1-1-2006

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

Ivelin Georgiev
Dartmouth College

Ryan H. Lilien
Dartmouth College

Bruce R. Donald
Dartmouth College

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

Ivelin Georgiev[1,⋆], Ryan H. Lilien[1,2,3,⋆], and Bruce R. Donald[1,3,4,5,⋆⋆]

[1] Dartmouth Computer Science Department, Hanover, NH 03755, USA
[2] Dartmouth Medical School, Hanover, NH
[3] Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH
[4] Dartmouth Department of Chemistry, Hanover, NH
[5] Dartmouth Department of Biological Sciences, Hanover, NH

Abstract. 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. 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 post-processing 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.

A revised version of this paper will appear in the Proceedings of the Tenth Annual International Conference on Research in Computational Molecular Biology (RECOMB), Venice Lido, Italy, April 2006.

DARTMOUTH COMPUTER SCIENCE TECHNICAL REPORT 2006-570
http://www.cs.dartmouth.edu/reports/abstracts/TR2006-570

⋆ These authors contributed equally to the work.

⋆⋆ Corresponding author, Bruce.R.Donald@dartmouth.edu. This work is supported by grants to B.R.D. from the National Institutes of Health (R01 GM-65982), and the National Science Foundation (EIA-0305444).

1 Introduction

1.1 Computational Protein Design

The ability to engineer proteins has many biomedical applications. 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. 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 [35, 13, 12, 11, 1, 25, 20, 14, 32]. In [24], a number of bound and unbound structures are compared, and the conclusion is drawn that only a small number of residues undergo conformational change, and that the structural changes are primarily side-chains, and not backbone. Hence, many protein design algorithms use a rigid backbone and model side-chain flexibility with a rotamer library, containing a discrete set of low-energy commonly-observed side-chain conformations [21, 28]. The major challenge for protein redesign algorithms is the efficient evaluation of the exponential number of candidate protein conformations, resulting not only from mutating residues along the peptide chain, but also by employing these discrete 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.

Non-ensemble-based algorithms for protein redesign are based on the assumption that protein folding and binding can be accurately predicted by examining the 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 [27], different heuristic approaches (random sampling, neural network, and genetic algorithm) have been proposed [35, 13, 12, 11, 22]. A provable and efficient deterministic algorithm, which has become the dominant choice for non-ensemble-based protein design, is Dead-End Elimination (DEE) [6, 15]. 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. Complexity can be further reduced through the use of an A* branch-and-bound algorithm to enumerate conformations in order of increasing energy [16].

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*). 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 enjoying a combinatorial pruning factor for rotamers that are provably not part of the minimized-GMEC. The extension of the DEE framework to include energy minimization is useful not only in identifying the minimized-GMEC, but also as a filter

in an ensemble scoring method for protein redesign based on energy-minimized conformations (such as K^* [17, 18]).

1.2 NRPS Redesign and K^*

Traditional ribosomal peptide synthesis is complemented by non-ribosomal 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 [33, 3, 31]. Among the products of NRPS enzymes are natural antibiotics (penicillin, vancomycin), antifungals, antivirals, immunosuppressants, and antineoplastics. The redesign of NRPS enzymes can lead to the synthesis of novel NRPS products, such as new libraries of antibiotics [2].

The main techniques for NRPS enzyme redesign are *domain-swapping* [34, 30, 7, 23, 19], *signature sequences* [33, 8, 3], and *active site manipulation from a structure-based mutation search utilizing ensemble docking* (the K^* method [17, 18]). A review of these methods and a discussion of the advantages of structure-based redesign (the K^* method) over the other two techniques can be found in [17, 18].

The K^* algorithm [17, 18] has been demonstrated for NRPS redesign, but is a general algorithm that is, in principle, capable of redesigning any protein. 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 (K^*) to the binding constant for a given 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 [17, 18].

1.3 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 demonstrate 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 minimized-GMEC will be pruned. We show that our method is useful not only in (a) identifying the minimized-GMEC (a non-ensemble-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/A^* algorithm in a search to switch the binding affinity of the Phe-specific adenylation domain of the non-ribosomal peptide synthetase 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 A^* search and our previous work on K^* [17, 18]. For brevity, we will henceforth refer to this algorithm as the *Hybrid MinDEE-K^** algorithm. The Hybrid MinDEE-K^* algorithm has two stages. 1) the MinDEE filter provably prunes the majority of conformations, and 2) remaining conformations are generated in order of increasing lower bounds on their minimized energies using the A^* branch-and-bound search. As each conformation is generated by A^*, it is energy-minimized and added to the partition function. The partition function approximation algorithm halts the A^* search as soon as the next A^* conformation's lower-energy bound exceeds a computed threshold. The resulting method is a provably-accurate approximation algorithm to compute partition functions and subsequently binding constants for protein-ligand binding. Our new algorithm prunes a combinatorial number of conformations, an improvement over the constant-factor speedup in the original K^* [17, 18]. 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-K^* algorithm has a much higher pruning efficiency than the original K^* algorithm. Moreover, it takes only *30 seconds* for MinDEE to determine which rotamers can be provably pruned. We make the following contributions in this paper:

1. Derivation of minimized-DEE (MinDEE), a novel, provable, and efficient DEE-like algorithm that incorporates energy minimization, with applications in both non-ensemble- and ensemble-based protein redesign. Analogously to the extensions to traditional DEE [6, 15, 9, 26], we also derive extensions to MinDEE to improve its pruning efficiency;
2. Introduction of a MinDEE/A^* algorithm that identifies the minimized-GMEC and returns a set of low-energy conformations;
3. Introduction of a hybrid MinDEE-K^* ensemble-based scoring and search algorithm, improving on our previous work on K^* [17, 18] by replacing a constant-factor with a combinatorial-factor provable pruning condition;
4. Derivation of provably-accurate ε-approximation algorithms for partition function computation; and
5. 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.

2 Derivation of the Minimized-DEE Criterion

2.1 The Original DEE Criterion

In this section we briefly review the *traditional-DEE* theorem [6, 26, 10, 15]. 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 $E_{_T} = E_{t'} + \sum_i E(i_r) + \sum_i \sum_{j>i} E(i_r, j_s)$, where $E_{t'}$ is the template self-energy

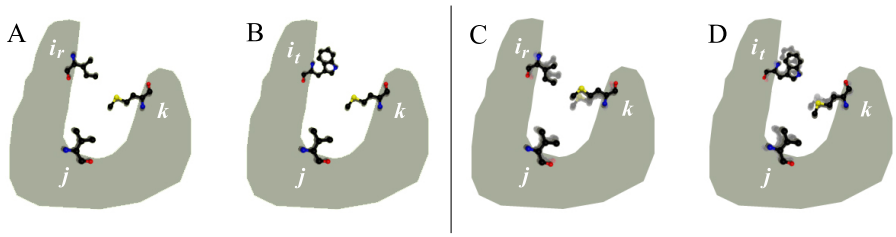

Fig. 1. Energy-Minimized DEE. Without energy minimization the swapping of rotamer i_r for i_t (Panel A to Panel B) leaves unchanged the conformations and self and pairwise energies of residues j and k. When energy minimization is allowed, the swapping of rotamer i_r for rotamer i_t (Panel C to Panel 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).

(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 $(i \neq j)$, $\max\limits_{s \in R_j} E(i_t, j_s) \geq E(i_t, j_g)$ and $\min\limits_{s \in R_j} E(i_g, j_s) \leq E(i_g, j_g)$, 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:

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

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.' Extensions to this initial DEE criterion allow for additional pruning while maintaining correctness with respect to identifying the rigid-GMEC [6, 15, 9, 10, 26].

2.2 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. 1). 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* [36, 29] 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[1] 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:

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

Analogous definitions exist for pairwise terms (Fig. 4 in the Appendix). 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

$$A^\bullet = (A_1^\bullet, A_2^\bullet, \cdots, A_n^\bullet) = \operatorname*{argmin}_{B=(B_1,B_2,\cdots,B_n) \in \prod_{i=1}^{n} \mathcal{V}(A_i)} E(B) \tag{2}$$

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:

$$E_{_T}(A) = E_{t'} + \sum_i E_{\odot}(i_r|A) + \sum_i \sum_{j>i} E_{\odot}(i_r, j_s|A). \tag{3}$$

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

$$\begin{aligned} E_{t'} + E_{\odot}(i_t|G_{i_g \to i_t}) + \sum_{j \neq i} E_{\odot}(i_t, j_g|G_{i_g \to i_t}) + \sum_{j \neq i} E_{\odot}(j_g|G_{i_g \to i_t}) \\ + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\odot}(j_g, k_g|G_{i_g \to i_t}) \geq E_{t'} + E_{\odot}(i_g|G) \\ + \sum_{j \neq i} E_{\odot}(i_g, j_g|G) + \sum_{j \neq i} E_{\odot}(j_g|G) + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\odot}(j_g, k_g|G). \end{aligned} \tag{4}$$

The $E_{t'}$ terms (Sec. 2.1) correspond to the rigid portion of the molecule; they are independent of rotamer choice, are equal, and can be canceled. We make the following

[1] 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).

trivial upper and lower-bound observations:

$$E_{\odot}(i_t|A) \leq E_{\oplus}(i_t);\ E_{\odot}(i_t, j_g|A) \leq \max_{s \in R_j} E_{\oplus}(i_t, j_s); \tag{5}$$

$$E_{\odot}(j_g|A) \leq E_{\oplus}(j_g);\ E_{\odot}(j_g, k_g|A) \leq E_{\oplus}(j_g, k_g); \tag{6}$$

$$E_{\ominus}(i_g) \leq E_{\odot}(i_g|A);\ \min_{s \in R_j} E_{\ominus}(i_g, j_s) \leq E_{\odot}(i_g, j_g|A); \tag{7}$$

$$E_{\ominus}(j_g) \leq E_{\odot}(j_g|A);\ E_{\ominus}(j_g, k_g) \leq E_{\odot}(j_g, k_g|A). \tag{8}$$

Substituting Eqs. (5-8) into Eq. (4), we obtain:

$$E_{\oplus}(i_t) + \sum_{j \neq i} \max_s E_{\oplus}(i_t, j_s) + \sum_{j \neq i} E_{\oplus}(j_g) + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\oplus}(j_g, k_g) \geq$$
$$E_{\ominus}(i_g) + \sum_{j \neq i} \min_s E_{\ominus}(i_g, j_s) + \sum_{j \neq i} E_{\ominus}(j_g) + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\ominus}(j_g, k_g). \tag{9}$$

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

$$E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s) - \sum_{j \neq i} \max_s E_{\oslash}(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_{\oslash}(j_s, k_u) >$$
$$E_{\oplus}(i_t) + \sum_{j \neq i} \max_s E_{\oplus}(i_t, j_s). \tag{10}$$

Proposition 1. *When Eq. (10) holds, rotamer i_r is provably not part of the minimized-GMEC.*

Proof. When Eq. (10) holds, we can substitute the left-hand side of Eq. (10) for the first two terms of Eq. (9), and simplify the resulting equation to:

$$E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s) - \sum_{j \neq i} \max_s E_{\oslash}(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_{\oslash}(j_s, k_u)$$
$$+ \sum_{j \neq i} E_{\oslash}(j_g) + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\oslash}(j_g, k_g) > E_{\ominus}(i_g) + \sum_{j \neq i} \min_s E_{\ominus}(i_g, j_s). \tag{11}$$

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. (11) and reduce:

$$E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s) > E_{\ominus}(i_g) + \sum_{j \neq i} \min_s E_{\ominus}(i_g, j_s). \tag{12}$$

Thus, when the MinDEE pruning condition Eq. (10) holds, $i_r \neq i_g$ and we can provably eliminate rotamer i_r as not being part of the energy-minimized GMEC. □

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. 10) can be computed as efficiently as the traditional-DEE pruning condition (Eq. 1). 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. We have also generalized the extensions to traditional-DEE pruning [6, 15, 9, 10, 26] for MinDEE, see Appendix A.

3 Minimized-DEE/A^* Search Algorithm (Non-Ensemble-Based Redesign)

3.1 Traditional-DEE with A^*

In [16], an 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.[2]

The following derivation of the DEE/A^* combined search closely follows [16]. The 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 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: $E(i_r) - E(i_t) + \sum_{j \neq i} \min_s E(i_r, j_s) - \sum_{j \neq i} \max_s E(i_t, j_s) > E_w$.

3.2 MinDEE with A^*

The traditional-DEE/A^* algorithm [16] can be extended to include energy minimization by substituting our newly derived MinDEE (Sec. 2.2) for traditional-DEE. So that no

[2] 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 (see Fig. 3 in [18, p. 745]).

conformations within E_w of the energy-minimized GMEC are pruned, the MinDEE equation (Eq. 10) becomes:

$$E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s) - \sum_{j \neq i} \max_s E_{\oslash}(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_{\oslash}(j_s, k_u) \\ - E_{\oplus}(i_t) - \sum_{j \neq i} \max_s E_{\oplus}(i_t, j_s) > E_w \ . \tag{13}$$

We modify the definition of the 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:

$$g = \sum_{i=1}^{d} (E_{\ominus}(i_r) + \sum_{j=i+1}^{d} E_{\ominus}(i_r, j_s)), \quad h = \sum_{j=d+1}^{n} E_j, \tag{14}$$

where

$$E_j = \min_s \left(E_{\ominus}(j_s) + \sum_{i=1}^{d} E_{\ominus}(i_r, j_s) + \sum_{k=j+1}^{n} \min_u E_{\ominus}(j_s, k_u) \right). \tag{15}$$

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/A^* search generates conformations in order of increasing *lower bounds* on the conformation's *minimized* energy.

We combine our modified MinDEE criterion (Eq. 13) with the modified A^* functions (Eqs. 14-15) 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 MinDEE/A^* search must guarantee that upon completion all conformations within E_w of the minimized-GMEC are returned. Since in the 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/A^* search. Let $B(s)$ be the lower bound on the energy of conformation s (see Appendix B, 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 A^* search.

Proposition 2. *The MinDEE/A^* search can be halted once the lower bound $B(c)$ on the energy of the next conformation c returned by A^*, satisfies $B(c) > E_m + E_w$. 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 A^*) and all conformations not yet returned. Since 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 $B(c) > E_m + E_w$ holds, then $E(s) > E_m + E_w$. 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 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/A^* algorithm incorporates energy minimization with provable guarantees, and is thus more capable of returning conformations with lower energy states than traditional-DEE.

4 Hybrid MinDEE-K^* Algorithm (Ensemble-Based Redesign)

We now present an extension and improvement to the original K^* algorithm [17, 18] by using a version of the MinDEE criterion plus A^* branch-and-bound search. The K^* ensemble-based scoring function approximates the protein-ligand binding constant with the following quotient: $K^* = \frac{q_{PL}}{q_P q_L}$, 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. 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 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 [17, 18]. In this section we illustrate how the newly-derived MinDEE and A^* algorithms 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 K^* scoring function is based on *energy-minimized* conformations. Since pruned conformations never have to be examined, the Hybrid MinDEE-K^* algorithm provides a combinatorial improvement in runtime over the previously described constant-factor ε-approximation algorithm [17, 18] (where a lower-bound on *each* conformation's minimum energy was quickly examined to determine if full energy minimization was required).

The MinDEE criterion (Eq. 10) can prune rotamers across mutation sequences.[3] By pruning *across* mutations with MinDEE, we risk pruning conformations that could otherwise contribute substantially to the computed partition functions, thus violating our

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

provably-good approximation to the full partition functions (Sec. 4.1). Hence, we derive a modified version of MinDEE, called *Single-Sequence* MinDEE (*SSMinDEE*), that is capable of pruning rotamers only within a single mutation sequence. The MinDEE criterion (Eq. 10) is valid for SSMinDEE; the only distinction is that the set of rotamers for each residue position consists only of the rotamers for the amino acid type of the given residue in the current mutation sequence, rather than of the rotamers for all possible amino acid types for that residue, as is the case with the *multiple-sequence minimized-DEE* described in Sec. 2.2.

4.1 Efficient Partition Function Computation Using A^* Search

Using the A^* algorithm with SSMinDEE, 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 B). 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 stop the A^* search once $B(c)$ becomes greater than the threshold. The 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. Thus, q^* is an ε-approximation to q, i.e., $q^* \geq (1-\varepsilon)q$.

Since the application of the MinDEE criterion (Eq. 10) 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 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 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 A^* satisfies $B(c) \geq -RT\left(\ln(q^*\rho - k\exp(-E_0/RT)) - \ln n\right)$,*

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:

$$q = q^* + q' + p^*. \tag{16}$$

Thus,

$$q \leq q^* + n\exp(-E_t/RT) + k\exp(-E_0/RT). \tag{17}$$

Hence, if

$$q^* \geq (1-\varepsilon)(q^* + n\exp(-E_t/RT) + k\exp(-E_0/RT)), \tag{18}$$

then $q^* \geq (1-\varepsilon)q$. Solving Eq. (18) for E_t, we obtain the desired stop-threshold:

$$-RT\left(\ln(q^*\rho - k\exp(-E_0/RT)) - \ln n\right) \leq E_t. \tag{19}$$

We can halt the search once a conformation's energy lower bound becomes greater than the stop-threshold (Eq. 19), 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. 13), 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 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 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

$$q^* \geq (1-\delta)(q^* + k\exp(-E_0/RT)), \tag{20}$$

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 re-introduced 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

$$q^* \geq (1-\varepsilon)(q^* + (k-l)\exp(-E_0/RT)). \tag{21}$$

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

$$l \geq k - \frac{q^*\rho}{\exp(-E_0/RT)}, \tag{22}$$

where again $\rho = \frac{\varepsilon}{1-\varepsilon}$. Note that an ε-approximation may be achieved before all conformations have been extracted; Eq. (22) guarantees such an accuracy when all non-pruned conformations have been extracted by 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.

Proposition 3 represents an *intra-mutation* energy filter (Fig. 2) for pruning within a single mutation sequence. We now derive a provably-accurate partition-function approximation for pruning *across* mutation sequences.

4.2 Inter-Mutation Filter

We first review some of the definitions from [17, 18]. The main motivation for the inter-mutation filter is that we must compute provably-accurate scores only for the top fraction of the mutation sequences. 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 K^* mutation sequence; $\gamma = 0.0$ computes an ε-approximation for all K^* mutation sequences. Let us assume that 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 i^{th} mutation sequence to be $K_i^* = \frac{q_{PL}}{q_P q_L}$, while $K_o^* = \frac{{}^o q_{PL}}{{}^o q_P {}^o q_L}$. We let q_{PL}^* be the partial partition function for the bound protein-ligand state,

```
Initialize: n ← Number of Rotameric Conformations; q* ← 0
while (n > 0)
      c ← GetNextAStarConf()
      if B(c) ≤ −RT (ln(q*ρ − k exp(−E0/RT)) − ln n)
            q* ← q* + exp (−ComputeMinEnergy(c)/RT)
            n ← n − 1
      else  Return q*
if q*ρ < k exp(−E0/RT)
      RepeatSearch(q*, ρ, k, E0)
else  Return q*
```

Fig. 2. Intra-Mutation Filter for Computing a Partition Function with Energy Minimization Using the A* Search. q^* is the running approximation to the partition function. The function $\mathtt{B}(\cdot)$ computes the energy lower bound for the given conformation (see Appendix B). The function $\mathtt{ComputeMinEnergy}(\cdot)$ returns a conformation's energy after energy minimization. The function $\mathtt{GetNextAStarConf}()$ returns the next conformation from the A* search. The function $\mathtt{RepeatSearch}(\cdot)$ sets up and repeats the mutation search if an ε-approximation is not achieved after the generation of all A^* conformations; the search is repeated at most once. Upon completion, q^* represents an ε-approximation to the true partition function q, such that $q^* \geq (1-\varepsilon)q$.

computed from the m already-generated conformations. We define $K_o^\dagger = \frac{{}^o q_{PL}}{{}^o q_P}$. Finally, let $\psi = \max\left(\gamma\varepsilon K_o^\dagger q_P,\ q^*_{PL}\rho\right)$ and $\rho = \frac{\varepsilon}{1-\varepsilon}$.

Proposition 4. *If the lower bound $B(c)$ on the minimized energy of the $(m+1)^{\text{st}}$ conformation returned by 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, $q_L = {}^o q_L$. Let us assume that we have a sequence for which $K_i^* \geq \gamma K_o^*$ holds. Thus,

$$\frac{q_{PL}}{q_P q_L} \geq \gamma \frac{{}^o q_{PL}}{{}^o q_P {}^o q_L}\ ,$$
$$q_{PL} \geq \gamma K_o^\dagger q_P\ . \tag{23}$$

First, we note again that

$$q' \leq n\exp(-E_t/RT); \tag{24}$$
$$p^* \leq k\exp(-E_0/RT). \tag{25}$$

From the definition of q_{PL}, we obtain

$$q_{PL} = q^*_{PL} + q' + p^*. \tag{26}$$

Now, if

$$n\exp(-E_t/RT) + k\exp(-E_0/RT) \leq \varepsilon K_o^\dagger \gamma q_P, \tag{27}$$

then by Eqs. (24) and (25) we have

$$q' + p^* \leq \varepsilon K_o^\dagger \gamma q_{_P}, \tag{28}$$

and by Eq. (23),

$$q' + p^* \leq \varepsilon \; q_{_{PL}}, \tag{29}$$

and finally, by Eq. (26), we obtain

$$q^*_{_{PL}} \geq (1-\varepsilon) q_{_{PL}}, \tag{30}$$

which is the definition of the partition function ε-approximation. Thus, if Eq. (27) holds, then we will have an ε-approximation to the true partition function $q_{_{PL}}$. Solving Eq. (27) for E_t, we obtain the stop-threshold:

$$E_t \geq -RT \left(\ln \left(\gamma \varepsilon K_o^\dagger q_{_P} - k \exp(-E_{_0}/RT) \right) - \ln n \right). \tag{31}$$

The first conformation that has an energy above the stop-threshold (Eq. 31) halts the partition function computation, since we already have an ε-approximation. Thus, combining Eq. (31) and the *intra*-mutation stop-threshold (Eq. 19), our stopping condition for the computation of $q_{_{PL}}$ becomes

$$B(c) > -RT \left(\ln \left(\psi - k \exp(-E_{_0}/RT) \right) - \ln n \right), \tag{32}$$

where $\psi = \max \left(\gamma \varepsilon K_o^\dagger q_{_P}, \; q^*_{_{PL}} \rho \right)$ 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 A^*, we can modify Eq. (22) 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 [17, 18] which is valid when the energy lower bounds for all of the conformations are evaluated. We should note that Eq. (32) 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 [17, 18], we define $\tilde{K_i^*} = \frac{q^*_{_{PL}}}{q^*_{_P} q_{_L}}$ 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 $\tilde{K_i^*} \in \left[K_i^*(1-\varepsilon), \frac{1}{1-\varepsilon} K_i^* \right]$. If $K_i^* \geq \gamma K_o^*$ holds for a mutation

sequence i, then by Proposition 4, $q_{PL} > q^*_{PL} \geq (1-\varepsilon)q_{PL}$. Also, since q_P is already computed using Proposition 3, $q_P > q^*_P \geq (1-\varepsilon)q_P$. Since $K^*_i = \frac{q_{PL}}{q_P q_L}$, we have

$$\left[K^*_i(1-\varepsilon) \leq \tilde{K^*_i} \leq \frac{1}{1-\varepsilon} K^*_i \right].$$

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

4.3 Algorithm

We now have all the necessary tools for our ensemble-based Hybrid MinDEE-K^* algorithm. The volume filter (see Sec. 5) in the original K^* is applied first to eliminate under- and over-packed mutation sequences; this is followed by the combinatorial SS-MinDEE filter and the A^* energy filter using the ε-approximation algorithms in Sec. 4.1 and 4.2 (see Table 1), which improve on the mere constant-factor speedup provided by the energy filter in the original K^* [17, 18]. By implementing a steric filter (see Sec. 5), similar to the one in [17, 18], as a part of the A^* search, we prevent some high-energy conformations (corresponding to steric clashes) with good lower bounds from being returned by 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. Finally, the K^* score for a given mutation is computed as the ratio of the bound and unbound partition functions: $K^* = \frac{q_{PL}}{q_P q_L}$. Our Hybrid MinDEE-$K^*$ 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 score predictions.

5 Methods

Structural Model. Our structural model is the same as the one used in the original K^* [17, 18]. In our experiments, the structural model consists of the 9 active site residues (D235, A236, W239, T278, I299, A301, A322, I330, C331) of GrsA-PheA (PDB id: 1AMU) [4], the steric shell (the 30 residues with at least one atom within 8 Å of a residue in the active site), 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 PheA protein. The residues of the active site modeled as flexible using rotamers and subject to energy minimization include: 235D, 236A, 239W, 278T, 299I, 301A, 322A, 330I, and 331C. The steric shell includes all residues not modeled as flexible and that contain at least one atom within 8 Å of the active site. The steric shell residues include: 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 [4], and also in [17, 18], 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 Lovell *et al*. rotamer library [21]. Each rotameric-based conformation is minimized using steepest-descent minimization (see Appendix B) and the AMBER energy function (electrostatic, vdW, and dihedral energy terms) [37, 5].

Energy Precomputation for Lower Bounds, $\mathbf{B}(\cdot)$. The MinDEE criterion (Eq. 10) uses both min and max *precomputed* energy terms to determine which rotamers are not part of the minimized-GMEC. There is no need to re-compute the min and max energies every time Eq. (10) is evaluated. See Appendix B 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, which requires that correct 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 1.5Å 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 to the simple coupled Goldstein criterion ([9] and Fig. 4d in the Appendix).

6 Results and Discussion

In this section, we compare the results of GMEC-based protein redesign without (traditional-DEE/A^*) and with (MinDEE/A^*) energy minimization. We also compare the redesign results when energy minimization is used without (MinDEE/A^*) and with (Hybrid MinDEE-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.

6.1 Comparison to Biological Activity Assays

Similarly to [17, 18], we simulated the biological activity assays of L-Phe and L-Leu against the wildtype PheA enzyme and the double mutant T278M/A301G [33]. In [33], T278M/A301G was shown to have the desired switch of specificity from Phe to Leu by performing activity assays. The activity for both the wildtype and the mutant protein sequences was normalized, so that the substrate with the larger activity was assigned a specificity of 100%, while the other substrate was assigned specificity relative to the first one. The wildtype PheA had a specificity of 100% for Phe and approximately 7% for Leu; the double mutant had a specificity of 100% for Leu and approximately 40% for Phe. The computed Hybrid MinDEE-K^* normalized scores qualitatively agreed with these results, showing the desired switch of specificity for T278M/A301G. The wildtype sequence had a normalized K^* score of 100% for Phe and 0.01% for Leu; the double mutant had a normalized score of 100% for Leu and 20% for Phe.

Table 1. Conformational Pruning with Hybrid MinDEE-K^*. 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, single-sequence minimized-DEE, steric, and energy (with A^*) pruning. The A^* energy filter is based on the ε-approximation algorithms in Secs. 4.1 and 4.2. 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.

	Conf. Remaining	Pruning Factor (%)
Initial	6.8×10^8	-
Volume Filter	2.04×10^8	3.33 (70.0)
SSMinDEE Filter	8.83×10^6	23.12 (95.7)
Steric Filter	5.76×10^6	1.53 (34.7)
A^* Energy Filter	2.78×10^5	20.7 (95.2)

6.2 Comparison to Traditional-DEE

For comparison, the simple coupled Goldstein traditional-DEE criterion [9] was used in a redesign search for changing the specificity of the wildtype PheA enzyme from Phe to Leu, using the experimental setup in Sec. 5. A comparison to the rotamers in the minimized-GMEC A236M/A322M (Sec. 6.3), revealed that 2 of these 9 rotamers were in fact *pruned* by traditional-DEE. As an example, the minimized-GMEC was energy-minimized from a conformation that included rotamer 5 [21] of Met at residue 236. This particular rotamer (χ angles $-177°$, $180°$, and $75°$) was pruned by traditional-DEE. We then energy-minimized A236M/A301G, the rigid-GMEC obtained by traditional-DEE/A^* and determined that its energy was higher (by appx. 5 kcal/mol) than the energy for the minimized-GMEC obtained by MinDEE/A^*. Moreover, a total of 104 different conformations minimized to an energy lower than the minimized rigid-GMEC energy. 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.

6.3 Redesign for Leu

Hybrid MinDEE-K^* The experimental setup for Leu redesign with Hybrid MinDEE-K^* is as described in Sec 5. The 2-point mutation search took approximately 10 hours on a cluster of 24 processors. Only 30% of the mutation sequences passed the volume filter, while MinDEE pruned over 95% 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 twenty (see Table 1). A brute-force version of Hybrid MinDEE-K^* that did not utilize any of the filters, would take approximately 2,450 times longer (appx. 1,023 days) for the same experimental setup for redesign.

An initial comparison to the original K^* results showed only a small overlap between the top-ranking mutations for Hybrid MinDEE-K^* and the original K^*[17, 18].

Since the two algorithms use different energy-minimization modules (see Appendix B), we then applied the better Hybrid MinDEE-K^* minimization scheme to the original K^*, in order to facilitate a fair comparison. 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.

The two top-scoring sequences are A301G/I330W and A301G/I330F for both Hybrid MinDEE-K^* and the original-K^*. These novel mutation sequences were tested in the wetlab and were shown to have the desired switch of specificity from Phe to Leu [17, 18]. Moreover, the other known successful redesign T278M/A301G [33] is ranked 4^{th} by both Hybrid MinDEE-K^* and the modified version of the original K^* algorithm (this sequence was ranked 12^{th} by the unmodified original K^* in [17, 18]). Furthermore, all of the top 17 Hybrid MinDEE-K^* sequences contain the mutation A301G, which is found in all known native Leu adenylation domains [3]. These results show that our algorithms can give reasonable predictions for redesign.

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 modified original K^* algorithm fully-evaluated approximately 30% more conformations than the 2.78×10^5 conformations fully evaluated by Hybrid MinDEE-K^* (see Table 1). Thus, Hybrid MinDEE-K^* is much more efficient at obtaining the desired results.

MinDEE/A^* We now discuss results from our non-ensemble-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 Sec. 5. 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 $E_w = 8.5$ kcal/mol. A total of 813 conformations, representing 45 unique mutation sequences, had actual minimized energies within 8.5 kcal/mol of the minimized-GMEC energy, 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 approximately 14 days on a single processor, with more than $120,000$ 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.

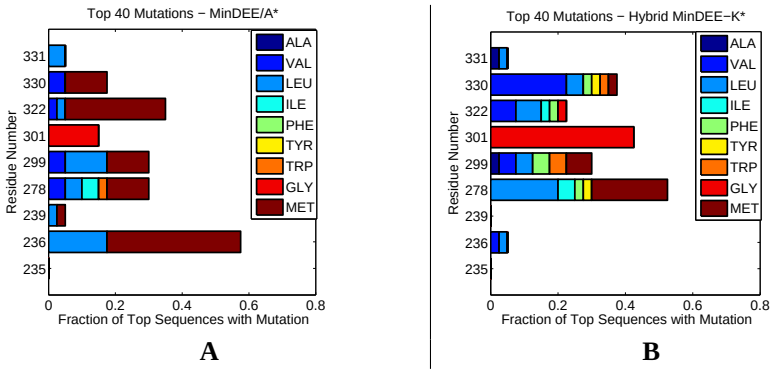

Fig. 3. Distribution of Mutations. The distribution of the mutation types for the top 40 mutation sequences for **(A)** MinDEE/A^* and **(B)** Hybrid MinDEE-K^* algorithms is shown as the fraction of each mutating type for each active site residue. The types and frequencies for the mutations are quite different for the two methods, which indicates that the difference in the information content for non-ensemble and ensemble-based algorithms can be substantial. The variability of mutation types for each active site residue is 2.38 (types per residue, for the residues that were mutated at least once in the given mutation sequences) for MinDEE/A^* and 3.86 for Hybrid MinDEE-K^*, suggesting the existence of multiple beneficial mutation types for the mutated residues.

Like A301G/I330W and A301G/I330F, the top 5 MinDEE/A^* sequences are unknown in nature. To assess the switch of specificity from Phe to Leu for the novel mutation sequences, we extracted the minimum-energy conformation for these top 5 Leu-binding sequences. Each of these 5 conformations was then energy-minimized when bound to Phe. Whereas the Leu-bound energies were negative and low, the corresponding Phe-bound energies were positive and high. Thus, the top mutation sequences represented by their minimum energy conformation are predicted to bind more stably to Leu than to Phe, as desired.

Only 9 of the 45 MinDEE/A^* mutation sequences passed the volume filter of Hybrid MinDEE-K^*. Moreover, only 5 of the MinDEE/A^* sequences could be found in the top 40 Hybrid MinDEE-K^* sequences, 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. 3). We can conclude that, currently, MinDEE appears useful as a filter in the Hybrid MinDEE-K^* algorithm; however, the incorporation of additional information, such as a comparison to negative design (the energies to bind the wild-type substrate), may promote MinDEE as a valuable stand-alone non-ensemble-based algorithm for protein redesign.

7 Limitations and Extensions

The MinDEE criterion can efficiently prune a large number of the possible conformations (see Sec. 6.3). However, because of the use of min and max energy terms, the prun-

ing 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 (while 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. 10), depending on the proximity of the residues involved.

8 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-K^* algorithm showed a significant improvement in pruning efficiency, as compared to the original 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, remove bias in the rotamer library by factoring the Jacobian into the partition function over torsion-angle space, and incorporate backbone flexibility. An accurate and efficient algorithm for redesigning natural products should prove useful as a technique for drug design.

Acknowledgments We 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.

References

[1] D. Bolon and S. Mayo. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA*, 98:14274–14279, 2001.

[2] D. Cane, C. Walsh, and C. Khosla. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science*, 282:63–68, 1998.

[3] G. Challis, J. Ravel, and C. Townsend. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.*, 7:211–224, 2000.

[4] E. Conti, T. Stachelhaus, M. Marahiel, and P. Brick. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of Gramicidin S. *EMBO J.*, 16:4174–4183, 1997.

[5] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[6] J. Desmet, M. Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

[7] S. Doekel and M. Marahiel. Dipeptide formation on engineered hybrid peptide synthetases. *Chem. Biol.*, 7:373–384, 2000.

[8] K. Eppelmann, T. Stachelhaus, and M. Marahiel. Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry*, 41:9718–9726, 2002.

[9] R. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.*, 66:1335–1340, 1994.

[10] D. Gordon and S. Mayo. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.*, 19:1505–1514, 1998.

[11] H. Hellinga and F. Richards. Construction of new ligand binding sites in proteins of known structure: I. Computer-aided modeling of sites with pre-defined geometry. *J. Mol. Biol.*, 222:763–785, 1991.

[12] A. Jaramillo, L. Wernisch, S. Héry, and S. Wodak. Automatic procedures for protein design. *Comb. Chem. High Throughput Screen.*, 4:643–659, 2001.

[13] W. Jin, O. Kambara, H. Sasakawa, A. Tamura, and S. Takada. De novo design of foldable proteins with smooth folding funnel: Automated negative design and experimental verification. *Structure*, 11:581–591, 2003.

[14] A. Keating, V. Malashkevich, B. Tidor, and P. Kim. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. *Proc. Natl. Acad. Sci. USA*, 98:14825–14830, 2001.

[15] I. Lasters and J. Desmet. The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Protein Eng.*, 6:717–722, 1993.

[16] A. Leach and A. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33:227–239, 1998.

[17] R. Lilien, B. Stevens, A. Anderson, and B. R. Donald. A novel ensemble-based scoring and search algorithm for protein redesign, and its application to modify the substrate specificity of the Gramicidin Synthetase A phenylalanine adenylation enzyme. In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 46–57, San Diego, March 2004.

[18] R. Lilien, B. Stevens, A. Anderson, and B. R. Donald. A novel ensemble-based scoring and search algorithm for protein redesign, and its application to modify the substrate specificity of the Gramicidin Synthetase A phenylalanine adenylation enzyme. *Journal of Computational Biology*, 12(6–7):740–761, 2005.

[19] U. Linne, S. Doekel, and M. Marahiel. Portability of epimerization domain and role of peptidyl carrier protein on epimerization activity in nonribosomal peptide synthetases. *Biochemistry*, 40:15824–15834, 2001.

[20] L. Looger, M. Dwyer, J. Smith, and H. Hellinga. Computational design of receptor and sensor proteins with novel functions. *Nature*, 423:185–190, 2003.

[21] S. Lovell, J. Word, J. Richardson, and D. Richardson. The penultimate rotamer library. *Proteins*, 40:389–408, 2000.

[22] J. Marvin and H. Hellinga. Conversion of a maltose receptor into a zinc biosensor by computational design. *PNAS*, 98:4955–4960, 2001.

[23] H. Mootz, D. Schwarzer, and M. Marahiel. Construction of hybrid peptide synthetases by module and domain fusions. *Proc. Natl. Acad. Sci. USA*, 97:5848–5853, 2000.

[24] R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman. Side-chain flexibility in proteins upon ligand binding. *Proteins*, 39(3):261–8, 2000.

[25] F. Offredi, F. Dubail, P. Kischel, K. Sarinski, A. Stern, C. Van de Weerdt, J. Hoch, C. Prosperi, J. François, S. Mayo, and J. Martial. De novo backbone and sequence design of an idealized α/β-barrel protein: evidence of stable tertiary structure. *J. Mol. Biol.*, 325:163–174, 2003.

[26] N. Pierce, J. Spriet, J. Desmet, and S. Mayo. Conformational splitting: a more powerful criterion for dead-end elimination. *J. Comput. Chem.*, 21:999–1009, 2000.

[27] N. Pierce and E. Winfree. Protein design is *NP*-hard. *Protein Eng.*, 15:779–782, 2002.

[28] J. Ponder and F. Richards. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193:775–791, 1987.

[29] C. Rienstra, L. Tucker-Kellogg, C. Jaroniec, M. Hohwy, B. Reif, M. McMahon, B. Tidor, T. Lozano-Pérez, and R. Griffin. De novo determination of peptide structure with solid-state magic-angle spinning NMR spectroscopy. *Proc. Natl. Acad. Sci. USA*, 99:10260–10265, 2002.

[30] A. Schneider, T. Stachelhaus, and M. Marahiel. Targeted alteration of the substrate specificity of peptide synthetases by rational module swapping. *Mol. Gen. Genet.*, 257:308–318, 1998.

[31] D. Schwarzer, R. Finking, and M. Marahiel. Nonribosomal peptides: from genes to products. *Nat. Prod. Rep.*, 20:275–287, 2003.

[32] J. Shifman and S. Mayo. Modulating calmodulin binding specificity through computational protein design. *J. Mol. Biol.*, 323:417–423, 2002.

[33] T. Stachelhaus, H. Mootz, and M. Marahiel. The specificiy-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, 6:493–505, 1999.

[34] T. Stachelhaus, A. Schneider, and M. Marahiel. Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science*, 269:69–72, 1995.

[35] A. Street and S. Mayo. Computational protein design. *Structure*, 7:R105–R109, 1999.

[36] L. Tucker-Kellogg. *Systematic Conformational Search with Constraint Satisfaction*. PhD thesis, Massachusetts Institute of Technology, 2002.

[37] S. Weiner, P. Kollman, D. Case, U. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.

APPENDIX

In Appendix A, four extensions to the *minimized-DEE* criterion are presented along with the corresponding extensions to the *traditional-DEE* criterion. Appendix B provides details on energy precomputation for computing the lower energy bounds $B(\cdot)$.

A Extensions to DEE with Energy Minimization

An excellent review of the advanced pruning techniques used in the extensions to the traditional-DEE pruning conditions appears in [26]. These methods allow more individual rotamers to be pruned during DEE and extend the DEE criterion to identify dead-ending rotamer pairs. For example, the proper use of Dead-Ending Pairs [15] (Fig. 4i) allows additional rotamers to be identified as being dead-ending and thus not part of the GMEC. Analogously to Sec 2.2, we have derived *minimized-DEE equivalents* to Dead-Ending Pairs (Fig. 4j) and for 3 other advanced pruning techniques from traditional-DEE; see Fig. 4 for a summary.

B Energy Precomputation for Lower Bounds

We first derive a lower bound for the energy of a minimized conformation, similarly to [17, 18]. We then present improvements on the energy precomputation algorithm, as compared to [17, 18].

B.1 Computing a Lower Bound on Minimized Energies

In our structural model, (Sec. 5), 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

$$E_S = A_{00} + \sum_{i \leq h} A_{0i} + \sum_{i \leq h} A_{i0} + \sum_{i \leq h} \sum_{i < j \leq h} A_{ij}. \tag{33}$$

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

$$E_{\min} = D_{00} + \sum_{i \leq h} D_{0i} + \sum_{i \leq h} D_{i0} + \sum_{i \leq h} \sum_{i < j \leq h} D_{ij}. \tag{34}$$

Traditional-DEE	
(a) $E(i_r) - E(i_t) + \sum_{j \neq i} \min_s E(i_r, j_s) - \sum_{j \neq i} \max_s E(i_t, j_s) > 0$	[6]
(c) $E(i_r) - E(i_t) + \sum_{j,j \neq i} \min_s \left(E(i_r, j_s) - E(i_t, j_s)\right) > 0$	[9]
(e) $E(i_r) - \sum_{x=1,T} C_x E(i_{t_x}) + \sum_{j,j \neq i} \min_s \left(E(i_r, j_s) - \sum_{x=1,T} C_x E(i_{t_x}, j_s) \right) > 0$	[9]
(g) $E(i_r) - E(i_t) + \sum_{j,\ j \neq h \neq i} \left(\min_s \left(E(i_r, j_s) - E(i_t, j_s)\right) \right) + \left(E(i_r, h_v) - E(i_t, h_v)\right) > 0$	[26]
(i) $E([i_r j_s]) - E([i_u j_v]) + \sum_{h \neq i,j} \min_t E([i_r j_s], h_t) - \sum_{h \neq i,j} \min_t E([i_u j_v], h_t) > 0$	[6, 15]
Minimized-DEE	
(b) $E_\ominus(i_r) - E_\oplus(i_t) + \sum_{j \neq i} \min_s E_\ominus(i_r, j_s) - \sum_{j \neq i} \max_s E_\oplus(i_t, j_s)$ $- \sum_{j \neq i} \max_s E_\oslash(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_\oslash(j_s, k_u) > 0$	
(d) $E_\ominus(i_r) - E_\oplus(i_t) - \sum_j \max_s E_\oslash(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_\oslash(j_s, k_u)$ $+ \sum_{j \neq i} \min_s \left(E_\ominus(i_r, j_s) - E_\oplus(i_t, j_s)\right) > 0$	
(f) $E_\ominus(i_r) - \sum_{x=1,T} C_x E_\oplus(i_{t_x}) - \sum_{j \neq i} \max_s E_\oslash(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > i} \max_{s,u} E_\oslash(j_s, k_u)$ $+ \sum_{j \neq i} \min_s \left(E_\ominus(i_r, j_s) - \sum_{x=1,T} C_x E_\oplus(i_{t_x}, j_s) \right) > 0$	
(h) $E_\ominus(i_r) - E_\oplus(i_t) - \sum_{j \neq i} \max_s E_\oslash(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_\oslash(j_s, k_u)$ $+ \sum_{j < k,\ j \neq i,h} \left(\min_s \left(E_\ominus(i_r, j_s) - E_\oplus(i_t, j_s)\right) \right) + \left(E_\ominus(i_r, h_v) - E_\oplus(i_t, h_v)\right) > 0;$	
(j) $E_\ominus([i_r j_s]) - E_\oplus([i_u j_v]) + \sum_{h \neq i,j} \min_t E_\ominus([i_r j_s], h_t) - \sum_{h \neq i,j} \max_t E_\oplus([i_u j_v], h_t)$ $- \sum_{h \neq i,j} \max_t E_\oslash(h_t) - \sum_{h \neq i,j} \sum_{k \neq i,j,k > h} \max_{t,w} E_\oslash(h_t, k_w) > 0$	

Fig. 4. Dead-End Elimination Pruning Conditions. A summary of the previously described traditional-DEE pruning conditions (top) and our newly derived minimized-DEE pruning conditions (bottom). (a) is the initial criterion for traditional-DEE [6], and (b) is the generalization for minimized-DEE (Eq. 10). The *simple* (d) and *general coupled* (f) minimized-DEE pruning conditions are analogous (resp.) to the corresponding *Goldstein* pruning conditions (c, e) of traditional-DEE [9]. General Goldstein (e), in traditional-DEE, compares the energy of i_r to a weighted average of the interaction energies among T candidate pruning rotamers i_{t_x}. $C_x \geq 0$ is the weight given to the energy computed using rotamer i_{t_x}. The traditional conformational splitting criterion [26] and the analogous MinDEE condition are given in (g) and (h), respectively. In the minimized-DEE generalization (j) of traditional Dead-Ending Pairs (i), $E_\Diamond([i_r j_s]) = E_\Diamond(i_r) + E_\Diamond(j_s) + E_\Diamond(i_r, j_s)$ $(i \neq j)$, $E_\Diamond([i_r j_s], h_t) = E_\Diamond(i_r, h_t) + E_\Diamond(j_s, h_t)$ $(i, j \neq h)$ where $E_\Diamond \in \{ E_\ominus, E_\oplus \}$.

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 K^* is indexed over all residues *and* over all rotamers for each residue, since the same rotamer can be in several different conformations, depending on the type of the neighboring residues (see Sec. 2.2). 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* [36, 29], so that one rotamer will not minimize into another. We use a voxel of $\pm 9^\circ$ for each χ angle.

B.2 Application of the Pairwise Energy Matrix

Energy precomputation is employed both for pruning with MinDEE (Sec. 2.2) and for the ε-approximation algorithms (Secs. 4.1 and 4.2). The MinDEE criterion (Eq. 10) uses both the lower- and the upper-bound (Appendix B.3) 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. (10) 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. (34). 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.

B.3 Algorithm Improvements

Analogously to the definition of matrix D in Appendix B.1, 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:

$$E_{\max} = F_{00} + \sum_{i \le h} F_{0i} + \sum_{i \le h} F_{i0} + \sum_{i \le h} \sum_{i < j \le h} F_{ij}. \tag{35}$$

Analogously to the definition of V (see Appendix B.1), 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 K^* algorithm [17, 18] 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, $l_m \leq l_s$. 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. 34) may actually result from incompatible starting points. Moreover, using multiple starting points for *full* energy-minimization is computationally infeasible (see Appendix B.2). 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 (Secs. 4.1 and 4.2) 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 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 [37, 5], 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.