Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Algorithms for protein design

Pablo Gainza[1], Hunter M Nisonoff[1] and Bruce R Donald[1,2,3]

Computational structure-based protein design programs are becoming an increasingly important tool in molecular biology. These programs compute protein sequences that are predicted to fold to a target structure and perform a desired function. The success of a program's predictions largely relies on two components: first, the input biophysical model, and second, the algorithm that computes the best sequence(s) and structure(s) according to the biophysical model. Improving both the model and the algorithm in tandem is essential to improving the success rate of current programs, and here we review recent developments in algorithms for protein design, emphasizing how novel algorithms enable the use of more accurate biophysical models. We conclude with a list of algorithmic challenges in computational protein design that we believe will be especially important for the design of therapeutic proteins and protein assemblies.

Addresses
[1] Department of Computer Science, Duke University, Durham, NC, United States
[2] Department of Biochemistry, Duke University Medical Center, Durham, NC, United States
[3] Department of Chemistry, Duke University, Durham, NC, United States

Corresponding author: Donald, Bruce R (brd+cosb15@cs.duke.edu)

Current Opinion in Structural Biology 2016, **39**:16–26

This review comes from a themed issue on **Engineering and design**

Edited by **Dan Tawfik** and **Raghavan Varadarajan**

http://dx.doi.org/10.1016/j.sbi.2016.03.006

Introduction

Computational structure-based protein design is one of the most promising tools for engineering proteins with new functions, including the development of therapeutic proteins and protein assemblies [1–4]. Despite important successes, however, many of the current computational protein design tools often have low success rates, and designed proteins sometimes fail to achieve the functional properties of native proteins. New advances in protein design methodologies are required to improve the functional properties and success rate of computationally designed proteins.

The problem of engineering a new functional protein using computational methods is typically divided in two challenging stages. The first stage is selecting a *target* tertiary/quaternary protein *fold* that will be *designed* for a specific function. Often the selected fold is one that performs a similar function and can later be redesigned to a new one [5–8,9••]. In other cases, a protein that has a completely different function is used as a *scaffold* and repurposed for a new one [10–12]. And, increasingly, protein engineers are incorporating empirical folding and structural principles [13–18] to design proteins from scratch (*de novo* design) [12–19]. The second stage is to design a protein sequence, together with side chain rotamers and residue conformations [20•], that will adopt the overall target fold (often allowing some *backbone flexibility* [13–18,20•,21,22•,23•,24–26,27••]) and perform a desired function (e.g., binding with specificity to a target molecule). This latter stage has been historically referred to as *protein design* [28]. Many computational protein engineering protocols implement different variations of these two stages, and these have resulted in many successfully engineered new proteins [5–8,9••,10–19,29–35]. Here we focus on protein design.

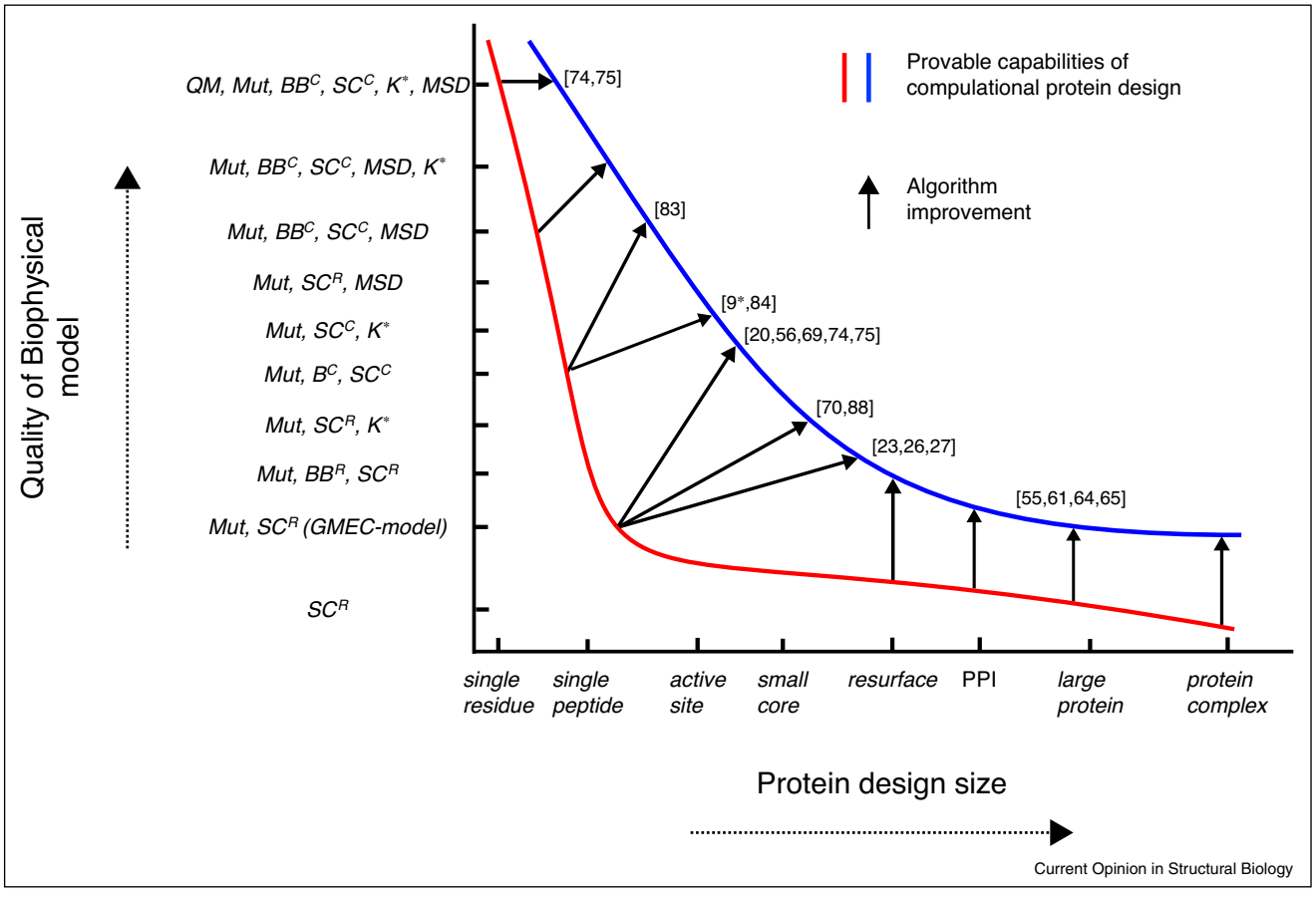
Protein design can be formulated as a well-defined computational problem by reducing it to an optimization over a family of parameterized structure-based protein redesign problems. In this well-posed version, an optimization *algorithm* (also known as a search algorithm) computes and outputs the best protein amino acid sequence(s) and structure(s) in a space defined by a *biophysical input model*. This biophysical model defines the sequence and structural search space (e.g., template input structure, the allowed flexibility, the amino acid sequences allowed, etc.), the optimization objective (e.g., single state, multi-state, ensemble-based, etc.), and the scoring potential for protein energetics (i.e., the energy function [36,37]). To our knowledge, all structure-based protein design programs conform to this formulation [13,38–42]. For example, one of the most frequently used biophysical models for backbone flexibility in Rosetta [13] consists of a target structure, an ensemble of allowed backbone moves (e.g., backbone dihedral changes), a rotamer library, the energy function, and a predefined sequence space [13,43]. This biophysical model describes a space which is then searched using Rosetta's *iterative relaxation/design* algorithm [13]. Iterative relaxation/design iteratively intercalates two steps: first, a design step, where the backbone is held constant while the conformations and amino acid identities of the side chains are optimized; and second, a relaxation step, where the sequence is held constant, while the backbone and side chains are optimized using a hybrid stochastic/gradient descent optimization [13,44,45].

The accuracy of a computational protein design relies largely on the biophysical model and it is thus highly desirable to improve this model. Biophysical model improvements, however, often come at the cost of exponentially increasing the computational complexity of the search problem. Since computational hardware cannot grow at the same rate, the only practical solution to search more complex biophysical models is through novel algorithms. Therefore, substantial improvement in computational protein design necessitates the development of novel algorithms (see Figure 1). For this reason, we focus on algorithms for protein design, and review those that we believe represent new algorithmic breakthroughs and that have potential for the design of therapeutic proteins and protein assemblies. We focus on developments since 2010 (foundational and earlier algorithms are discussed in [37,46,47•]) in four areas: optimization algorithms for protein design, algorithms to search improved flexibility models, multi-state design, and ensemble-based design. Because of constraints on the length of this survey, we exclude related algorithms that are important for therapeutic and assembly protein design that have also been highly productive recently, such as docking algorithms (for a review see [48]), scaffold search algorithms (e.g., [49,50]), and algorithms to optimize libraries for *in vitro* evolution of designed proteins (e.g., [51,52]).

Provable versus heuristic algorithms

Protein design, like many other problems in the field of computational structural biology, belongs to a hard class of computational problems [47•]. Consider, for example, a simple yet common biophysical model for the protein

Figure 1



Cartoon of the equicomplexity curves for computational protein design. Curves show a trade-off between biophysical model quality and protein design size given fixed computational resources (time and space). Algorithm improvements (black arrows) expand the boundaries of these trade-offs to allow higher quality biophysical models for larger protein design sizes. The *y*-axis maps the multi-dimensional input biophysical models that can be potentially used in protein design to a one-dimensional axis. The axis is ordered from the simplest models at the bottom (with a pairwise molecular-mechanics energy function) to the most advanced models at the top. Several examples of input biophysical models are shown: *Mut*: All amino acid mutations allowed at all designed residue positions; SC^R: discrete side-chain flexibility; BB^R: discrete backbone flexibility; SC^C: continuous side-chain flexibility; BB^C: continuous backbone flexibility; *MSD*: multi-state design; *K**: ensemble-based free energy calculations; *QM*: the most advanced energy function models. *PPI*: Protein–protein interaction. Example references of algorithms that improve the curves and that are cited in this review are shown next to each arrow. *Ref [9••] corresponds to the design of a peptide inhibitor of a PPI in the *x*-axis.

design problem, which we call the global minimum energy conformation model (GMEC-model). In this model the backbone of the protein is not allowed to move, the interaction energies between atoms are pairwise additive, the amino acid side chains are allowed to change only between discrete states (known as rotational isomers, or *rotamers*) [37,46,47•], and the goal is to compute the GMEC and its corresponding amino acid sequence. It has been shown that, even under this simple GMEC-model, the protein design problem belongs to a hard class of computational problems known as NP-hard (reviewed in [47•]), which implies that an efficient algorithm for all instances of the problem is unlikely. Moreover, other useful models (discussed in 'Algorithms for multistate protein design' and 'Algorithms to model partition functions and free energy of binding in design' sections) can belong to even more challenging classes of computational problems.

Because of the complexity inherent to these hard classes of problems, many protein design programs use *heuristic* optimization algorithms to compute low-energy sequences in the search space described by the biophysical model [46]. Heuristic algorithms use stochastic processes to search the space described by the model and are usually fast. Their speed makes them attractive because protein designers can expand the biophysical model (i.e., expand backbone flexibility, rotamer libraries, the number of protein sequences considered, or the number of states in a multistate design), without incurring a significant performance penalty. For these reasons, the field of heuristic algorithms has been highly prolific both before 2010 [46] and since 2010 (reviewed here).

However, heuristic algorithms cannot guarantee the optimality of the solution relative to the input biophysical model. This can be a particular disadvantage when models are expanded: a heuristic algorithm may not be able to accurately search the expanded model [53••]. Provable protein design algorithms, in contrast, guarantee that if the algorithm runs to completion, the computed sequences are the best ones, or are provably close to the best one, as defined by the model. Although provable algorithms can be empirically slower than heuristic algorithms, their guarantees present several advantages.

First, provable algorithms can (and often do) compute lower energy sequences than heuristic algorithms. For example, in a recent study [53••], Simoncini *et al.* compared their provable optimization algorithm implemented in the Toulbar2 program [54,55] versus the heuristic simulated annealing (SA) algorithm implemented in the Rosetta program using a GMEC-model. They found that, in a set of 100 test protein designs, SA often fails to compute the optimal answer even after running hundreds of times. In another study, Donald and co-workers compared the performance of a commonly used heuristic to model the continuous low-energy regions around rotamers, against provably searching the space [56•]. The heuristic method consisted of an initial discrete search followed by *post hoc* minimization. They found that the provable answer was far lower in energy and more similar to native proteins than the heuristically computed sequences.

When a protein designer defines a biophysical model for a specific, empirical protein design problem she usually cannot predict the computational complexity of the model (although algorithms such as [57] can upper bound the complexity, but these are exceptions to the rule). Since protein design search spaces quickly grow exponentially along multiple dimensions (i.e., sequence space, side chain conformation space, backbone conformation space), the input biophysical model might be too complex for any method to optimize accurately, heuristic or provable. Therefore, a second, and very important advantage of provable algorithms, is that they provide a signaling mechanism on the complexity of the input biophysical model. The importance of this signaling mechanism was evinced in the work by Simoncini *et al.* [53••], where they found a clear tendency for the difference between the SA solution and the GMEC computed by the provable algorithm to increase as the size of the problem grows. If a heuristic algorithm, such as SA, is used to optimize a biophysical model that exceeds the capabilities of provable algorithms, then it is impossible to know the size of the difference, and the designer has no information on the quality of the heuristic search. In contrast, provable algorithms in these cases warn the designer of the need to either first, restrict the biophysical model or second, improve the algorithms.

Finally, when the output of provable algorithms fails to perform experimentally as predicted, the error can be isolated and attributed solely to the biophysical model, and the biophysical model can then be improved. With heuristic algorithms, it is difficult to know whether the design process failed because of the algorithm or the biophysical model. An interesting example of the ambiguity introduced by heuristic algorithms was recently reported in a closely related field within computational structural biology, in Nuclear Magnetic Resonance (NMR) structure determination [58••]. With sparse data (e.g., for membrane proteins) this problem is a form of protein structure prediction, and has similarities to backbone conformation search and scoring in protein design. Martin *et al.* showed that a recently published structure of Diacylglycerol Kinase, solved by NMR was vastly different from the crystal structure likely because the SA optimization protocol used to interpret the NMR data failed to find up to 18 other low-energy structures with completely different topologies that also fit the data. Although this discrepancy occurred in the field of NMR structure determination, it is a warning about

the ability of SA to explore backbone conformational space, and we suspect that similar dangers can occur when using SA or other heuristic algorithms in computational protein design.

Thus, there are trade-offs to using provable versus heuristic algorithms: heuristic algorithms can more quickly search complex biophysical models yet cannot guarantee accuracy, while provable algorithms guarantee accuracy but are typically slower. Since 2010 there have been considerable developments in algorithms for both categories and we review these in the following sections.

Progress in optimization algorithms for the GMEC-model

The GMEC-model is one of the most used protein design models and it is often employed as a benchmark to test the performance of new protein design optimization algorithms, provable or heuristic. Most of the heuristic algorithms that are currently used for the protein design GMEC-model, however, were developed before 2010 (and are surveyed in [37,46]). One recent, promising approach, however, is the CLEVER algorithm for protein design, which builds on previous cluster expansion algorithms developed in Amy Keating's laboratory [59,60]. Cluster expansion for protein design is a technique that maps the complex three-dimensional atomic-level energy function, which is a function of atomic coordinates, to a simpler linear function dependent only on the sequence [60]. Thus, cluster expansion maps the biophysical input model to a much simpler one. Then, integer linear programming solvers can be used to efficiently find the optimal sequence in the new model. Since this mapping is only an approximation, the error resulting from a cluster expansion can potentially be large. In more recent work, Hahn *et al.* analyzed the sources of error in cluster expansion and developed a series of techniques to reduce them [59]. Although it is likely very difficult to eliminate all approximation errors in cluster expansion, its speed makes it an attractive approach for protein design.

In contrast to progress in heuristic algorithms, several notable advances in provable protein design optimization have been reported over the past few years. Tommi Jaakkola and co-workers developed the max-product linear programming algorithm (MPLP) for computer vision and protein design applications [61]. MPLP uses a message passing-based [47•], block-coordinate descent algorithm to approximate tight lower bounds on the energy of partial protein design solutions [61,62•]. Then, using information from the bounding process, the algorithm groups residues into clusters, which iteratively results in tighter MPLP bounds until the exact, best solution is computed [61]. Other algorithms have adopted bounding techniques, similar to MPLP, and used them in different branch-and-bound frameworks. The BroMAP algorithm [63] also uses a message passing algorithm as the bounding technique, followed by a branch-and-bound algorithm. At each step in the branch-and-bound process, BroMAP prunes solutions that it can prove will not lead to the optimal solution. Donald and co-workers developed the Dynamic A^* algorithm [64•], which is based on the traditional A^* algorithm for protein design (reviewed in [37,46]). Dynamic A^* radically improves the performance of A^* through both better bounding and by introducing dynamic residue ordering to the design process. Zeng and co-workers [65] developed an AND/OR branch-and-bound method that exploits the sparse nature of protein design biophysical models. In an AND/OR tree the protein design optimization problem can be split into different components on-the-fly, and each subtree can be solved independently, which can result in a significant performance improvement. Other recent developments use advanced computer science tools, such as dynamic programming [57,66•] or solvers for the Boolean satisfiability problem [67].

Promising improvements in the performance of protein design optimization algorithms for the GMEC-model were recently reported by Thomas Schiex and co-workers [53••,54,55]. Their protein design optimization algorithm exploits weighted constraint satisfaction (WCSP) techniques, including fast soft local consistencies for bounding, an advanced branch and bound implementation, and sophisticated ordering techniques, to compute the GMEC significantly faster than competing approaches [54]. These algorithms are implemented in the Toulbar2 program [53••,54,55], with support only for discrete flexibility models, and in the Osprey protein design program [39,68] (with full support for continuous flexibility models, described in the next section [64•,69••]).

Although the algorithms described in this section were tested on the discrete rotamer GMEC-model they often are integral components of other algorithms that are used to search more complex biophysical models [56•,70]. Often, the improvements in optimization power described in this Section are necessary to allow for search over these more complex models in a reasonable amount of time.

Algorithms to search improved models of protein flexibility

Protein design energy functions can be sensitive to small changes in atom coordinates. A protein sequence that is predicted to be energetically disfavorable under a rigid backbone biophysical model, for example, could be very favorable with some slight adjustments to the backbone or side chain angles. Moreover, certain protein secondary structure elements, such as backbone loops, can change their conformations after introducing mutations, and protein design algorithms must model these new conformations to find the lowest energy conformation of a new sequence. Thus, protein design algorithms can improve

the quality of their design predictions by expanding/improving their flexibility models.

A widely used method for improved flexibility is the iterative relaxation/design method in Rosetta [13]. This method intercalates sequence design steps (which hold constant the backbone coordinates) with backbone relaxation steps (which hold constant the amino acid sequence [13]). The relaxation step was recently improved by Tyka *et al.* in the FastRelax [45] and BatchRelax [71] algorithms, which improve the speed of these methods by annealing the weight of the van der Waals term during the minimization process.

Separating backbone flexibility and sequence design into different steps may not always find the best sequences under the flexibility model because the best sequence might only be reached after a concerted backbone movement and residue mutation. Thus, designers have attempted to search backbone and sequence space simultaneously [20•,21,22•].

The importance of searching backbone flexibility and sequence space simultaneously has been validated retrospectively in biologically relevant computational experiments. For example, Kortemme and co-workers [21] showed that incorporating backbone flexibility greatly improved the ability to predict mutational tolerance in HIV-1 protease and reverse transcriptase. In addition, in [22•] they showed that coupling backbone movements with side-chain movements improved the ability to predict mutations in enzymes and to predict sequences for a protein to bind a particular ligand.

Searching backbone flexibility and sequence space simultaneously represents an important algorithmic challenge for protein design algorithms along two areas: (i) algorithms are needed to improve the flexibility model by defining a restricted, yet comprehensive, conformational search space, and (ii) algorithms are needed to search the improved, yet often more complex, flexibility models. In the first area the main challenge lies in improved models of backbone flexibility. The size of the conformation space with backbone flexibility can be huge and it would be useful to have methods to restrict the space to subsets of good backbone conformations. This problem is particularly hard for protein loops because the space of possible loop conformations is usually large. To address this problem, Kortemme and co-workers developed the KIC algorithm [23•,24], which generates a library of backbone loops for both protein design and protein modeling. KIC uses a kinematic closure algorithm to analytically solve all possible solutions to a set of three 'pivot' residues and uses a Ramachandran plot to compute a library of additional torsional values for the remaining torsional angles in a loop of arbitrary size. Notably, KIC has been compared against loop backbones generated using molecular dynamics and shown to be more accurate [25]. The Baker laboratory incorporated KIC and several other backbone flexibility techniques into RosettaRemodel, a general framework for backbone flexibility in Rosetta [72]. Tripathy *et al.* [26] developed a method that uses NMR orientational restraints to generate ensembles of loop backbones. Their method, POOL, exploits sparse NMR orientational restraints and kinematic restrictions on the backbone to systematically search all loop conformations that satisfy the data. In cases where some NMR data on a loop of the wildtype structure is available, POOL can be used to accurately generate candidate loop conformations, and these can be used as input for the design. Floudas and co-workers [27••] provide an algorithm to predict loop structures by iteratively refining the bounds on the dihedral angles of the loop residues. Importantly, the algorithm is able to handle cases in which the coordinates of the flanking secondary structure elements are not known.

Once these discrete backbones are computed, then they can be incorporated into the biophysical model, and any algorithm to compute the GMEC can be used to compute the optimal conformation for the improved model. In the case of cluster expansion approximation where the three dimensional coordinates of atoms are mapped to a simple linear function dependent only on sequence, however, it is not straightforward to see that the approximation will be robust enough for backbone flexibility. To investigate this, Agpar *et al.* [73] recently evaluated the performance of cluster expansion with a backbone flexibility model and demonstrated that the approximation is robust enough for the improved model.

Discrete models of flexibility, however, are limited because protein energetics can be very sensitive to small changes. With discrete models it becomes hard to know beforehand how much discrete sampling is necessary for a specific design. Rather than model conformations as discrete conformations, Donald and co-workers have developed a suite of provable algorithms [20•,56•,69••,74••] that incorporate continuous flexibility to model the low-energy torsional regions around discrete conformations. They have found that modeling the continuous flexibility around discrete, low-energy conformations results in designed protein sequences that are much more similar to native protein sequences [56•]. These algorithms to model continuous flexibility [20•,56•,69••,74••] build on the A^* protein design algorithm [64•] to compute bounds on the energies of the continuous regions around discrete conformations, and guarantee to find the lowest energy conformations and sequences under the continuous flexibility model. The iMinDEE algorithm [56•] introduces a new, powerful provable technique to remove continuous rotamers in a protein design search that can be shown to not be part of the optimal solution. The DEEPer algorithm [20•] models simultaneous continuous flexibility in both backbone and side-chain movements. The EPIC

algorithm [74••] reduces the burden of using continuous flexibility in protein design by precomputing energy functions between pairs of residues as compact, quick-to-evaluate polynomials. The HOT and PartCR algorithms further accelerate the speed of continuous flexibility algorithms by merging/or partitioning discrete conformations, improving the speed of continuous flexibility algorithms 100-fold [69••]. The LUTE algorithm [75], conceptually similar to cluster expansion, performs a black-box reduction of protein design with continuous flexibility, to (two calls to) a discrete optimization algorithm for protein design (such as those described in 'Progress in optimization algorithms for the GMEC-model' section). LUTE also provides a similar reduction for design with non-pairwise energy functions, such as Poisson–Boltzmann solvation. By reducing these more complex optimizations to pairwise discrete optimization, LUTE can exploit many of their advantages while obtaining the speed of a discrete algorithm.

Algorithms for multistate protein design

The design of protein therapeutics and protein assemblies often requires optimizing over multiple *states*, which is known as multistate design (MSD). For example, the MSD goal might be to optimize binding affinity (which can be modeled by optimizing the difference in energy between the bound state of a complex and the unbound states of the monomers), to optimize for affinity to multiple targets, or to optimize for binding to one target while preventing binding to other targets. MSD has many practical applications. For example, Meiler and co-workers [76] recently showed that MSD resulted in better agreement with observed evolutionary sequence profiles of antibodies when compared with single state design. In another example [5,6], Anderson, Donald and co-workers used MSD to predict active site mutations in a drug target that confer resistance to a drug while maintaining the natural function of the enzyme.

In practice, MSD can be harder than traditional protein design optimization problems because multiple states must be simultaneously optimized, instead of just one, but since 2010 there have been several developments of new heuristic algorithms for MSD (earlier MSD algorithms are discussed in [77]). Building on cluster expansion, the Keating laboratory developed the CLASSY algorithm for MSD [29,78]. CLASSY uses a protocol called a 'CLASSY specificity sweep' which consists of multiple steps. In the first step, CLASSY designs a sequence for affinity to a target. On the second and subsequent rounds, new constraints are added to increase the difference in binding energy of the sequence to the target versus the off-targets. An optimization process is run anew to optimize for binding to the target. The procedure continues until a highly specific sequence is found. Fromer *et al.* [79,80] developed an algorithm for multispecificity using probabilistic graphical models [47•]. Their method combines the energy function of the multiple targets to a single energy function and models the graphical model for each target structure as a separate graphical model. Constraints between the different graphical models ensure that the sequences of all target structures are the same, and the loopy belief propagation optimization algorithm is then used to compute the optimal sequence. Allen and Mayo [81] adapted the FASTER algorithm, a heuristic optimization algorithm developed for protein design, to the multistate case. Their method, MSD-FASTER, iteratively computes sequences for each of the target states and then applies them together, until convergence. Chica and co-workers [82] developed an approach to separate backbone ensembles generated computationally into native versus non-native clusters using multiple linear regression against a training set with known experimental data. They then apply the FASTER algorithm to optimize the sequence with the largest Boltzmann-weighted average energy difference between native and non-native ensembles. Leaver-Fay *et al.* [77] developed a genetic algorithm for multistate design within a multistate program that allows the user to define custom multistate optimization objectives. Finally, Meiler and co-workers developed the RECON algorithm [76], which allows the multiple states to separately explore the sequence space, while slowly converging, under the assumption that this should simplify the energy landscape of the sequence space.

The problem of computing the MSD belongs to a hard class of computational problems (we conjecture that its class is harder than computing the GMEC). In practice MSD is computationally challenging because the number of sequences is exponential in the number of mutable residue positions, and the optimization objective considers multiple states for each sequence. To our knowledge, the Comets algorithm developed in the Donald laboratory [83••] is the only provable algorithm for multistate design with an all-atom energy functions that does not exhaustively enumerate the sequence space. Comets uses a search tree based on the A^* algorithm over sequences, where each internal node encodes a partially defined sequence, and each leaf encodes a fully defined sequence. Comets computes a multistate lower bound on the best energy of every internal node, and at each step the internal node with the best bound is expanded. This process continues until the optimal MSD sequences are computed.

Algorithms to model partition functions and free energy of binding in design

Proteins exist in solution as thermodynamic ensembles, and the binding affinity of two proteins depends on the free energy of these ensembles [9••,47•,84]. In particular, when two proteins bind each other, such as when a therapeutic protein binds its target, the binding affinity

is often affected by an associated loss in conformational entropy. Thus, protein design protocols must account for the entropy change [9••], and failure to do so can result in weakly binding proteins. This was validated recently in a computational study by Fleishman *et al.* [85], which compared the conformational flexibility of the side chains in the binding interface of computationally designed proteins, compared with that of native proteins. Their computational study showed that native binding proteins had restrictions on the side-chain plasticity of proteins, ensuring they undergo a small entropic loss upon binding. Roberts *et al.* [9••] also performed a comparison between the computational predictions of a GMEC-model versus the predictions of an ensemble-based model, in an experimentally validated design of peptide inhibitors for cystic fibrosis. They found that of the top 30 sequences designed by a single-structure (GMEC) model versus an ensemble model, the predictions by the two models shared only 4 sequences in common. The top experimentally validated sequences were top-ranked by the ensemble-based model but were not top-ranked sequences in the single-structure model [9••].

Computing protein ensembles, and the associated partition function for a protein, belongs to a difficult class of computer science problems, #*P*-hard. However, several approximation algorithms have been developed to compute rotamer-based partition functions that can be used to estimate conformational entropy upon binding. Fleishman *et al.* [85] developed a method similar to self consistent mean field to compute a penalty for rotamers that change conformation upon binding. Kamisetty *et al.* [86•] used probabilistic graphical models to formulate the partition function computation as an inference problem. Then, they used the GOBLIN algorithm, which is based on loopy belief propagation, to compute the free energy of binding in protein–protein interactions [86•]. Their method was also used to show that explicitly accounting for conformational entropy can significantly improve predictions of binding affinities. Gevorg Grigoryan developed the molecular-dynamics based Valocidy algorithm [87••], which uses ensembles to compute free energy. Grigoryan showed how to use Valocidy to calculate free-energy estimates in protein design by incorporating it into cluster expansion.

Despite the hardness of computing partition functions, some algorithms can approximate the free energy of binding with guarantees on the input for continuous rotameric or residue conformation [20•] models. The *K** algorithm [9••] computes approximations to the binding constant of two proteins by computing the ratio of the partition functions between bound and unbound states [47•]. To approximate the partition function of each state, *K** enumerates conformations in order of energy, using the *A** algorithm and stops once a provable, ε-approximation to each partition function has been computed. Jou *et al.* developed the BWM* algorithm, based on dynamic programming techniques, to rapidly enumerate conformational ensembles for use in partition function computation [57]. Viricel *et al.* adapted weighted constraint satisfaction techniques to compute ε-approximations to the partition function and *K** score more quickly than previous methods [70]. Silver *et al.* [88••] developed a method to compute configurational entropy using *A**-enumerated ensembles, similar to the *K** algorithm [84], followed by a novel entropy expansion. The method represents the entropy as a series of mutually exclusive terms, each corresponding to the marginal entropy of a particular degree of freedom or to coupling between degrees of freedom. This expansion offers a clear interpretation of the contribution of particular degrees of freedom to the total entropy of the molecule [88••].

Conclusions and important algorithmic challenges in protein design

The accuracy of structure-based computational protein design depends on the quality of the input biophysical model. Improving this model in turn requires new algorithms capable of searching the more complex search spaces created by the improved model. The perspicacious reader will note, however, that an arbitrary improvement to the biophysical model, no matter how beneficial in principle, may not admit a corresponding algorithmic improvement to re-expand the equicomplexity curve (Figure 1). The most significant improvements to algorithms for protein design are likely to come from algorithms that exploit the structure of new features in the biophysical model. Yet not all features will be equally exploitable by the algorithms. For example, modeling the solvent molecules in a protein's environment explicitly would most likely improve the accuracy of protein design, yet it is challenging to develop efficient design algorithms that can search these solvent models. Thus, there is an impedance matching between algorithms and models, and therefore both algorithms and models must be improved in tandem.

Since 2010 there have been important improvements to many important algorithmic challenges in protein design. There are many areas where protein design algorithms can be further improved, both within and outside those listed here. First, as shown by [53••] and argued in this review, significant progress in the speed of provable optimization algorithms is not only possible, but it is essential to ensure the quality of the designed sequences with respect to the input biophysical model. Developing algorithms that improve over the best state-of-the-art algorithms will enable enhanced modeling and, in consequence, better designs. In addition, as shown in multiple studies [9••, 20•,21,25,47•,56•] improving flexibility models in protein design can result in sequences that are lower energy, are more similar to those of native proteins, or perform better experimentally [9••]. Developing new algorithms that expand the equicomplexity curve (Figure 1) for expanded flexibility models, especially continuous flexibility, is likely to result in better protein designs.

The successful engineering of proteins for new function depends on a two stage strategy: selecting a tertiary or quaternary protein fold that can be designed for a specific function, followed by protein design of the fold. Recently, as the functional goals of protein design have become increasingly ambitious, the first stage (selecting the fold) has become more relevant because most protein folds are likely not designable for a specific function. Designers often utilize scaffold search algorithms or increasingly more commonly, *de novo* strategies to find folds, followed sequentially by protein design to compute a sequence for these folds. This sequential strategy, known as a *greedy* strategy in computer science, suffers from an important limitation from decoupling the two stages: because the stages are sequential, the fold selection process may not be able to foresee whether a protein sequence will be able to adopt the fold and perform a function (e.g., binding). Thus, filtering criteria at the fold selection stage will remove solutions that could be successfully designed to a functioning protein. The engineering of proteins would greatly benefit from a tighter coupling between fold search and protein design. Several recent strategies address in part this limitation [15,89] from a structural perspective: they predict whether the fold is designable based on observed protein structures. Others, such as inverse rotamers [90], address this from a functional (i.e., binding or catalysis) perspective. However, a tighter integration between fold selection and protein design is needed to improve the functional capabilities of computational protein engineering, and such integration will likely require new algorithms. These and other advances, coupled with state-of-the-art techniques for *in vitro*, *in vivo* and preclinical validation [91], could transform computational protein design into an essential tool for the development of novel therapeutics.

Conflict of interest statement

Nothing declared.

Acknowledgements

We sincerely apologize to the many authors of outstanding work that was overlooked or not included in this review because of space limitations. We thank Marcel Frankel, Anna Lowegard, and Jonathan Jou for comments on the manuscript, as well as Mark A Hallen for helpful discussions. This work is supported by the following grant from the National Institutes of Health: R01 GM-78031 to BRD.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest

•• of outstanding interest

1. Schreiber G, Fleishman SJ: **Computational design of protein–protein interactions**. *Curr Opin Struct Biol* 2013, **23**:903-910.

2. Khare SD, Fleishman SJ: **Emerging themes in the computational design of novel enzymes and protein:protein interfaces**. *FEBS Lett* 2013, **587**:1147-1154.

3. Der BS, Kuhlman B: **Strategies to control the binding mode of *de novo* designed protein interactions**. *Curr Opin Struct Biol* 2013, **23**:639-646.

4. Zhang J, Zheng F, Grigoryan G: **Design and designability of protein-based assemblies**. *Curr Opin Struct Biol* 2014, **27**:79-86.

5. Reeve SM, Gainza P, Frey KM, Georgiev I, Donald BR, Anderson AC: **Protein design algorithms predict viable resistance to an experimental antifolate**. *Proc Natl Acad Sci U S A* 2015, **112**:749-754.

6. Frey KM, Georgiev I, Donald BR, Anderson AC: **Predicting resistance mutations using protein design algorithms**. *Proc Natl Acad Sci U S A* 2010, **107**:13707-13712.

7. Chen C-Y, Georgiev I, Anderson AC, Donald BR: **Computational structure-based redesign of enzyme activity**. *Proc Natl Acad Sci U S A* 2009, **106**:3764-3769.

8. Choi Y, Hua C, Sentman CL, Ackerman ME, Bailey-Kellogg C: **Antibody humanization by structure-based computational protein design**. *mAbs* 2015, **7**:1045-1057.

9. • Roberts KE, Cushing PR, Boisguerin P, Madden DR, Donald BR: **Computational design of a PDZ domain peptide inhibitor that rescues CFTR activity**. *PLoS Comput Biol* 2012, **8**:e1002477.
The authors present an ensemble-based, continuous flexibility algorithm to approximate the binding constant of a protein–protein complex. Then, they use this algorithm to design inhibitors of the CAL protein to disrupt the interaction between CAL and the cystic fibrosis transmembrane regulator (CFTR). The inhibitors were tested experimentally and shown to both bind tightly to their target, and rescue CFTR activity in cell-based assays.

10. Azoitei ML, Correia BE, Ban Y-EA, Carrico C, Kalyuzhniy O, Chen L, Schroeter A, Huang P-S, McLellan JS, Kwong PD, Baker D, Strong RK, Schief WR: **Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold**. *Science* 2011, **334**:373-376.

11. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, Baker D: **Computational design of ligand-binding proteins with high affinity and selectivity**. *Nature* 2013, **501**:212-216.

12. Procko E, Berguig GY, Shen BW, Song Y, Frayo S, Convertine AJ, Margineantu D, Booth G, Correia BE, Cheng Y *et al.*: **A computationally designed inhibitor of an Epstein–Barr viral Bcl-2 protein induces apoptosis in infected cells**. *Cell* 2014, **157**:1644-1656.

13. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D: **Design of a novel globular protein fold with atomic-level accuracy**. *Science* 2003, **302**:1364-1368.

14. Doyle L, Hallinan J, Bolduc J, Parmeggiani F, Baker D, Stoddard BL, Bradley P: **Rational design of α-helical tandem repeat proteins with closed architectures**. *Nature* 2015, **528**:585-588.

15. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D: **Principles for designing ideal protein structures**. *Nature* 2012, **491**:222-227.

16. King IC, Gleixner J, Doyle L, Kuzin A, Hunt JF, Xiao R, Montelione GT, Stoddard BL, DiMaio F, Baker D: **Precise assembly of complex beta sheet topologies from *de novo* designed building blocks**. *eLife* 2016, **4**:e11012.

17. Lin Y-R, Koga N, Tatsumi-Koga R, Liu G, Clouser AF, Montelione GT, Baker D: **Control over overall shape and size in *de novo* designed proteins**. *Proc Natl Acad Sci U S A* 2015, **112**:E5478-E5485.

18. Joh NH, Wang T, Bhate MP, Acharya R, Wu Y, Grabe M, Hong M, Grigoryan G, DeGrado WF: ***De novo* design of a transmembrane Zn^{2+}-transporting four-helix bundle**. *Science* 2014, **346**:1520-1524.

19. Correia BE, Bates JT, Loomis RJ, Baneyx G, Carrico C, Jardine JG, Rupert P, Correnti C, Kalyuzhniy O, Vittal V, Connell MJ, Stevens E, Schroeter A, Chen M, MacPherson S, Serra AM, Adachi Y, Holmes MA, Li Y, Klevit RE, Graham BS, Wyatt RT, Baker D, Strong RK, Crowe JE, Johnson PR, Schief WR: **Proof of principle for epitope-focused vaccine design**. *Nature* 2014, **507**:201-206.

20. • Hallen MA, Keedy DA, Donald BR: **Dead end elimination with perturbations (DEEPer): a provable protein design algorithm**

with continuous sidechain and backbone flexibility. *Proteins* 2013, **81**:18-39.
The authors present a novel algorithm to handle simultaneous continuous backbone and side-chain flexibility. Additionally, an improved pruning criterion is proposed to speedup DEE/*A**-based calculations. The enhanced backbone flexibility consistently resulted in lower-energy conformations, when compared against previous methods.

21. Humphris-Narayanan E, Akiva E, Varela R, Conchir S, Kortemme T: **Prediction of mutational tolerance in HIV-1 protease and reverse transcriptase using flexible backbone protein design**. *PLoS Comput Biol* 2012, **8**:e1002639.

22. • Ollikainen N, de Jong RM, Kortemme T: **Coupling protein side-chain and backbone flexibility improves the re-design of protein–ligand specificity**. *PLoS Comput Biol* 2015, **11**:e1004335.
The authors provide a method to couple both side-chain and backbone flexibility during a stochastic search. They demonstrate the importance of modeling subtle changes that arise during the coupling through sequence recovery of an enzyme ligand binding model.

23. • Stein A, Kortemme T: **Improvements to robotics-inspired conformational sampling in Rosetta**. *PLoS ONE* 2013, **8**:e63090.
The authors build on a previous method to model flexible loops using robotics-based kinematic closure. A number of improved sampling methods are proposed, which when combined, better sample native poses with sub-angstrom accuracy

24. Mandell DJ, Coutsias EA, Kortemme T: **Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling**. *Nat Methods* 2009, **6**:551-552.

25. Babor M, Mandell DJ, Kortemme T: **Assessment of flexible backbone protein design methods for sequence library prediction in the therapeutic antibody herceptin–HER2 interface**. *Protein Sci* 2011, **20**:1082-1089.

26. Tripathy C, Zeng J, Zhou P, Donald BR: **Protein loop closure using orientational restraints from NMR data**. *Proteins* 2012, **80**:433-453.

27. •• Subramani A, Floudas CA: **Structure prediction of loops with fixed and flexible stems**. *J Phys Chem B* 2012, **116**:6670-6682.
The authors present a method to obtain loop conformations through iterative bounding the dihedral angle space of each residue. Each iteration involves dihedral angle sampling, followed by rotamer optimization and all-atom energy minimization, and a clustering step in which densely populated clusters are used to refine the bounds on the dihedral angles allowed for each amino acid. The algorithm was tested both for scenarios in which the surrounding secondary structure configurations are known, and in the case in which only the *identities* of the surrounding secondary structure (α-helix or β-sheet), but not the *coordinates*, are known.

28. Pabo C: **Molecular technology: Designing proteins and peptides**. *Nature* 1983, **301** 200–200.

29. Grigoryan G, Reinke AW, Keating AE: **Design of protein-interaction specificity gives selective bZIP-binding peptides**. *Nature* 2009, **458**:859-864.

30. Guntas G, Hallett RA, Zimmerman SP, Williams T, Yumerefendi H, Bear JE, Kuhlman B: **Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins**. *Proc Natl Acad Sci U S A* 2015, **112**:112-117.

31. Lewis SM, Wu X, Pustilnik A, Sereno A, Huang F, Rick HL, Guntas G, Leaver-Fay A, Smith EM, Ho C, Hansen-Estruch C, Chamberlain AK, Truhlar SM, Conner EM, Atwell S, Kuhlman B, Demarest SJ: **Generation of bispecific IgG antibodies by structure-based design of an orthogonal Fab interface**. *Nat Biotechnol* 2014, **32**:191-198.

32. Jardine J, Julien J-P, Menis S, Ota T, Kalyuzhniy O, McGuire A, Sok D, Huang P-S, MacPherson S, Jones M, Nieusma T, Mathison J, Baker D, Ward AB, Burton DR, Stamatatos L, Nemazee D, Wilson IA, Schief WR: **Rational HIV immunogen design to target specific germline B cell receptors**. *Science* 2013, **340**:711-716.

33. Do Kwon Y, Pancera M, Acharya P, Georgiev IS, Crooks ET, Gorman J, Joyce MG, Guttman M, Ma X, Narpala S, Soto C, Terry DS, Yang Y, Zhou T, Ahlsen G, Bailer RT, Chambers M, Chuang G-Y, Doria-Rose NA, Druz A, Hallen MA, Harned A, Kirys T, Louder MK, O'Dell S, Ofek G, Osawa K, Prabhakaran M, Sastry M, Stewart-Jones GBE, Stuckey J, Thomas PV, Tittley T, Williams C, Zhang B, Zhao H, Zhou Z, Donald BR, Lee LK, Zolla-Pazner S, Baxa U, Schn A, Freire E, Shapiro L, Lee KK, Arthos J, Munro JB, Blanchard SC, Mothes W, Binley JM, McDermott AB, Mascola JR, Kwong PD: **Crystal structure, conformational fixation and entry-related interactions of mature ligand-free HIV-1 env**. *Nat Struct Mol Biol* 2015, **22**:522-531.

34. Reardon PN, Sage H, Dennison SM, Martin JW, Donald BR, Alam SM, Haynes BF, Spicer LD: **Structure of an HIV-1-neutralizing antibody target, the lipid-bound gp41 envelope membrane proximal region trimer**. *Proc Natl Acad Sci U S A* 2014, **111**:1391-1396.

35. Georgiev IS, Rudicell RS, Saunders KO, Shi W, Kirys T, McKee K, O'Dell S, Chuang G-Y, Yang Z-Y, Ofek G, Connors M, Mascola JR, Nabel GJ, Kwong PD: **Antibodies VRC01 and 10E8 neutralize HIV-1 with high breadth and potency even with Ig-framework regions substantially reverted to germline**. *J Immunol* 2014, **192**:1100-1106.

36. Boas FE, Harbury PB: **Potential energy functions for protein design**. *Curr Opin Struct Biol* 2007, **17**:199-204.

37. Lippow SM, Tidor B: **Progress in computational protein design**. *Curr Opin Biotechnol* 2007, **18**:305-311.

38. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D: **Progress in modeling of protein structures and interactions**. *Science* 2005, **310**:638-642.

39. Gainza P, Roberts KE, Georgiev I, Lilien RH, Keedy DA, Chen C-Y, Reza F, Anderson AC, Richardson DC, Richardson JS, Donald BR: **OSPREY: protein design with ensembles, flexibility, and provable algorithms**. *Methods Enzymol* 2013, **523**:87-107.

40. Grigoryan G, Zhou F, Lustig SR, Ceder G, Morgan D, Keating AE: **Ultra-fast evaluation of protein energies directly from sequence**. *PLoS Comput Biol* 2006, **2**:e63.

41. Dahiyat BI, Mayo SL: ***De novo* protein design: fully automated sequence selection**. *Science* 1997, **278**:82-87.

42. Baker D: **Prediction and design of macromolecular structures and interactions**. *Philos Trans R Soc Lond B Biol Sci* 2006, **361**:459-463.

43. Liu Y, Kuhlman B: **RosettaDesign server for protein design**. *Nucleic Acids Res* 2006, **34**:W235-W238.

44. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D: **Relaxation of backbone bond geometry improves protein energy landscape modeling**. *Protein Sci* 2014, **23**:47-55.

45. Tyka MD, Keedy DA, André I, DiMaio F, Song Y, Richardson DC, Richardson JS, Baker D: **Alternate states of proteins revealed by detailed energy landscape mapping**. *J Mol Biol* 2011, **405**:607-618.

46. Samish I: **Search and sampling in structural bioinformatics**. *Struct Bioinform* 2009:207-236.

47. • Donald BR: *Algorithms in Structural Molecular Biology*. Cambridge, MA: MIT Press; 2011.
This book provides a comprehensive overview of algorithms in structural biology with many applications to protein design.

48. Park H, Lee H, Seok C: **High-resolution protein–protein docking by global optimization: recent advances and future challenges**. *Curr Opin Struct Biol* 2015, **35**:24-31.

49. Gonzalez G, Hannigan B, DeGrado WF: **A real-time all-atom structural search engine for proteins**. *PLoS Comput Biol* 2014, **10**:e1003750.

50. Zhou J, Grigoryan G: **Rapid search for tertiary fragments reveals protein sequence–structure relationships**. *Protein Sci* 2015, **24**:508-524.

51. Parker AS, Griswold KE, Bailey-Kellogg C: **Optimization of combinatorial mutagenesis**. *J Comput Biol* 2011, **18**:1743-1756.

52. Jacobs TM, Yumerefendi H, Kuhlman B, Leaver-Fay A: **SwiftLib: rapid degenerate-codon-library optimization through dynamic programming**. *Nucleic Acids Res* 2015, **43** e34-e34.

53. •• Simoncini D, Allouche D, de Givry S, Delmas C, Barbe S, Schiex T: **Guaranteed discrete energy optimization on large protein design problems**. *J Chem Theory Comput* 2015, **11**:5980-5989.
This paper demonstrates how Cost Function Network methods provably find the optimal conformation for protein design problems of unprecedented sizes for deterministic algorithms. They then use the deterministic algorithm to show that stochastic methods, such as simulated annealing in Rosetta, struggle to locate the GMEC as the search space increases. Specifically, they found that as the protein design problem size grows, the probability of Rosetta finding the optimal answer approaches zero. Up to 30% of the residues in the top predictions from stochastic methods had a different sequence identity when compared to the true GMEC.

54. Allouche D, Andre I, Barbe S, Davies J, de Givry S, Katsirelos G, O'Sullivan B, Prestwich S, Schiex T, Traore S: **Computational protein design as an optimization problem**. *Artif Intell* 2014, **212**:59-79.

55. Traore S, Allouche D, Andre I, Givry Sd, Katsirelos G, Schiex T, Barbe S: **A new framework for computational protein design through cost function network optimization**. *Bioinformatics* 2013, **29**:2129-2136.

56. • Gainza P, Roberts KE, Donald BR: **Protein design using continuous rotamers**. *PLoS Comput Biol* 2012, **8**:e1002335.
The authors introduce a more efficient method to model continuous flexibility during the protein design search using provable algorithms. They also present a large-scale study demonstrating that the continuous-rotamer model often results in different rotamer and sequence assignments that also have lower energy when compared to the rigid-rotamer model.

57. Jou JD, Jain S, Georgiev I, Donald BR: **BWM*: a novel, provable, ensemble-based dynamic programming algorithm for sparse approximations of computational protein design**. *J Comput Biol* 2015. In Press.

58. •• Martin JW, Zhou P, Donald BR: **Systematic solution to homo-oligomeric structures determined by NMR**. *Proteins* 2015, **83**:651-661.
The paper demonstrates the importance of using a systematic search of possible protein folds rather than relying on stochastic methods for protein structure determination by NMR with large numbers of ambiguous constraints.

59. Hahn S, Ashenberg O, Grigoryan G, Keating AE: **Identifying and reducing error in cluster-expansion approximations of protein energies**. *J Comput Chem* 2010, **31**:2900-2914.

60. Grigoryan G, Zhou F, Lustig SR, Ceder G, Morgan D, Keating AE: **Ultra-fast evaluation of protein energies directly from sequence**. *PLoS Comput Biol* 2006, **2**:e63.

61. Sontag D, Meltzer T, Globerson A, Jaakkola TS, Weiss Y: *Tightening LP relaxations for MAP using message passing*. 2012arXiv:1206.3288.

62. • Sontag D, Choe DK, Li Y: **Efficiently searching for frustrated cycles in MAP inference**. In *Proceedings of the Twenty-eighth Conference on Uncertainty in Artificial Intelligence (UAI-12)*. Corvallis, OR: AUAI Press; 2012, 795-804.
The dual decomposition of the Lagrangian formulation of finding the maximum a posteriori (MAP) conformation (equivalent to the protein design GMEC problem) over a Markov Random Field allows for the exploitation of fast algorithms, such as message passing, that operate locally. These algorithms can provide bounds on the optimal solution, which are useful for many protein design search algorithms such as A^*. This paper presents techniques to make provably good improvements to the constraints set during the optimization, allowing for tightened bounds for computing the MAP.

63. Hong E-J, Lippow SM, Tidor B, Lozano-Pérez T: **Rotamer optimization for protein design through map estimation and problem-size reduction**. *J Comput Chem* 2009, **30**:1923-1945.

64. • Roberts KE, Gainza P, Hallen MA, Donald BR: **Fast gap-free enumeration of conformations and sequences for protein design**. *Proteins* 2015:1859-1877.
The authors present a number of improvements to the A^* algorithm, which is the most prevalent provable algorithm in the field of protein design. The improvements involve using a dynamic ordering of the search tree and improving the bounds used to estimate the energies of partial conformations. These improvements expand the search space that is possible for deterministic algorithms.

65. Zhou Y, Wu Y, Zeng J: **Computational protein design using AND/OR branch-and-bound search**. In *Research in Computational Molecular Biology. Lecture Notes in Computer Science 9029*. Edited by Przytycka TM. Springer International Publishing; 2015:354-366.

66. • Peng J, Hosur R, Berger B, Xu J: *iTreePack: Protein complex side-chain packing by dual decomposition*. 2015arXiv:1504.05467.
This paper integrates two promising approaches, dual decomposition and tree-width decomposition, to solve the side-chain placement problem. For protein–protein complexes, where the tree-width is often prohibitively large, the authors propose a dual relaxation of the problem to generate subproblems of smaller tree-width that can be solved efficiently. The new algorithm is both more efficient and more accurate than a previous tree-width decomposition algorithm that heuristically eliminates interactions to decrease the tree-width, demonstrating the importance of using provable guarantees.

67. Ollikainen N, Sentovich E, Coelho C, Kuehlmann A, Kortemme T: **Sat-based protein design**. *IEEE/ACM International Conference on Computer-aided Design-Digest of Technical Papers, 2009. ICCAD 2009*. IEEE; 2009:128-135.

68. Traore S, Roberts KE, Allouche D, Donald BR, Andre I, Schiex T, Barbe S: **Fast search algorithms for computational protein design**. *J Comput Chem* 2015. In press.

69. •• Roberts KE, Donald BR: **Improved energy bound accuracy enhances the efficiency of continuous protein design**. *Proteins* 2015, **83**:1151-1164.
Modeling continuous rotamers with provable accuracy is difficult, due predominantly to loose lower-bounds on pairwise energies that are then used for pruning and during the search. The authors propose two methods to improve these bounds with provable guarantees, and demonstrate that the improvements greatly enhance both pruning and speed.

70. Viricel C, Simoncini D, Allouche D, Givry Sd, Barbe S, Schiex T: **Approximate counting with deterministic guarantees for affinity computation**. In *Modelling, Computation and Optimization in Information Systems and Management Sciences*. Edited by Thi HAL, Dinh TP, Nguyen NT. Springer International Publishing; 2015:165-176. Advances in Intelligent Systems and Computing 360.

71. Tyka MD, Jung K, Baker D: **Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers**. *J Comput Chem* 2012, **33**:2483-2491.

72. Huang P-S, Ban Y-EA, Richter F, Andre I, Vernon R, Schief WR, Baker D: **RosettaRemodel: a generalized framework for flexible backbone protein design**. *PLoS ONE* 2011, **6**:e24109.

73. Apgar JR, Hahn S, Grigoryan G, Keating AE: **Cluster expansion models for flexible-backbone protein energetics**. *J Comput Chem* 2009, **30**:2402-2413.

74. •• Hallen MA, Gainza P, Donald BR: **Compact representation of continuous energy surfaces for more efficient protein design**. *J Chem Theory Comput* 2015, **11**:2292-2306.
The authors propose a method to overcome the main bottleneck in continuous-flexibility design by representing the continuous energy as a local polynomial expansion, that can be precomputed. The algorithm can represent the energy surface of both molecular mechanics and quantum-mechanical energy functions and provides a substantial speed-up in continuous protein design.

75. Hallen MA, Jou JD, Donald BR: **Lute (local unpruned tuple expansion): accurate continuous flexible protein design with general energy functions for rigid-rotamer-like efficiency**. In *Proceedings of the Annual International Conference on Research in Computational Molecular Biology (RECOMB), vol 9649*. Switzerland: Springer Verlag International Publishing; 2016, .

76. Sevy AM, Jacobs TM, Crowe JE Jr, Meiler J: **Design of protein multi-specificity using an independent sequence search reduces the barrier to low energy sequences**. *PLoS Comput Biol* 2015, **11**:e1004300.

77. Leaver-Fay A, Jacak R, Stranges PB, Kuhlman B: **A generic program for multistate protein design**. *PLoS ONE* 2011, **6**:e20937.

78. Negron C, Keating AE: **Multistate protein design using CLEVER and CLASSY**. *Methods Enzymol* 2013, **523**:171-190.

79. Fromer M, Yanover C, Linial M: **Design of multispecific protein sequences using probabilistic graphical modeling**. *Proteins* 2010, **78**:530-547.

80. Fromer M, Yanover C, Harel A, Shachar O, Weiss Y, Linial M: **SPRINT: side-chain prediction inference toolbox for multistate protein design**. *Bioinformatics* 2010, **26**:2466-2467.

81. Allen BD, Mayo SL: **An efficient algorithm for multistate protein design based on FASTER**. *J Comput Chem* 2010, **31**:904-916.

82. Davey JA, Damry AM, Euler CK, Goto NK, Chica RA: **Prediction of stable globular proteins using negative design with non-native backbone ensembles**. *Structure* 2015, **23**:2011-2021.

83. •• Hallen MA, Donald BR: **Comets (constrained optimization of multistate energies by tree search): a provable and efficient algorithm to optimize binding affinity and specificity with respect to sequence**. *J Comput Biol* 2015. In Press.

The authors present a comprehensive framework and algorithm to optimize multistate design problems. The algorithm uses a tree search to prevent exhaustive search over sequences, providing efficiency and provable guarantees.

84. Georgiev I, Lilien RH, Donald BR: **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**. *J Comput Chem* 2008, **29**:1527-1542.

85. Fleishman SJ, Khare SD, Koga N, Baker D: **Restricted sidechain plasticity in the structures of native proteins and complexes**. *Protein Sci* 2011, **20**:753-757.

86. • Kamisetty H, Ramanathan A, Bailey-Kellogg C, Langmead CJ: **Accounting for conformational entropy in predicting binding free energies of protein–protein interactions**. *Proteins* 2011, **79**:444-462.

The authors employ a probabilistic graphical model representation, viewing protein–protein complexes as a Markov Random Field. They employ the well-known loopy-belief propagation algorithm, which does not have provable guarantees but instead efficiently attempts to solves the linear-programming relaxation of the problem satisfying local consistency. This variational inference technique can provide estimates on the free-energy of the protein, which the authors use to demonstrate the importance of modeling conformational entropy in protein–protein interactions.

87. •• Grigoryan G: **Absolute free energies of biomolecules from unperturbed ensembles**. *J Comput Chem* 2013, **34**:2726-2741.

The author presents a novel simulation-based technique for computing the absolute free-energy of a macromolecular state. The method is also applied to train a sequence-based energy function applicable for protein design, which requires a search over an exponential number of sequences that is prohibitively expensive for the purely simulation-based technique. The work has great importance to multistate design, where the goal often to optimize the difference in free energies between a number of states.

88. •• Silver NW, King BM, Nalam MN, Cao H, Ali A, Kiran Kumar Reddy G, Rana TM, Schiffer CA, Tidor B: **Efficient computation of small-molecule configurational binding entropy and free energy changes by ensemble enumeration**. *J Chem Theory Comput* 2013, **9**:5098-5115.

Using a DEE/*A** based method, the free-energy of multiple ligands are computed using a rigid receptor. Compare to [84]. The authors demonstrate that only a small-fraction of the conformational space is needed to be explored before accurate free-energy predictions can be made. Using these predictions and a novel decomposition of the configurational entropy into increasing, additive higher-order terms, the authors demonstrate that configurational entropy loss is an important predictor of effective binding and that the majority of the entropy can be calculated from first-order uncoupled degrees of freedom.

89. Zhang J, Grigoryan G: **Mining tertiary structural motifs for assessment of designability**. *Methods Enzymol* 2013, **523**:21.

90. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, Röthlisberger D, Baker D: **New algorithms and an in silico benchmark for computational enzyme design**. *Protein Sci* 2006, **15**:2785-2794.

91. Rudicell RS, Kwon YD, Ko S-Y, Pegu A, Louder MK, Georgiev IS, Wu X, Zhu J, Boyington JC, Chen X, Shi W, Yang Z, Doria-Rose NA, McKee K, O'Dell S, Schmidt SD, Chuang G-Y, Druz A, Soto C, Yang Y, Zhang B, Zhou T, Todd J-P, Lloyd KE, Eudailey J, Roberts KE, Donald BR, Bailer RT, Ledgerwood J, Program NCS, Mullikin JC, Shapiro L, Koup RA, Graham BS, Nason MC, Connors M, Haynes BF, Rao SS, Roederer M, Kwong PD, Mascola JR, Nabel GJ: **Enhanced potency of a broadly neutralizing HIV-1 antibody in vitro improves protection against lentiviral infection in vivo**. *J. Virol.* 2014, **88**:12669-12682 10.1128/JVI.02213-14.