT)) - \ln n) \leq E_t. \tag{9}$$

We can halt the search once a conformation's energy lower bound becomes greater than the stop-threshold [eq. (9)], since then q^* is already an ε-approximation to q. ■

The application of the MinDEE criterion gives a combinatorial-factor speedup by caching the minimum lower energy bound for the set of all pruned conformations. Since the conformations pruned by MinDEE can potentially contribute significantly to the partition function, we bound their contribution, thus guaranteeing a provably-accurate approximation to the full partition function. The conformation tree could, in principle, be reduced by pruning an *arbitrary* subset of the rotamers, so long as a guarantee on the accuracy is still maintained through a bound on the contribution of the pruned conformations. However, in practice, the amount of pruning and the resulting approximation accuracy depend on which rotamers are chosen for pruning. Using MinDEE to determine the set of pruned rotamers guarantees that the pruned conformations will have high lower energy bounds by requiring that no conformations within E_w of the minimized-GMEC energy are pruned [eq. (3)], whereas an arbitrary rotameric set could easily contain conformations with very good (i.e., low) energies. Proposition 3 turns pruning with MinDEE into a provable heuristic. Note that: (1) the magnitude of p^* is determined by the lower energy bounds of the pruned conformations, and (2) the number of conformations that $\boldsymbol{A}^*$ must extract to guarantee a provably-accurate approximation to the partition function depends on the magnitude of p^*. By using MinDEE pruning instead of an arbitrary set of rotamers, we increase the pruning efficiency. Since conformations that contain steric clashes do not contribute to the partition function for the given mutation sequence, we can further reduce p^* by including in P only the pruned conformations whose lower energy bound does not contain a rotamer that always clashes sterically (such a reduction in P, and hence, k, can be computed during the MinDEE phase, since rotamers whose precomputed minimum-energy bounds indicate steric clashes, necessarily imply that all conformations containing these rotamers are also steric clashes).

If at some point in the search, the stop-threshold condition has not been reached and there are no remaining conformations for $\boldsymbol{A}^*$ to extract ($n = 0$), then $q' = 0$ by definition, and $q = q^* + p^*$. Hence, if $q^*\rho \geq k\exp(-E_0/RT)$, then $q^* \geq (1-\varepsilon)(q^* + k\exp(-E_0/RT))$, so $q^* \geq (1 - \varepsilon)q$ is already an ε-approximation to q; otherwise, we have

$$q^* \geq (1 - \delta)(q^* + k\exp(-E_0/RT)), \tag{10}$$

for some approximation accuracy $\delta > \varepsilon$. Thus, the set of pruned rotamers must be reduced to guarantee the desired approximation accuracy. To assure that an ε-approximation is achieved when the search is repeated, a subset of the k pruned conformations in P must be reintroduced into the computation. Let l be the number of conformations from P (the set of pruned conformations) that are not to be pruned, such that $p^* \leq (k - l)\exp(-E_0/RT)$. We will conservatively assume that the l conformations do not contribute to q^*, although they no longer contribute to p^* either. At the end of the second mutation search, we must have

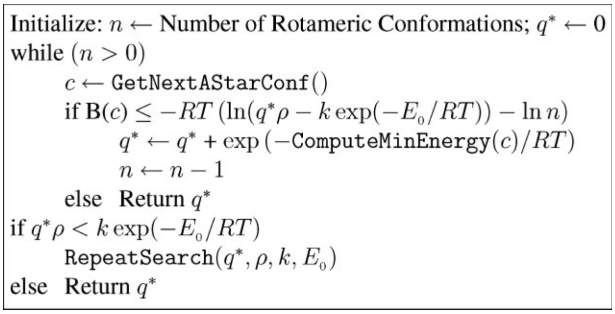
$$q^* \geq (1 - \varepsilon)(q^* + (k - l)\exp(-E_0/RT)). \tag{11}$$

Solving for l, we obtain the following condition, which guarantees the desired ε-approximation accuracy:

$$l \geq k - \frac{q^*\rho}{\exp(-E_0/RT)}, \tag{12}$$

where again $\rho = \frac{\varepsilon}{1-\varepsilon}$. Note that an ε-approximation may be achieved before all conformations have been extracted; eq. (12) guarantees such an accuracy when all non-pruned conformations have been extracted by $\boldsymbol{A}^*$. To guarantee that at least l out of the k pruned conformations will be allowed during the repeated computation, we can choose a subset Q of the rotamers pruned by MinDEE, such that not pruning Q keeps at least l additional conformations.

In the algorithm for partition function computation described in this section, conformation pruning is performed only within a mutation sequence (Fig. 4). In Appendix B, we derive an improvement to this partition function algorithm that further improves the



```
Initialize: $n \leftarrow$ Number of Rotameric Conformations; $q^* \leftarrow 0$
while $(n > 0)$
    $c \leftarrow$ GetNextAStarConf()
    if $\mathrm{B}(c) \le -RT\,(\ln(q^*\rho - k\exp(-E_0/RT)) - \ln n)$
        $q^* \leftarrow q^* + \exp(-\mathtt{ComputeMinEnergy}(c)/RT)$
        $n \leftarrow n - 1$
    else  Return $q^*$
if $q^*\rho < k\exp(-E_0/RT)$
    RepeatSearch$(q^*, \rho, k, E_0)$
else  Return $q^*$
```

Figure 4. Efficient partition function computation with energy minimization using the A^* search. q^* is the running approximation to the partition function. The function `B(·)` computes the energy lower bound for the given conformation (see Appendix C). The function `ComputeMinEnergy(·)` returns a conformation's energy after energy minimization. The function `GetNextAStarConf()` returns the next conformation from the A^* search. The function `RepeatSearch(·)` sets up and repeats the mutation search if an ε-approximation is not achieved after the generation of all $\boldsymbol{A}^*$ conformations; the search is repeated at most once. Upon completion, q^* represents an ε-approximation to the true partition function q, such that $q \ge q^* \ge (1-\varepsilon)q$.

efficiency of the partition function computation by allowing conformation pruning across mutation sequences. The improved algorithm in Appendix B also yields a provably-good approximation (see Proposition 4 therein).

Algorithm

We now have all the necessary tools for our ensemble-based Hybrid MinDEE-$\boldsymbol{K}^*$ algorithm. The volume filter (see Methods section) in the original $\boldsymbol{K}^*$ is applied first to eliminate under- and over-packed mutation sequences. For each of the remaining unpruned sequences, the $K^* = \frac{q_{\mathrm{PL}}}{q_{\mathrm{P}} q_{\mathrm{L}}}$ scores are computed, using the partition function algorithms of section "Efficient partition function computation using A^* search" and Appendix B to efficiently compute the q_{PL}, q_{P}, and q_{L} partition functions. The application of the MinDEE and $\boldsymbol{A}^*$ algorithms in the partition function computation improves on the mere constant-factor speedup provided by the energy filter in the original $\boldsymbol{K}^*$ algorithm.[25] By implementing a steric filter (see Methods section), similar to the one in ref. 25, as a part of the $\boldsymbol{A}^*$ search, we prevent some high-energy conformations (corresponding to steric clashes) with good lower bounds from being returned by $\boldsymbol{A}^*$, gaining an additional combinatorial speedup. Only the conformations that pass all of these filters are energy-minimized and used in the computation of the partition function for the conformational ensemble. In contrast to the original $\boldsymbol{K}^*$ algorithm[25] where, for a given mutation sequence, pruning was performed *during* the (worst-case exponential) conformation enumeration, Hybrid MinDEE-$\boldsymbol{K}^*$ uses the polynomial-time MinDEE criterion *before* the enumeration occurs. Our Hybrid MinDEE-$\boldsymbol{K}^*$ algorithm efficiently prunes the majority of the mutation sequences and conformations from more expensive evaluation, while still giving provable guarantees about the accuracy of its binding score predictions [eq. (A21) below]. Finally, the unpruned mutation sequences are ranked in order of their computed $\boldsymbol{K}^*$ scores.

Methods

Structural Model

Our structural model is the same as the one used in the original $\boldsymbol{K}^*$.[25] In our experiments, the structural model consists of nine active site residues (D235, A236, W239, T278, I299, A301, A322, I330, C331) of GrsA-PheA (PDB id: 1AMU),[37] a steric shell (30 residues with at least one atom within 8 Å from the substrate), the amino acid substrate, and the AMP cofactor. The steric shell facilitates the computation of the energy between the active site residues and neighboring regions of the protein (the residue-to-template energy) and constrains the movement of the active site residues to only sterically-allowable conformations relative to the body of the GrsA-PheA protein. All nine active site residues are modeled as flexible using rotamers and are subject to energy minimization. The steric shell includes residues 186Y, 188I, 190T, 210L, 213F, 214F, 230A, 234F, 237S, 238V, 240E, 243M, 279L, 300T, 302G, 303S, 320I, 321N, 323Y, 324G, 325P, 326T, 327E, 328T, 329T, 332A, 333T, 334T, 515N, and 517K. In 1AMU,[37] and also in ref. 25, residues 235D and 517K make H-bonds to the amino acid backbone of the ligand, thereby stabilizing the substrate in a productive orientation for catalysis. Flexible residues are represented by rotamers from the Richardsons' rotamer library.[9] The energy function consists of the AMBER electrostatic, vdW, and dihedral energy terms,[38,39] and the EEF1 pairwise solvation energy term.[40] A dielectric of 20 and a solvation energy scaling factor of 0.05 was used for the computational experiments. Each rotameric-based conformation is minimized using steepest-descent minimization (see Appendix C).

Energy Precomputation for Lower Bounds, B(·)

The MinDEE criterion [eq. (2)] uses both min and max *precomputed* energy terms to determine which rotamers are not part of the minimized-GMEC. There is no need to recompute the min and max energies every time eq. (2) is evaluated. See Appendix C for a detailed discussion.

Approximation Accuracy

We use an ε-value of 0.03, thus guaranteeing that the computed partial partition functions will be not less than 97% of the corresponding full partition functions. We use a value of 0.01 for γ (see Appendix B), which requires that correct $\boldsymbol{K}^*$ scores be computed for all mutation sequences whose score is at most two orders of magnitude less than the best score.

Filters

Volume Filter

Mutation sequences that are over- or under-packed by more than 30Å^3 compared to the wildtype PheA are pruned.

Steric Filter

Conformations in which a pair of atoms' vdW radii overlap by more than 1.5 Å prior to minimization are pruned.

Sequence-Space Filter

The active site residues are allowed to mutate to the set (GAVLIFYWM) of hydrophobic amino acids.

MinDEE

We use an implementation of the MinDEE analog[29] to the simple coupled Goldstein criterion.[33]

Results and Discussion

In this section, we compare the results of GMEC-based protein redesign without (traditional-DEE/$\boldsymbol{A}^*$) and with (MinDEE/$\boldsymbol{A}^*$) energy minimization. We also compare the redesign results when energy minimization is used without (MinDEE/$\boldsymbol{A}^*$) and with (Hybrid MinDEE-$\boldsymbol{K}^*$) conformational ensembles. We further compare our ensemble-based redesign results both to our previous computational predictions of protein designs and to biological activity assays of predicted protein mutants.

Comparison to Biological Activity Assays

Similarly to ref. 25, we simulated the biological activity assays of L-Phe and L-Leu against the wildtype PheA enzyme and the double mutant T278M/A301G.[17] In ref. 17, T278M/A301G was shown to have decreased specificity for Phe and increased specificity for Leu, as compared to the wildtype enzyme. The computed Hybrid MinDEE-$\boldsymbol{K}^*$ scores qualitatively agreed with these results: the Hybrid MinDEE-$\boldsymbol{K}^*$ score for wildtype with Phe was 17-fold higher than T278M/A301G with Phe; the Hybrid MinDEE-$\boldsymbol{K}^*$ score for wildtype with Leu was 12-fold lower than T278M/A301G with Leu.

Comparison to Traditional-DEE

For comparison, the simple coupled Goldstein traditional-DEE criterion[33] was used in a redesign search for changing the specificity of the wildtype PheA enzyme from Phe to Leu, using the experimental setup in section "Methods". A comparison to the rotamer assignments in the minimized-GMEC A236M/A322M MinDEE/A^* section revealed that A301, the minimized-GMEC identity at residue position 301, was in fact pruned by traditional-DEE. We then energy-minimized A236M/A301G, the rigid-GMEC obtained by traditional-DEE/$\boldsymbol{A}^*$ and determined that its energy was higher (by $\sim$ 6 kcal/mol) than the energy for the minimized-GMEC obtained by MinDEE/$\boldsymbol{A}^*$. Moreover, a total of 396 different conformations minimized to an energy lower than the minimized rigid-GMEC energy (see Fig. 6). These results confirm our claim that traditional-DEE is not provably-accurate with energy-minimization; they also show that conformations pruned by traditional-DEE may minimize to a lower energy state than the rigid-GMEC.

Redesign for Leu

Hybrid MinDEE-$\boldsymbol{K}^*$

The experimental setup for Leu redesign with Hybrid MinDEE-$\boldsymbol{K}^*$ is as described in section "Methods". The 2-point mutation search took $\sim$ 9 hr on a cluster of 24 processors. Only 30% of the mutation sequences passed the volume filter, while MinDEE pruned 98% of the remaining conformations. The use of the ε-approximation algorithms reduced the number of conformations that had to be subsequently generated and energy-minimized by an additional factor of 50 (see Table 1). A brute-force version of Hybrid MinDEE-$\boldsymbol{K}^*$ that did not utilize any of the filters, would take $\sim$ 8700 times longer ($\sim$ 3, 262 days) for the same experimental setup for redesign.

Table 1. Conformational Pruning with Hybrid MinDEE-$\boldsymbol{K}^*$.

	Conf. remaining	Pruning factor (%)
Initial	6.8×10^8	–
Volume Filter	2.04×10^8	3.33 (70.0)
MinDEE Filter	4.13×10^6	49.43 (98.0)
Steric Filter	3.86×10^6	1.07 (6.5)
$\boldsymbol{A}^*$ Energy Filter	7.82×10^4	49.41 (98.0)

The initial number of conformations for the GrsA-PheA 2-residue Leu mutation search is shown with the number of conformations remaining after the application of volume, MinDEE, steric, and energy (with $\boldsymbol{A}^*$) pruning. The $\boldsymbol{A}^*$ energy filter is based on the ε-approximation algorithms in section "Efficient partition function computation using A^* search" and Appendix B. The pruning factor represents the ratio of the number of conformations present before and after the given pruning stage. The pruning-% (in parentheses) represents the percentage of remaining conformations eliminated by the given pruning stage.

To determine the per-sequence pruning efficiency of Hybrid MinDEE-$\boldsymbol{K}^*$, we further computed the fraction of fully-evaluated conformations (the number of conformations that pass all of the Hybrid MinDEE-$\boldsymbol{K}^*$ filters, divided by the total number of conformations) separately for each sequence. Figure 5 shows the fraction of fully-evaluated conformations versus the computed log $\boldsymbol{K}^*$ scores for each of the unpruned sequences, for the protein-ligand bound-state partition function computation. As expected, the fraction of fully-evaluated conformations that contribute significantly to the computation of the provably-accurate ε-approximation to the partition function is very small (less than 0.5%) for all sequences, confirming again the efficiency of Hybrid MinDEE-$\boldsymbol{K}^*$. However, there is no correlation between the magnitude of the sequence scores and the fraction of fully-evaluated conformations.

The two top-scoring sequences are A301G/I330W and A301G/I330F for both Hybrid MinDEE-$\boldsymbol{K}^*$ and the original-$\boldsymbol{K}^*$. These novel mutation sequences were tested in the wetlab and were shown to have the desired switch of specificity from Phe to Leu (for details of the wetlab experiments, see ref. 25). Moreover, the other known successful redesign T278M/A301G[17] is ranked 3rd by Hybrid MinDEE-$\boldsymbol{K}^*$ (this sequence was ranked 12th by the original-$\boldsymbol{K}^*$ in ref. 25). Furthermore, all of the top 13 Hybrid MinDEE-$\boldsymbol{K}^*$ sequences contain the mutation A301G, which is found in all known native Leu adenylation domains.[18] These results show that our algorithms can give reasonable predictions for redesign.

*Comparison to Original-K**. An initial comparison to the original-$\boldsymbol{K}^*$ results showed only a small overlap between the top-ranking mutations for Hybrid MinDEE-$\boldsymbol{K}^*$ and the original-$\boldsymbol{K}^*$.[25] To facilitate a fair comparison between the two algorithms, we used

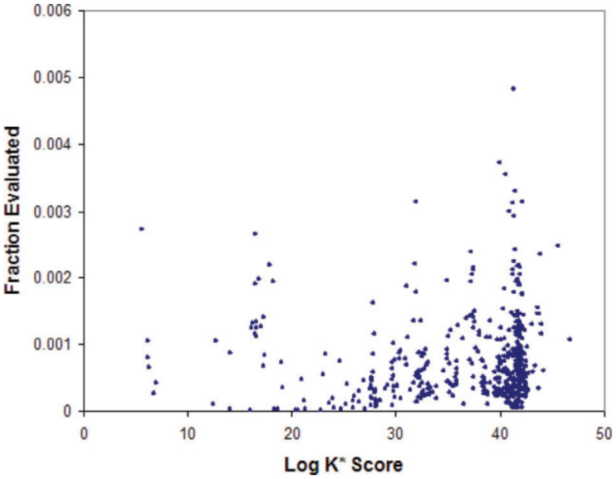


Figure 5. Fraction of fully-evaluated conformations for the Hybrid MinDEE-***K**** bound-state ensembles (GrsA-PheA active site redesign). For each of the unpruned mutation sequences, the log of the computed ***K**** score is shown vs. the fraction of fully-evaluated conformations used to compute an ε-approximation to the partition function for the bound protein-ligand complex. The fraction of fully-evaluated conformations for a given sequence is the ratio of the number of conformations that pass all of the Hybrid MinDEE-***K**** pruning filters (see Table 1) divided by the total number of conformations for that sequence. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

the same energy function (as described in Methods section, but without solvation energies) and energy-minimization module (see Appendix C) for both Hybrid MinDEE-***K**** and the original-***K****. This comparison revealed that both the mutation-sequence rankings and the scores for a given mutation sequence are very similar for the two algorithms: the top 19 sequences are identical, while all of the top 40 sequences for Hybrid MinDEE-***K**** can be found in the top 40 sequences for ***K****, and vice versa; the trend is similar for the remaining sequences, as well. This fact shows that, all other factors being equal, both algorithms converge to very similar results, despite the different (but still provably-accurate) filters used. To compare the efficiency of the two algorithms, we measured the number of fully-evaluated conformations, since the full energy minimization of the conformations is the most computationally-expensive part of both algorithms. The original-***K**** algorithm fully-evaluated ∼ 30% more conformations than Hybrid MinDEE-***K****. Thus, Hybrid MinDEE-***K**** is much more efficient at obtaining the desired results.

*MinDEE/**A****

We now discuss results from our GMEC-based experiments using MinDEE/***A****. To redesign the wildtype PheA enzyme so that its substrate specificity is switched towards Leu, we used the experimental setup described in section "Methods". The MinDEE filter on the bound protein:ligand complex pruned 206 out of the 421 possible rotamers for the active site residues, reducing the number of conformations that were subsequently supplied to ***A**** by a factor of 2,330. We then extracted and minimized all conformations over the 2-point mutation sequences using the ***A**** search until the halting condition defined in Proposition 2 was reached, for $E_w = 12.5$ kcal/mol. A total of 7261 conformations, representing 221 unique mutation sequences, had actual minimized energies within 12.5 kcal/mol of the minimized-GMEC energy (see Fig. 6), which confirms that a mutation sequence can be found in multiple low-energy states. The top-ranked MinDEE/***A**** mutation sequence is A236M/A322M; the minimized-GMEC is obtained from this sequence. The entire redesign process took ∼ 4 days on a single processor (the MinDEE pruning stage took less than a minute, and the remainder of the time was spent in the ***A**** enumeration stage), with more than 60, 000 extracted conformations before the search could be provably halted. Thus, the provable accuracy of the results comes at the cost of this computational overhead, since the number of extracted conformations is much larger than the actual number of conformations within E_w of the minimized-GMEC energy. Note, however, that a redesign effort without a MinDEE filter and a provably-accurate halting condition would be computationally infeasible.

Since a mutation sequence can be found in multiple low-energy states (see above), it is interesting to determine how similar these states are. We therefore selected the set of conformations generated by MinDEE/***A**** for the minimized-GMEC sequence A236M/A322M for further analysis. For this sequence, Figure 7 shows the all-atom RMSD (active site residues only) for the minimized-GMEC with each of the 337 conformations within 12.5 kcal/mol of the minimized-GMEC energy. As Figure 7 shows, the similarity of the structures varies significantly, with 75% of the structures clustered within the range 0.6–1.1 RMSD (average of 0.83). Although the correlation between the RMSD values and the conformational energies is weak (R^2 of 0.24), there is a general trend for

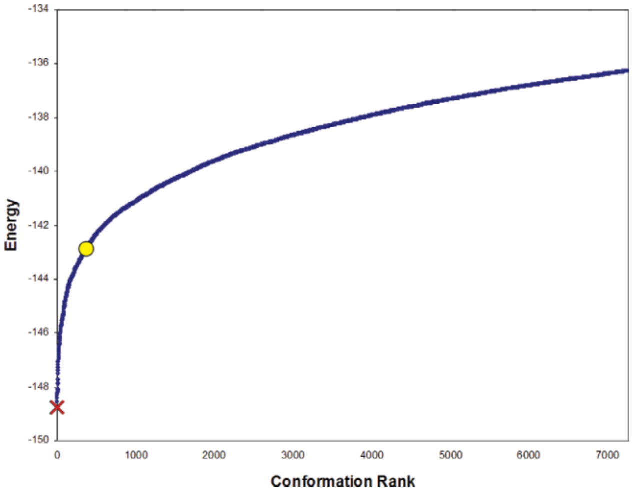


Figure 6. Energies of all conformations within 12.5 kcal/mol of the minimized–GMEC energy. The energies (after minimization) of the minimized–GMEC (red cross) and the rigid–GMEC (yellow circle) are shown. The minimized–GMEC A236M/A322M is (by definition) the lowest–energy conformation, while the rigid–GMEC is ranked 397th. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

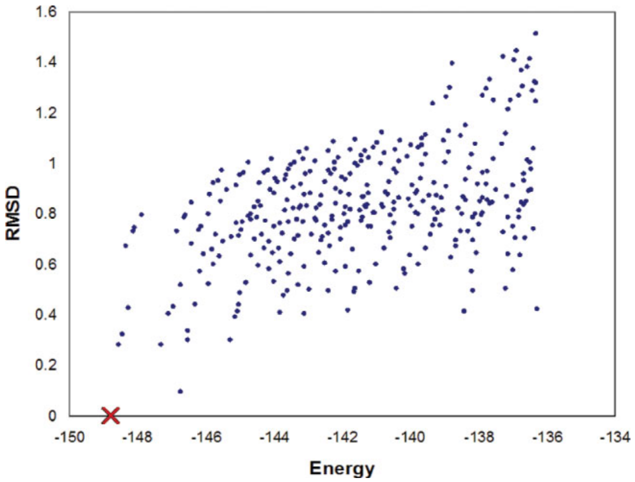


Figure 7. All-atom RMSD (active site residues only) versus energy for all A236M/A322M conformations generated by MinDEE/$\boldsymbol{A}^*$. A total of 337 conformations for the A236M/A322M sequence have energies within 12.5 kcal/mol of the MinDEE/$\boldsymbol{A}^*$ minimized-GMEC. The all-atom RMSD with the minimized-GMEC (red cross) for each of these conformations is shown versus the corresponding computed conformational energy.

conformations with a larger deviation from the minimized-GMEC structure to also have higher energies.

As another measure of similarity between the low-energy conformations for the A236M/A322M sequence, we computed the frequency for each observed rotamer identity at each active site residue position (Fig. 8). As Figure 8 shows, with the exception of T278 and C331 which assume all allowed rotamers for the corresponding amino acid types from the Richardsons' rotamer library, all other residues preferentially assume only a small subset of the possible rotamers (cf. ref. 9), thus indicating some (though not high) rotamer diversity between the different structures. This rotamer diversity, in combination with the rotameric energy minimization allowed in our model, are the reasons for the structure variability observed in Figure 7.

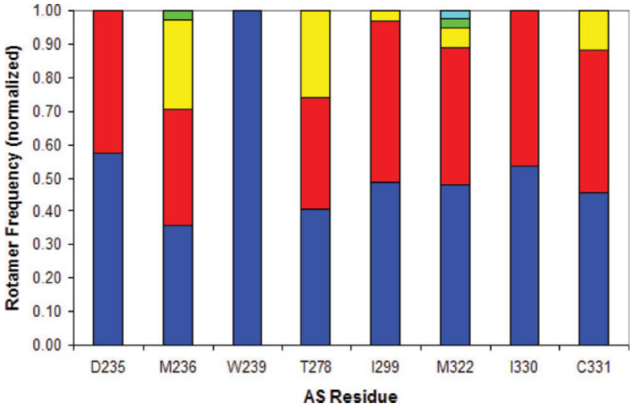


Figure 8. Rotamer diversity for the A236M/A322M conformations generated by MinDEE/$\boldsymbol{A}^*$. For each active site residue, the normalized frequency for each observed rotamer (number of occurrences divided by the total number of structures) is shown: the highest-occurring rotamer is in blue, the second-highest is in red, followed by yellow, green, and light blue. For clarity, A301 is not shown here since Ala has only one rotamer.

Only 2 of the top 40 MinDEE/$\boldsymbol{A}^*$ mutation sequences can be found in the top 40 Hybrid MinDEE-$\boldsymbol{K}^*$ sequences, and vice versa,

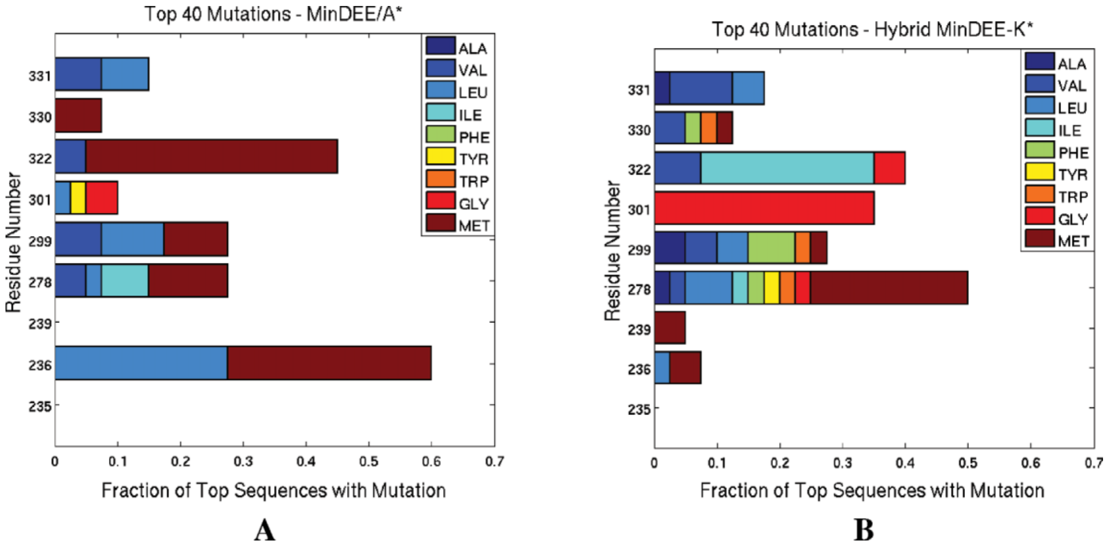


Figure 9. Distribution of mutations. The distribution of the mutation types for the top 40 mutation sequences for (A) MinDEE/$\boldsymbol{A}^*$ and (B) Hybrid MinDEE-$\boldsymbol{K}^*$ algorithms is shown as the fraction of each mutating type for each active site residue. The types and frequencies for the mutations are quite different for the two methods, which indicates that the difference in the information content for GMEC- and ensemble-based algorithms can be substantial.

indicating that ensemble-scoring yields substantially different predictions from single-structure scoring using the minimized-GMEC, where only the minimized bound state of a single conformation is considered (see Fig. 9).

Limitations and Extensions

The MinDEE criterion can efficiently prune a large number of the possible conformations (see section "Redesign for Leu"). However, because of the use of min and max energy terms, the pruning efficiency of MinDEE cannot be as high as that of traditional-DEE. This trade-off in efficiency results from the provable guarantees that MinDEE can (but traditional-DEE cannot) make when energy minimization is employed. An increase of the pruning capabilities of MinDEE would require the derivation and computation of tighter upper and lower energy bounds. Since (with a rigid backbone) the conformational changes due to switching the identity of a single rotamer should decrease in magnitude as the proximity to the modified rotamer decreases, it may also be possible to increase the pruning factor by scaling the terms in the MinDEE condition [eq. (2)], depending on the proximity of the residues involved.

The goal of our ensemble-based Hybrid MinDEE-$\boldsymbol{K}^*$ algorithm is to find mutation sequences with better binding constants for the novel substrate than the wildtype enzyme. An assessment of catalytic activity is not explicitly included in the algorithm. In general, it would be interesting to generalize $\boldsymbol{K}^*$ to stabilize the transition state. Since the transition state is not known structurally, $\boldsymbol{K}^*$ maintains backbone contacts of the substrate in proximity to the nucleotide cofactor. As was shown in ref. 26, the top $\boldsymbol{K}^*$-predicted mutations in a GrsA-PheA redesign improved the catalytic specificity ($k_{\text{cat}}/K_{\text{M}}$) as well.

Several limitations of our computational model warrant a discussion. Since using a continuous representation for the partition functions is currently not feasible, our algorithm discretizes the conformational space. Rotamer discretization has been shown to work well in practice.[6,7,14,41–43] A further limitation of our model is the use of a rigid backbone. However, our algorithm aims to simultaneously find the best mutations and to stabilize the sidechain placements for the given backbone, rather than assuming the backbone will remain rigid. All dead-end elimination algorithms, and the majority of structure-based protein design algorithms in general, use a model with a rigid backbone. The incorporation of backbone flexibility, however, will likely improve the computational predictions, and thus represents interesting future work.[44]

Conclusions

When energy-minimization is required, the traditional-DEE criterion makes no guarantees about pruning rotamers belonging to the minimized-GMEC. In contrast, a rotamer is only pruned by MinDEE if it is provably not part of the minimized-GMEC. We showed experimentally that the minimized-GMEC can minimize to lower energy states than the rigid-GMEC, confirming the feasibility and significance of our novel MinDEE criterion. When used as a filter in ensemble-based redesign, MinDEE efficiently reduced the conformational and sequence search spaces, leading both to predictions consistent with previous redesign efforts and novel sequences that are unknown in nature. Our Hybrid MinDEE-$\boldsymbol{K}^*$ algorithm showed a significant improvement in pruning efficiency, as compared to the original $\boldsymbol{K}^*$ algorithm. Redesign searches for two other substrates, Val and Tyr, have also been performed, confirming the generality of our algorithms.

Protein design using traditional-DEE uses neither ensembles nor rotamer minimization. In our experiments, we reported the relative benefits of incorporating ensembles and energy-minimization into a provable redesign algorithm. A major challenge for protein redesign algorithms is the balance between the efficiency and accuracy with which redesign is performed. While the ability to prune the majority of mutation/conformation search space is extremely important, increasing the accuracy of the model is a prerequisite for successful redesign. It would be interesting to implement finer rotamer sampling and more accurate (and hence more expensive) energy functions, and remove bias in the rotamer library by factoring the Jacobian into the partition function over torsion-angle space. MinDEE can also be generalized to incorporate backbone flexibility.[44] An accurate and efficient algorithm for redesigning the enzymes that synthesize natural products should prove useful as a technique for drug design.

Acknowledgments

The authors thank Prof. A. Anderson, Dr. S. Apaydin, Mr. J. MacMaster, Mr. A. Yan, Mr. B. Stevens, and all members of the Donald Lab for helpful discussions and comments on drafts.

Appendix

In Appendix A, we present a detailed proof of Proposition 1 from section "DEE with energy minimization: MinDEE". Appendix B presents an improvement to the algorithm of section "Efficient partition function computation using A^* search" for more efficient partition function computation. Appendix C provides details on the energy precomputation for computing the lower energy bounds $B(\cdot)$.

Appendix A: MinDEE Derivation

In this section, we present a detailed proof of Proposition 1. For clarity, we restate Proposition 1 here:

Proposition 1. *When Eq. (2) holds, rotamer i_r is provably not part of the minimized-GMEC.*

Proof. For a given protein, we define a *rotamer vector* $A = (A_1, A_2, \cdots, A_n)$ to specify the rotamer at each of the n residue positions; $A_i = r$ when rotamer r is assumed by residue i. We then define the *conformation vector* $A^\bullet = (A_1^\bullet, A_2^\bullet, \cdots, A_n^\bullet)$ such that $A_i^\bullet$ is the conformation of residue i in the voxel-constrained minimized conformation, i.e., $A_i^\bullet \in \mathcal{V}(A_i)$ and

$$A^\bullet = \left(A_1^\bullet, A_2^\bullet, \cdots, A_n^\bullet\right) = \underset{B=(B_1,B_2,\cdots,B_n)\in\prod_{i=1}^n \mathcal{V}(A_i)}{\operatorname{argmin}} E(B) \qquad \text{(A1)}$$

where $E(B)$ is the energy of the system specified by conformation vector B. For the energy-minimized conformation starting from rotamer vector A, we define the self-energy of rotamer i_r as $E_{\odot}(i_r|A) = E(A_i^{\bullet})$ and the pairwise interaction energy of the rotamer pair i_r, j_s as $E_{\odot}(i_r, j_s|A) = E(A_i^{\bullet}, A_j^{\bullet})$ where $E(A_i^{\bullet})$ is the self-energy of residue i in conformation $A_i^{\bullet}$ and $E(A_i^{\bullet}, A_j^{\bullet})$ is the pairwise energy between residues i and j in conformations $A_i^{\bullet}$ and $A_j^{\bullet}$. We can then express the minimized energy of A, $E_T(A)$ as:

$$E_T(A) = E_{t'} + \sum_i E_{\odot}(i_r|A) + \sum_i \sum_{j>i} E_{\odot}(i_r, j_s|A). \tag{A2}$$

Let G represent the rotamer vector that minimizes into the minimized-GMEC and $E_T(G)$ be the energy of the minimized-GMEC. Let $G_{i_g \rightarrow i_t}$ be the rotamer vector G where rotamer i_g is replaced with i_t. We know that $E_T(G_{i_g \rightarrow i_t}) \geq E_T(G)$, so we can pull residue i out of the two summations, obtaining:

$$\begin{aligned} E_{t'} + E_{\odot}(i_t|G_{i_g\rightarrow i_t}) + \sum_{j\neq i} E_{\odot}(i_t, j_g|G_{i_g\rightarrow i_t}) + \sum_{j\neq i} E_{\odot}(j_g|G_{i_g\rightarrow i_t}) \\ + \sum_{j\neq i} \sum_{k\neq i, k>j} E_{\odot}(j_g, k_g|G_{i_g\rightarrow i_t}) \geq E_{t'} + E_{\odot}(i_g|G) \\ + \sum_{j\neq i} E_{\odot}(i_g, j_g|G) + \sum_{j\neq i} E_{\odot}(j_g|G) + \sum_{j\neq i} \sum_{k\neq i, k>j} E_{\odot}(j_g, k_g|G). \end{aligned} \tag{A3}$$

The $E_{t'}$ terms (section "The original DEE criterion") correspond to the rigid portion of the molecule; they are independent of rotamer choice, are equal, and can be canceled. We make the following trivial upper and lower-bound observations (the $E_{\ominus}(\cdot)$, $E_{\oplus}(\cdot)$, and $E_{\oslash}(\cdot)$ terms are as defined in section "DEE with energy minimization: MinDEE"):

$$E_{\odot}(i_t|A) \leq E_{\oplus}(i_t); E_{\odot}(i_t, j_g|A) \leq \max_{s\in R_j} E_{\oplus}(i_t, j_s); \tag{A4}$$

$$E_{\odot}(j_g|A) \leq E_{\oplus}(j_g); E_{\odot}(j_g, k_g|A) \leq E_{\oplus}(j_g, k_g); \tag{A5}$$

$$E_{\ominus}(i_g) \leq E_{\odot}(i_g|A); \min_{s\in R_j} E_{\ominus}(i_g, j_s) \leq E_{\odot}(i_g, j_g|A); \tag{A6}$$

$$E_{\ominus}(j_g) \leq E_{\odot}(j_g|A); E_{\ominus}(j_g, k_g) \leq E_{\odot}(j_g, k_g|A). \tag{A7}$$

Substituting eqs. (A4–A7) into eq. (A3), we obtain:

$$\begin{aligned} E_{\oplus}(i_t) + \sum_{j\neq i} \max_s E_{\oplus}(i_t, j_s) + \sum_{j\neq i} E_{\oplus}(j_g) + \sum_{j\neq i} \sum_{k\neq i, k>j} E_{\oplus}(j_g, k_g) \geq \\ E_{\ominus}(i_g) + \sum_{j\neq i} \min_s E_{\ominus}(i_g, j_s) + \sum_{j\neq i} E_{\ominus}(j_g) + \sum_{j\neq i} \sum_{k\neq i, k>j} E_{\ominus}(j_g, k_g). \end{aligned} \tag{A8}$$

When the MinDEE pruning condition eq. (2) holds, we can substitute the left-hand side of eq. (2) for the first two terms of eq. (A8), and simplify the resulting equation to:

$$\begin{aligned} E_{\ominus}(i_r) + \sum_{j\neq i} \min_s E_{\ominus}(i_r, j_s) - \sum_{j\neq i} \max_s E_{\oslash}(j_s) \\ - \sum_{j\neq i} \sum_{k\neq i, k>j} \max_{s,u} E_{\oslash}(j_s, k_u) + \sum_{j\neq i} E_{\oslash}(j_g) + \sum_{j\neq i} \sum_{k\neq i, k>j} E_{\oslash}(j_g, k_g) \\ > E_{\ominus}(i_g) + \sum_{j\neq i} \min_s E_{\ominus}(i_g, j_s). \end{aligned} \tag{A9}$$

We then substitute the following two bounds $\sum_{j\neq i} \max_s E_{\oslash}(j_s) \geq \sum_{j\neq i} E_{\oslash}(j_g)$ and $\sum_{j\neq i} \sum_{k\neq i, k>j} \max_{s,u} E_{\oslash}(j_s, k_u) \geq \sum_{j\neq i} \sum_{k\neq i, k>j} E_{\oslash}(j_g, k_g)$ into eq. (A9) and reduce:

$$E_{\ominus}(i_r) + \sum_{j\neq i} \min_s E_{\ominus}(i_r, j_s) > E_{\ominus}(i_g) + \sum_{j\neq i} \min_s E_{\ominus}(i_g, j_s). \tag{A10}$$

Thus, when the MinDEE pruning condition eq. (2) holds, $i_r \neq i_g$ and we can provably eliminate rotamer i_r as not being part of the energy-minimized GMEC. ■

Appendix B: Improved Partition Function Computation

We now describe an improvement to the algorithm of section "Efficient partition function computation using A^* search" for more efficient partition function computation. In section "Efficient partition function computation using A^* search", provably-accurate $\boldsymbol{K}^*$ scores are computed for all mutation sequences. However, since we are only interested in mutation sequences with high $\boldsymbol{K}^*$ scores (i.e., sequences that are good binders), we need only require that a provably-accurate score be computed for the top fraction of the mutation sequences. To achieve this, we will allow conformational pruning *across* mutation sequences. Hence, for clarity, we will refer to the partition function computation described in this Appendix as *inter-mutation*, while the computation described in section "Efficient partition function computation using A^* search" (where conformational pruning could be performed *only within* a sequence) will be referred to as *intra-mutation*. Below, we use the following idea (cf ref. 25). When using $\boldsymbol{K}^*$ to perform a mutation search, we can bootstrap the pruning condition for improved efficiency (by caching partition functions, we can exploit $\boldsymbol{K}^*$ bounds from other mutations in the same search). Our search algorithm has the desirable property that provably-accurate ε-approximations are computed for top-ranking mutations, while the bounds we can prove on the quickly-computed $\boldsymbol{K}^*$ values for lower-ranked mutations do not enjoy the same degree of accuracy. This idea is briefly formulated and then exploited below.

We first review some of the definitions from ref. 25. We let $\gamma \in [0, 1]$ be a parameter that defines the set of mutation sequences for which an ε-approximation is to be computed. We require that an ε-approximation be guaranteed for a mutation sequence i only when $K_i^* \geq \gamma K_o^*$, where K_i^* is the score for sequence i and K_o^* is the best score observed so far in the search. When $\gamma = 1.0$, an ε-approximation is guaranteed only for the best-scoring $\boldsymbol{K}^*$ mutation sequence; $\gamma = 0.0$ computes an ε-approximation for all $\boldsymbol{K}^*$ mutation sequences. Let us assume that $\boldsymbol{A}^*$ has already generated the first m conformations and that there are n remaining conformations that have not been generated yet. We use the definitions for q', p^*, E_0, and k from Proposition 3 above. We assume that we have already computed q_{P} using the intra-mutation filter only (Proposition 3), and now describe how to efficiently compute q_{PL}.

We define the score for the ith mutation sequence to be $K_i^* = \frac{q_{\mathrm{PL}}}{q_{\mathrm{P}} q_{\mathrm{L}}}$, while $K_o^* = \frac{{}^o q_{\mathrm{PL}}}{{}^o q_{\mathrm{P}} {}^o q_{\mathrm{L}}}$. We let q^*_{PL} be the partial partition function for the bound protein-ligand state, computed from the m already-generated conformations. We define $K_o^\dagger = \frac{{}^o q_{\mathrm{PL}}}{{}^o q_{\mathrm{P}}}$. Finally, let $\psi = \max\,(\gamma \varepsilon K_o^\dagger q_{\mathrm{P}}, q^*_{\mathrm{PL}} \rho)$ and $\rho = \frac{\varepsilon}{1-\varepsilon}$.

Proposition 4. *If the lower bound $B(c)$ on the minimized energy of the $(m+1)^{\mathrm{st}}$ conformation returned by $\boldsymbol{A}^*$ satisfies $B(c) \geq -RT(\ln(\psi - k\exp(-E_0/RT)) - \ln n)$, then the partition function computation can be halted, with q^*_{PL} guaranteed to be an ε-approximation to the true partition function q_{PL} for a mutation sequence whose score K_i^* satisfies $K_i^* \geq \gamma K_o^*$.*

Proof. Since the ligand is invariant throughout the search, $q_{\mathrm{L}} = {}^o q_{\mathrm{L}}$. Let us assume that we have a sequence for which $K_i^* \geq \gamma K_o^*$ holds. Thus,

$$\begin{aligned} \frac{q_{\mathrm{PL}}}{q_{\mathrm{P}} q_{\mathrm{L}}} &\geq \gamma \frac{{}^o q_{\mathrm{PL}}}{{}^o q_{\mathrm{P}} {}^o q_{\mathrm{L}}}, \\ q_{\mathrm{PL}} &\geq \gamma K_o^\dagger q_{\mathrm{P}}. \end{aligned} \tag{A11}$$

First, we note again that

$$q' \leq n\exp(-E_t/RT); \tag{A12}$$

$$p^* \leq k\exp(-E_0/RT). \tag{A13}$$

From the definition of q_{PL}, we obtain

$$q_{\mathrm{PL}} = q^*_{\mathrm{PL}} + q' + p^*. \tag{A14}$$

Now, if

$$n\exp(-E_t/RT) + k\exp(-E_0/RT) \leq \varepsilon K_o^\dagger \gamma q_{\mathrm{P}}, \tag{A15}$$

then by eqs. (A12) and (A13) we have

$$q' + p^* \leq \varepsilon K_o^\dagger \gamma q_{\mathrm{P}}, \tag{A16}$$

and by eq. (A11),

$$q' + p^* \leq \varepsilon q_{\mathrm{PL}}, \tag{A17}$$

and finally, by eq. (A14), we obtain

$$q^*_{\mathrm{PL}} \geq (1-\varepsilon) q_{\mathrm{PL}}, \tag{A18}$$

which is the definition of the partition function ε-approximation. Thus, if eq. (A15) holds, then we will have an ε-approximation to the true partition function q_{PL}. Solving eq. (A15) for E_t, we obtain the stop-threshold:

$$E_t \geq -RT\big(\ln\big(\gamma \varepsilon K_o^\dagger q_{\mathrm{P}} - k\exp(-E_0/RT)\big) - \ln n\big). \tag{A19}$$

The first conformation that has an energy above the stop-threshold (eq. A19) halts the partition function computation, since we already have an ε-approximation. Thus, combining eq. (A19) and the intra-mutation stop-threshold (eq. 9), our stopping condition for the computation of q_{PL} becomes

$$B(c) > -RT(\ln(\psi - k\exp(-E_0/RT)) - \ln n), \tag{A20}$$

where $\psi = \max(\gamma \varepsilon K_o^\dagger q_{\mathrm{P}}, q^*_{\mathrm{PL}} \rho)$ and $B(\cdot)$ is the lower bound on the minimized energy of a conformation. ■

If the desired approximation accuracy is not achieved at the end of the mutation search, after all conformations have been extracted by $\boldsymbol{A}^*$, we can modify eq. (12) to incorporate the inter-mutation filter, obtaining the number of conformations l from P (the set of pruned conformations) that must be allowed in the repeated search:

$$l \geq k - \frac{\psi}{\exp(-E_0/RT)}.$$

We have derived the stop-threshold that guarantees an ε-approximation to the partition function when conformations are generated in order of increasing lower bounds on the conformation's energy. This generalizes the inter-mutation proof in ref. 25 which is valid when the energy lower bounds for all of the conformations are evaluated. We should note that eq. (A20) was derived assuming $K_i^* \geq \gamma K_o^*$ holds, so we can guarantee an ε-approximation to q_{PL} only for this case. When $K_i^* < \gamma K_o^*$, then we might not obtain an ε-approximation for the given mutation sequence, but we do not require a provably-good approximation for such low-scoring sequences.

Similarly to ref. 25, we define $\tilde{K_i^*} = \frac{q^*_{\mathrm{PL}}}{q^*_{\mathrm{P}} q_{\mathrm{L}}}$ to be an ε-approximation to the full score of a mutation sequence (the score if the full partition functions are used, instead of the partial ones) when $\tilde{K_i^*} \in [K_i^*(1-\varepsilon), \frac{1}{1-\varepsilon} K_i^*]$. If $K_i^* \geq \gamma K_o^*$ holds for a mutation sequence i, then by Proposition 4, $q_{\mathrm{PL}} > q^*_{\mathrm{PL}} \geq (1-\varepsilon) q_{\mathrm{PL}}$. Also, since q_{P} is already computed using Proposition 3, $q_{\mathrm{P}} > q_P^* \geq (1-\varepsilon) q_{\mathrm{P}}$. Since $K_i^* = \frac{q_{\mathrm{PL}}}{q_{\mathrm{P}} q_{\mathrm{L}}}$, we have

$$\left[K_i^*(1-\varepsilon) \leq \tilde{K_i^*} \leq \frac{1}{1-\varepsilon} K_i^*\right]. \tag{A21}$$

Thus, the algorithm guarantees that an ε-approximation to the full *score* is computed when $K_i^* \geq \gamma K_o^*$.

Appendix C: Energy Precomputation for Lower Bounds

We first derive a lower bound for the energy of a minimized conformation, closely following ref. 25. We then present improvements on the energy precomputation algorithm, as compared to ref. 25.

Computing a Lower Bound on Minimized Energies

In our structural model, (section "Methods"), some residues are treated as rigid, while others have a rigid backbone but flexible side-chains. Let h be the number of flexible residues in our system. Let A be a $(h + 1) \times (h + 1)$ precomputed residue-indexed energy matrix that describes the energy interactions of a given residue i within itself (A_{i0}), with the backbone (A_{0i}), and with other residues (A_{ij}); the matrix element A_{00} is reserved for the energy interactions between the atoms of the backbone only. We term A_{00} to be the template energy, A_{0i} is the residue-to-template energy, A_{i0} is the intra-residue energy, and A_{ij} is the pairwise energy for residue i. The energy of the system can be computed as

$$E_S = A_{00} + \sum_{i \le h} A_{0i} + \sum_{i \le h} A_{i0} + \sum_{i \le h} \sum_{i<j \le h} A_{ij}. \tag{A22}$$

To compute the energy of a minimized conformation, we use a matrix M, whose elements are analogous to the elements of A, but the precomputed energies correspond to the energy-minimized structure. If we obtain the lower bounds on the energy terms in M and store these bounds in a matrix D, then we can define the lower bound $E_{\min}$ on the energy of a minimized system as

$$E_{\min} = D_{00} + \sum_{i \le h} D_{0i} + \sum_{i \le h} D_{i0} + \sum_{i \le h} \sum_{i<j \le h} D_{ij}. \tag{A23}$$

The computation of $E_{\min}$ can be done in time $O(h^2)$ with a precomputed pairwise energy matrix. The use of a precomputed residue-indexed lower-bound pairwise energy matrix avoids the computation of $O(a^2)$ energy terms, where $a \gg h$ is the total number of atoms in the system.

The precomputed energy matrix in the original $\boldsymbol{K}^*$ is indexed over all residues and over all rotamers for each reside. Thus, for a system with h flexible residues and m rotamers for each residue, we precompute a $(hm + 1) \times (hm + 1)$ residue-indexed lower-bound pairwise energy matrix V whose elements V_{00}, V_{0i}, V_{i0}, and V_{ij} are analogous to the elements of D. To compute the lower bounds on the minimized template, intra-residue, residue-to-template, and pairwise energy terms, we allow rotamers to assume the best possible conformation for the given relative system (template, self-, or pairwise). However, the movement of the rotamer dihedrals is constrained to a hypercuboid region of conformation space, called a voxel,[34,35] so that one rotamer will not minimize into another. We use a voxel of ±9° for each χ angle.

Application of the Pairwise Energy Matrix

Energy precomputation is employed both for pruning with MinDEE (see section "DEE with energy minimization: MinDEE") and for the ε-approximation algorithms (see section "Efficient partition function computation using A^* search" and Appendix B). The MinDEE criterion (eq. 2) uses both the lower- and the upper-bound (see Appendix section "Improved energy bounds computation") precomputed energy terms to determine which rotamers are not part of the energy-minimized GMEC. Thus, there is no need to re-compute the minimum and maximum energies every time eq. (2) is evaluated.

Both the intra- and inter-mutation filters (Propositions 3 and 4, respectively) require that a lower bound on the energy-minimized conformation be computed. For this purpose, a lookup in the lower-bound pairwise energy matrix is performed and the terms involved in the given conformation are added, analogously to eq. (A23). The computation of a lower bound on the energy of a conformation permits a subset of the conformations to be pruned before the computationally-expensive full energy-minimization stage. The full energy minimization of a given system requires the simultaneous minimization of all of the flexible residues for the system, a much more costly process than the pairwise minimization performed for the precomputations. Moreover, once the pairwise matrices are precomputed, they can be used in any mutation search that involves the same residues. Thus, in a protein-ligand system, a redesign for a different ligand requires the re-computation only of the terms involving the ligand.

Improved Energy Bounds Computation

Analogously to the definition of matrix D in Appendix section "computing a lower bound on minimized energies", we define the matrix F to be the residue-indexed upper-bound pairwise energy matrix, which facilitates the computation of the upper-bound $E_{\max}$ on the maximized energy of a system:

$$E_{\max} = F_{00} + \sum_{i \le h} F_{0i} + \sum_{i \le h} F_{i0} + \sum_{i \le h} \sum_{i<j \le h} F_{ij}. \tag{A24}$$

Analogously to the definition of V (see Appendix section "computing a lower bound on minimized energies"), when we index over all rotamers for all residues, we can define the $(hm + 1) \times (hm + 1)$ residue-indexed upper-bound pairwise energy matrix U, whose elements U_{00}, U_{0i}, U_{i0}, and U_{ij} are upper-bounds on the corresponding energy terms.

The original $\boldsymbol{K}^*$ algorithm[25] used a steepest-descent minimization scheme to precompute lower-bound energy matrices. To improve the minimization results, we (1) refined the implementation of the steepest-descent algorithm, and (2) implemented a random sampling with steepest descent algorithm that explores the energy landscape within a voxel better than the local steepest-descent algorithm. Empirically, however, the computed minimum energy bounds using multiple random-sampling starting points appear to be over-optimistic and present a worse approximation to the actual conformation energies. The resulting lower bounds l_m from multiple minimization starting points are necessarily at least as low as the corresponding lower bounds l_s computed by minimizing only from the center of the voxels, $l_m \le l_s$. Choosing a good starting point for the energy minimization of a full conformation that could use the additional information of the pairwise l_m bounds is a difficult task, since the different addends involved in the computation of l_m [analogous to Eq. (A23)] may actually result from incompatible starting points. Moreover, using multiple starting points for *full* energy-minimization is computationally infeasible (see Appendix section "Application of the pairwise energy matrix"). Thus, using multiple minimization starting points for lower-bounds computation in fact increases the gap between lower bounds and actual energies (i.e., the lower bounds are less achievable). As a result, the ε-approximation algorithms (see section "Efficient partition function computation using A^* search" and Appendix B) require the full minimization of a larger number of conformations before the provable halting conditions (Propositions 3 and 4) are reached. Hence, we chose to compute the pairwise minimum energy bounds using steepest-descent minimization starting at the center of the voxel space.

While min energies may appear as a natural concept, the computation of max energies (pairwise-computed maximum energy bounds) presents both conceptual and practical challenges. A simple maximization algorithm cannot be used, since most rotamer systems will maximize into a steric clash, which would make max bounds biophysically inapplicable. Moreover, energy functions, such as AMBER,[38,39] are not well-defined for high energies. However, max bounds are used only in the MinDEE framework, where, indirectly, minimized conformations are compared to determine which ones are provably not the minimized-GMEC. We can thus think of the max energy for a given rotamer system as the worst minimization this system can achieve. Hence, we chose to compute max energies as $\max(M)$, where M is the set of energies obtained by steepest-descent minimization from multiple starting points (max of mins). In all our experiments we used 200 randomly-chosen starting points per voxel.

References

1. Street, A.; Mayo, S. Structure 1999, 7, R105.
2. Jin, W.; Kambara, O.; Sasakawa, H.; Tamura, A.; Takada, S. Structure 2003, 11, 581.
3. Jaramillo, A.; Wernisch, L.; Héry, S.; Wodak, S. Comb Chem High Throughput Screen 2001, 4, 643.
4. Hellinga, H.; Richards, F. J Mol Biol 1991, 222, 763.
5. Bolon, D.; Mayo, S. PNAS USA 2001, 98, 14274.
6. Looger, L.; Dwyer, M.; Smith, J.; Hellinga, H. Nature 2003, 423, 185.
7. Keating, A.; Malashkevich, V.; Tidor, B.; Kim, P. Proc Natl Acad Sci USA 2001, 98, 14825.
8. Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M. Proteins 2000, 39, 261.
9. Lovell, S.; Word, J.; Richardson, J.; Richardson, D. Proteins 2000, 40, 389.
10. Ponder, J.; Richards, F. J Mol Biol 1987, 193, 775.
11. Pierce, N.; Winfree, E. Protein design is *NP*-hard. Prot Eng 2002, 15, 779.
12. Chazelle, B.; Kingsford, C.; Singh, M. INFORMS J Comput Comput Biol Special Issue 2004, 16, 380.
13. Marvin, J.; Hellinga, H. PNAS 2001, 98, 4955.
14. Desmet, J.; Spriet, J.; Lasters, I. Proteins 2002, 48, 31.
15. Shah, P.; Hom, G.; Mayo, S. J Comput Chem 2004, 25, 1797.
16. Desmet, J.; De Maeyer, M.; Hazes, B.; Lasters, I. Nature 1992, 356, 539.
17. Stachelhaus, T.; Mootz, H.; Marahiel, M. Chem Biol 1999, 6, 493.
18. Challis, G.; Ravel, J.; Townsend, C. Chem Biol 2000, 7, 211.
19. Schwarzer, D.; Finking, R.; Marahiel, M. Nat Prod Rep 2003, 20, 275.
20. Stachelhaus, T.; Schneider, A.; Marahiel, M. Science 1995, 269, 69.
21. Schneider, A.; Stachelhaus, T.; Marahiel, M. Mol Gen Genet 1998, 257, 308.
22. Doekel, S.; Marahiel, M. Chem Biol 2000, 7, 373.
23. Mootz, H.; Schwarzer, D.; Marahiel, M. Proc Natl Acad Sci USA 2000, 97, 5848.
24. Eppelmann, K.; Stachelhaus, T.; Marahiel, M. Biochemistry 2002, 41, 9718.
25. Lilien, R.; Stevens, B.; Anderson, A.; Donald, B. R. J Comput Biol 2005, 12, 740.
26. Stevens, B.; Lilien, R.; Georgiev, I.; Donald, B. R.; Anderson, A. Biochemistry 2006, 45, 15495.
27. Cane, D.; Walsh, C.; Khosla, C. Science 1998, 282, 63.
28. Georgiev, I.; Lilien, R.; Donald, B. R. (2006b). A novel minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *International Conference on Research in Computational Molecular Biology (RECOMB)*, Venice, Italy, 2006.
29. Georgiev, I.; Lilien, R.; Donald, B. R. (2006a). Improved pruning algorithms and divide-and-conquer strategies for dead-end elimination, with application to protein design. Bioinformatics 2006, 22, e174. (In the Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB), Fortaleza, Brazil).
30. Pierce, N.; Spriet, J.; Desmet, J.; Mayo, S. J Comput Chem 2000, 21, 999.
31. Gordon, D.; Mayo, S. J Comput Chem 1998, 19, 1505.
32. Lasters, I.; Desmet, J. Prot Eng 1993, 6, 717.
33. Goldstein, R. Biophys J 1994, 66, 1335.
34. Tucker-Kellogg, L. (2002). Systematic Conformational Search with Constraint Satisfaction. PhD thesis, Massachusetts Institute of Technology, MA.
35. Rienstra, C.; Tucker-Kellogg, L.; Jaroniec, C.; Hohwy, M.; Reif, B.; McMahon, M.; Tidor, B.; Lozano-Pérez, T.; Griffin, R. Proc Natl Acad Sci USA 2002, 99, 10260.
36. Leach, A.; Lemon, A. Proteins 1998, 33, 227.
37. Conti, E.; Stachelhaus, T.; Marahiel, M.; Brick, P. EMBO J 1997, 16, 4174.
38. Weiner, S.; Kollman, P.; Case, D.; Singh, U.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. J Am Chem Soc 1984, 106, 765.
39. Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. J Am Chem Soc 1995, 117, 5179.
40. Lazaridis, T.; Karplus, M. Proteins: Struct Funct Genet 1999, 35, 133.
41. Dahiyat, B; Mayo, S. Science 1997, 278, 82.
42. Looger, L.; Hellinga, H. J Mol Biol 2001, 307, 429.
43. Dwyer, M.; Looger, L.; Hellinga, H. Science 2004, 304, 1967.
44. Georgiev, I.; Donald, B. R. Bioinformatics 2007, 23, i185.