How much can physics do for protein design?

Eleni Michael, Thomas Simonson

▶ To cite this version:

Eleni Michael, Thomas Simonson. How much can physics do for protein design?. Current Opinion in Structural Biology, 2022, 72, pp.46-54. 10.1016/j.sbi.2021.07.011 . hal-03663952

HAL Id: hal-03663952

https://polytechnique.hal.science/hal-03663952v1

Submitted on 16 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0959440X21001196
Manuscript_892570846e1602161b99ab10f064118d

How much can physics do for protein design?

Eleni Michael and Thomas Simonson*

Laboratoire de Biologie Structurale de la Cellule (CNRS UMR7654), Ecole Polytechnique, 91128 Palaiseau, France.

*Email: thomas.simonson@polytechnique.fr

Abstract

Physics and physical chemistry are an important thread in computational protein design, complementary to knowledge-based tools. They provide molecular mechanics scoring functions that need little or no *ad hoc* parameter re-adjustment; methods to thoroughly sample equilibrium ensembles, and different levels of approximation for conformational flexibility. They led recently to the successful redesign of a small protein using a physics-based folded-state energy. Adaptive Monte Carlo or molecular dynamics schemes were discovered where protein variants are populated according to their ligand binding free energy or catalytic efficiency. Molecular dynamics have been used for backbone flexibility. Implicit solvent models have been refined, polarizable force fields applied, and many physical insights obtained.

1 Introduction

Protein design is by nature pragmatic: success is measured by the hits obtained. The main ingredients are the energy or scoring function, the description of conformational space, the unfolded state model, and the algorithm to sample sequences and conformations. Physical chemistry and empiricism can be mixed in whatever proportion is needed. Scoring functions can emphasize molecular mechanics or purely empirical terms. The search for active variants can carefully mimic the physical behavior of a thermodynamic ensemble or simply target high scores with a heuristic search. Once predicted designs move into the experimental testing phase, experience shows that experimental firepower can rescue many physically-naive prediction models. So how much physical realism is enough, which ingredients are most critical, and what are the best ways forward?

© 2021 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

This review focusses more on new methodology than applications, and covers the period from early 2019 to early 2021. This period has produced many interesting advances. We first consider the problem of whole protein redesign. Recent work supports the possibility of a purely "physics-based", molecular mechanics energy function for the folded state. Implicit models of nonpolar solvation have been compared. Several studies explored routes towards physically realistic unfolded state descriptions, which remain tentative, however. Second, we consider methods for the design of protein-ligand complexes, including the redesign of existing enzymes for new substrates. Many aspects have been addressed within this broad area. Ligand pose selection and refinement are one important challenge. Several studies have focussed on tuning active site electric fields. One noteworthy advance is the use of adaptive landscape flattening to allow the design protocol to directly target the ligand binding free energy, including transition-state binding. This strategy is possible when sampling obeys the physically-correct, Boltzmann distribution.

In the third and fourth sections, we consider, respectively, the energy function and the problem of sampling sequences and structures. Aspects that appeared in the two previous sections are pursued further. Polarizable energy functions have been applied, not to protein design but to the closely-related problems of acid/base equilibria and side chain repacking. Implicit solvent models continue to improve, including their polar and nonpolar components. An implicit membrane model was applied to membrane protein redesign. Another element of physical realism is provided when predicted designs are post-processed with a more accurate model, especially models that use molecular dynamics with a fully flexible protein and an explicit solvent. Designs can then be rescored using a simplified free energy function (like MMPBSA) or by full-blown alchemical free energy perturbation simulations (FEP). The fourth, sampling topic includes work on ligand poses, backbone flexibility, and multistate design. An emerging theme is the use of molecular dynamics for conformational sampling *during* the exploration of sequence space, which is the most physically-realistic approach.

This review is not exhaustive, even for the period covered. Several related areas are only touched upon or mentioned. Thus, important experimental advances include high-throughput assays for whole protein design [1], protein-peptide binding [2], and stability mutations [3], as well as the new ProtoBank stability mutation database [3]. Many articles apply machine learning, like the notable Alphafold structure prediction method [4]. A few are considered below, but most are outside our scope [5, 6]. A journal special issue covers several design topics beyond our scope [7]. Enzyme design was reviewed in depth earlier [8, 9]. Interesting studies not directly tied to physics-based methods are excluded or only mentioned, including *de novo* ligand binding-site construction [10] and allostery design [11]. New methods within the Rosetta software were reviewed recently [12]. More directly related to the question in our title are advances in the Osprey software package [13], which implements a sophisticated side chain rotamer treatment as well as methods to compute ligand-binding free energies. Also related is the recent Proteus software release [14], which implements physics-based design. Finally, physics-based design methods and underlying theory were reviewed [14].

2 Protein stability and redesign

2.1 The unfolded state

Whole protein design or redesign depends on an energy function for the unfolded state. So far, applications have used heavily-parametrized, purely empirical functions. Physics-based models are an interesting future perspective. Several recent studies are of interest. Peran et al. provided a high-resolution description of unfolded states of a small protein from FRET and SAXS experiments, all-atom MC simulations, and polymer theory [15]. Under refolding conditions, the unfolded state was less compact than the native, and included some residual, native, helical structure. Local and nonlocal intra-protein interactions were inferred, both native and nonnative. Sequence-specific interactions introduced significant deviations from idealized homopolymer models. Two groups used coarse-grained models and mean field theory to study electrostatic interactions and pH effects in the unfolded state [16, 17]. Interesting observations included the role of the position of ionized groups within the sequence and the good performance of mean field theory for a disordered protein structure. The past year has seen several publications on the structure and dynamics of intrinsically disordered proteins, which could mimic the unfolded state [18, 19]. For example, extended peptide models have been used in calculations of stability changes, considered next.

2.2 Computing stability changes

A common test of computational protein design (CPD) is to predict stability changes associated with point mutations. These depend on a representation of the unfolded state. Most groups have employed simple, extended peptide models. In CPD, this approximation is associated with many others: implicit solvent, side chain rotamers, and so on. In the context of more realistic physical models, which use all-atom MD and alchemical free energy perturbation (FEP), there are far fewer approximations and the unfolded treatment can be expected to limit the overall accuracy (despite other, remaining approximations such as fixed atomic charges). FEP then gives a rough lower bound on what can be achieved with an extended peptide unfolded model, in the absence of any empirical parametrization. Two groups reported FEP tests recently. One considered 43 point mutations in lysozyme [20], and used an Ala-X-Ala tripeptide to represent the unfolded state when mutating a residue X. The mean unsigned error (MUE) was 1.4 kcal/mol and the correlation was 0.74 compared to experiment. The other group considered 87 mutations in five proteins [21], and used either a tri-, penta-, or heptapeptide for the unfolded state. Rosetta gave an MUE of 1.65 kcal/mol for this dataset. With FEP, the MUE was half as large: 0.85 kcal/mol, with a very small improvement going from a tri- to a heptapeptide unfolded model. An earlier, even larger FEP study gave similar errors [22].

Nisthal et al. produced an important experimental dataset [3], then used it to benchmark several models. Using a novel automated method, they generated stability data for almost all point mutations of a 56-residue domain from protein G (935 mutants out of 1064). By using a single approach and producing all variants, they avoided flaws inherent in the largest stability database, ProTherm, such as its predominance of large-to-small mutations. The same group maintains the new ProtoBank ressource, which includes the ProTherm data [3]. Most of the protein G single mutations had a small effect on stability. Several CPD models were tested on the data, including Rosetta. Since Rosetta does not directly provide energies in physical units, only correlations were measured. Values of 0.5–0.6 were obtained, lower than with all-atom FEP, as expected.

2.3 Whole protein redesign

The most important recent advance in whole protein design was the development a few years ago of high throughput experimental screens [1]. These allowed thousands of proteins and miniproteins to be redesigned or designed *de novo* with Rosetta, including miniproteins with activity against SARS-Cov-2 [23]. This work reported a success rate of about 6% (designed protein is expressed and correctly folded), although both higher and lower rates have been found for other templates and datasets. For example, another study redesigned two small proteins using Rosetta and flexible backbone exploration; a genetic algorithm was used to enrich the design set in sequences similar to natural ones [24]. 16

designs were experimentally tested; five could be expressed but none had a native-like circular dichroism spectrum.

Two computational advances are of interest for the question posed by our title. First, our group has studied the effect of several implicit solvent models on the ability to recover native-like sequences in whole protein redesign [14, 25]. We compared two treatments of nonpolar solvation: a surface area (SA) and a Lazaridis-Karplus (LK) term, and combined them with two Generalized Born (GB) variants for polar effects: one that maintained the many-body GB character and one that averaged it out. In all cases, the unfolded model was empirical and specifically parametrized. The best sequence recovery, comparable to Rosetta, was obtained with the more rigorous, many-body GB and the LK nonpolar term. The 2nd publication [25] includes a tutorial and scripts for unfolded model parametrization, using a well-defined, maximum-likelihood approach.

A second advance was the first successful redesign of a protein (a PDZ domain) using a nonempirical, physics-based energy function for the folded state [26]. A handful, not thousands of designs were validated, and those tested were chosen partly based on several empirical criteria. Nevertheless, all those tested were successful, and it is striking that a standard, molecular mechanics model was sufficient to produce fully-redesigned proteins.

2.4 Fitness landscapes

To end this section, we mention studies of fitness landscapes, very briefly due to space constraints. These lead to concepts like designability and evolutionary trajectories. They can also help produce and optimize designs. Thus, one study used fitness landscape information given by a neural network to perform whole protein redesign [27] (with the help of FoldIt players). The neural network predicted the probability of residue-residue distances and orientations from sets of aligned sequences, and provided gradients with respect to sequence mutations. By moving along the gradients, better distance distributions could be obtained in a series of iteration cycles, which led to an optimized sequence.

Importantly, fitness landscapes provide physical insights. Unlike full redesign calculations, most studies explore the landscape structure close to a native or a designed sequence: its ruggedness and slope, the magnitude of short- and long-range correlations, the important degrees of freedom. Chen & Wolynes derived several measures of landscape frustration and used them to show that a designed protein had a higher frustration than a natural one [28]. Two groups estimated the intrinsic dimension of sequence space for several protein families, and quantified the constraints imposed by protein symmetry [29] or by phylogenetic relations [30]. Ding et al. introduced a low-dimensional "latent space" representation, learned using methods from image processing, which allowed them to characterize the dimensionality of sequence space [31]. Finally, Blanco et al. reviewed both high-throughput experiments and simple theoretical models to probe fitness landscapes [32].

3 Protein-ligand binding

We focus on the redesign of existing binding pockets to accomodate new ligands, rather than *de novo* binding site design [10, 33]. Indeed, new physics-based methods were either developed for, or can be understood from redesign applications. We distinguish three subproblems. The first is to refine an initial ligand pose (such as a native pose), before the design step *per se*. The second is to post-process designed complexes using more accurate and costly methods; for example, MD with explicit solvent, combined with a Poisson-Boltzmann (PB) free energy function. The third and most important subproblem is to design a complex using the *binding affinity* as the design target. CPD has produced impressive successes, like miniproteins that bind the SARS-Cov-2 receptor binding domain [23]. Most of these successes (and many failures) were obtained by optimizing the *total energy* of the complex, rather than its binding affinity. Designing for affinity is a much harder problem, and the best current solution was discovered only recently, thanks to an adaptive landscape flattening approach [34–36]. Finally, in a separate subsection, we turn to the special case of enzyme redesign, where some specific new methodology has been proposed.

3.1 Pre- and post-processing steps

Protein-ligand redesign applies mutations around a binding pocket, while allowing ligand, side chains, and possibly backbone to adjust their conformations. The allowed ligand conformations are an essential input. Often, a few poses are used and, for each one, internal deformations are allowed, similar to side chain rotamers. Poses can be borrowed from a native ligand. They can also be obtained by docking methods, which have seen major advances in speed this year through GPU implementations [37, 38]. Another study reported a complete molecular mechanics force field for small molecules, optimized to reproduce crystal structures, then combined it with the Rosetta protein and solvent energy functions in a large-scale docking benchmark. Among over 1000 test calculations, over half recovered the experimental binding mode with sub-Ångstrom accuracy [39].

Gilabert et al. presented a Protein Energy Landscape Exploration, or PELE, to identify ligand binding poses and calculate absolute binding free energies [40]. They explored and evaluated different ligand binding poses, generated by random ligand translation and rotation moves, protein backbone perturbations along normal modes, side chain prediction, and global energy minimization. The energy function combined molecular mechanics and GB solvent. Mobley et al. developed a hybrid MD/MC method to identify binding modes or bound conformations of flexible ligands, and to compute their relative populations [41]. In addition, they proposed a new, "Molecular Darting" MC move, which allows one to reversibly hop between several, predefined binding modes [42], and could be incorporated into protein design schemes.

It is also important to post-process designed complexes with more realistic models. MD simulations can be readily applied to dozens of designs, allowing the rotamer, fixed backbone, and implicit solvent approximations to be removed. GB or PB or PBSA rescoring can be applied [43], and alchemical FEP can be used for a few designs. If the design used a molecular mechanics force field [14], post-processing can use the same one, facilitating the interpretation. Several recent studies illustrate these ideas. One study applied 0.5–1 μs of MD to each of 37 designs, to determine the mobility of ligand, solvent, and protein within the binding pocket [44]. Increased mobility indicated lower quality designs, and machine learning approaches built from the trajectories could further help discriminate successful/failed binders. Another study applied a Linear Interaction Energy (LIE) free energy function to MD trajectories of 28 Cyt P450–ligand complexes, obtaining mean errors around 1 kcal/mol [45]. No CPD was performed, but the LIE performance shows it is of use for CPD post-processing. Finally, our own group applied an earlier LIE model for PDZ–peptide binding to 15 designed peptides [46]. Deviations between the CPD and LIE affinities were about 1.5 kcal/mol (triple the estimated LIE error) and negative, with CPD over-binding. In addition, this study surveyed the accuracy of PBSA and related free energy functions in 15 studies. The reported accuracies appear to be below the potential of these methods, possibly due to implementation choices.

3.2 A method to design for affinity

Rigorous methods to design for binding affinity are very valuable. For small problems, so-called partition function methods can enumerate sequences within a given energy window of the Global Minimum Energy Conformation (GMEC). Doing this for a protein and a protein–ligand complex leads to a controlled approximation for the relative binding free energies of protein variants [47]. Recent work led to improved K* algorithms, with increased pruning of high energy sequences and conformations [48]. This led to a 10–100 fold speedup for a set of 41 (mostly single) mutations of the Ras–Raf binding interface. The rank order of mutant affinities was in good agreement with experiment (0.81 Spearman correlation), although precise affinity errors were not reported. The method was then used to redesign two positions in the same complex, for a total of 441 possible sequences. A point mutant was discovered with a 5-fold increase in the experimentally-measured affinity.

Despite such complex, powerful algorithms, partition function methods remain expensive and allow very limited sampling. An important development was the discovery [34–36] of a new, simpler and more efficient method to design specifically for binding affinity, based instead on adaptive landscape flattening (ALF). It can handle spaces of 100,000 sequence variants while providing relative binding free energies that are well-converged. Two applications were reported, plus a third that is in the enzyme subsection below. One application was the PDZ-peptide study above [46]. Peptides were designed to bind the Tiam1 PDZ domain; 15 designs were predicted to improve binding over a native peptide, but LIE showed the binding was overestimated and no actual hits were obtained. The other application redesigned the SARS-Cov-2 receptor domain to enhance ACE2 binding [49]. Both studies used an MC simulation that adaptively flattened the free energy landscape in sequence space of the designed entity (receptor domain, say), by optimizing a bias function B, in the absence of the ligand (absence of ACE2). Once the landscape is flattened, B closely approximates the sequence free energy of the apo receptor, up to a sign change. B was then included in a simulation of the receptor-ACE2 complex (holo state), where it subtracted out the unbound state. Thus, the biased MC simulation sampled sequences according to their binding free energy [34], and sequences with increased affinity were exponentially enriched. Notice that since this CPD method yields binding *free energies*, comparison to experimental affinities or FEP results is considerably simplified. Importantly, since the method is applicable to an enzyme binding its transition state, it can allow the design of catalytic efficiency, as shown in the next subsection.

3.3 Enzyme redesign

Many enzyme studies have sought to optimize the total energy of the enzyme-substrate complex [8], instead of optimizing catalysis, and this may contribute to the weak activity of some designs, even though negative or partly-negative results are, unfortunately, rarely published [50]. Several recent studies have sought to optimize catalysis more directly, by tuning the electric field in the active site [9, 51]. Beker et al. introduced the "inverse catalysis problem" for enzyme design: the determination of the catalytic fields from the transition state and reactant state wavefunctions [52]. This then allowed the direct extraction of the geometrical characteristics of the optimal catalytic site, as well as transition-state stabilization and ground-state destabilization effects. The method was validated by comparing to experimental data for Kemp eliminase mutants. Warshel and coworkers performed a pilot study of the enzyme haloalkane dehalogenase, using MD with the Empirical Valence Bond semi-classical model to directly estimate catalytic effects of several mutations, which were then validated by experiments [53]. One of the mutants had a higher-than-wildtype efficiency. Bonk et al. used MD and importance sampling to simulate the reaction pathway of the KARI enzyme, then used machine learning techniques to identify structural and dynamical characteristics that promote catalysis [5]. They noted the existence of a region in conformational space that promotes reactivity when populated. Its main defining features were the substrate conformation, substrate bond polarization and metal coordination geometry.

In fact, when designing an enzyme, it is now possible to target directly the binding affinity of the transition state, also known as the catalytic efficiency, which is equal to the 2nd order rate constant k_{cat}/K_M under Michaelis-Menten theory. Indeed, the ALF method to design for binding affinity applies to transition state binding. The ALF approach can also be used to design for binding specificity: for example, transition state binding *vs.* substrate binding. In that case, one optimizes the catalytic rate k_{cat}, instead of the efficiency k_{cat}/K_M. With these approaches, there is no need to explicitly consider electric fields. Rather, one acts directly on the rate, either via k_{cat} or k_{cat}/K_M. The first application to an enzyme appeared recently [54]. Methionyl-tRNA synthetase was the test system and the Proteus software was used. Known variants with activity towards the unnatural amino acid azidonorleucine were recovered, and new variants with activity towards the native substrate methionine were predicted, then confirmed experimentally. In followup work, new variants were obtained with activity towards the unnatural amino acid β-methionine [14]. Both studies used a physics-based, MMGBLK energy function.

4 Energy function developments

CPD energy functions often contain empirical terms, and new ones are constantly being added, tuned, and reparametrized [12, 55, 56]. As machine learning becomes more widespread and successful, this trend is on the rise. Physics-based energy functions, on the other hand, are transferable to all biomolecules, can be systematically improved, and give physical insights. Recent successful CPD applications and simulation work in general show that physical realism remains a powerful route, where progress continues. Two important threads are the rapid rise of polarizable molecular mechanics and the continuing refinement of implicit solvent models.

In CPD, many of the protein degrees of freedom are represented implicitly [14]: electronic polarization, bond and angle stretching and, in many cases, backbone motions [14]. As a result, the protein dielectric constant is normally set to a value much greater than one, like 4 or 8. How then to construct a consistent model that treats electronic polarization explicitly, but many other protein degrees of freedom implicitly? One study showed that modeling protein electronic polarization explicitly, in combination with a PB solvent, led to improved acid/base predictions for several proteins, even though all other protein degrees of freedom were treated implicitly [57]. Notice that side chain protonation/deprotonation is formally analogous to a mutation, and thus acid/base calculations are directly of interest to CPD. Another study quantified the effect of the polarizable Amoeba force field on side chain repacking [58], an important sub-task of CPD. For a collection of proteins important for hearing pathologies, the authors observed a systematically improved agreement with experimental X-ray data. Other polarizable force field developments have been reviewed [59].

Another important thread is refinement of implicit solvent models. Building on the Amoeba polarizable protein force field, the Tinker developers reported a new implicit solvent model that was developed and parametrized in combination with Amoeba [60]. The continuum electrostatic term was based on an analytic generalized Kirkwood approximation. The nonpolar term was based on novel cavitation and dispersion estimators. The model was completely parametrized and tested, both on small molecules and proteins. The Generalized Born electrostatic model has often been used for protein design, and this model continues to progress. A very good analysis of its physical basis was published recently [61]. An efficient GPU implementation of a combined GBSA model was provided within the CHARMM/OpenMM software [62]. Such fast implementations will be especially valuable for sampling backbone motions in CPD. GB optimization and benchmarking for protein-ligand binding were also reported [63], with a powerful global optimization algorithm leading to improved parameters.

In CPD, GB is normally combined with a nonpolar model, with SA models among the most popular. Our group showed that a Lazaridis-Karplus nonpolar term gave superior performance for native sequence recovery in whole protein design [14, 25]. In addition, GBLK gave larger sequence entropies, in much better agreement with natural sequence diversity. Two semi-empirical implicit membrane models were implemented in Rosetta [64, 65], and gave good performance for membrane protein prediction and design. While these models were not derived from first principles, they provide a considerable increase in realism compared to a simple hydrophobic slab.

Finally, mixed models continue to be explored, where a few water molecules are represented explicitly, while the bulk of solvent is treated implicitly. Notice that it is technically straightforward to include a few explicit waters in CPD models, and could improve accuracy for some problems. One group used Rosetta to include a few explicit waters at protein-protein and protein-ligand interfaces, and obtained improved native structure recovery in protein-protein discrimination tests [66]. Another used MC within the PELE tool to sample explicit waters buried in protein cavities or at interfaces, and evaluated performance by comparing to crystal water positions [67]. Another group developed hybrid MD/MC to study the equilibrium between buried and bulk water molecules, and improve the accuracy of protein-ligand binding free energy calculations [68].

5 Sampling backbone conformations

The importance of backbone flexibility was underlined recently by an enzyme design study, where CPD with certain backbone conformations led to recovery of the correct side chain organization in the active site, while most did not [69]. Including backbone flexibility has a major effect on the structure and complexity of CPD computations. Algorithmic aspects and methods that provably identify the GMEC were reviewed recently [7].

The simplest strategy to explicitly model backbone flexibility in CPD is to run calculations with a few distinct backbone conformations. It is common to run a design calculation with a fixed backbone, then relax the side chain and backbone structure with a fixed sequence ("design-then-relax"), then iterate, as in two recent studies [70, 71]. One can also run in parallel calculations with a collection of backbone conformations, determined in advance. For example, in an MC simulation, one could apply the same mutation to several backbones, each carrying the same sequence, then accept or reject the mutation based on an average energy. This multibackbone method was shown to increase sequence conservation, compared to a series of independent designs, for a protein that undergoes large backbone rearrangements [72]. It was also applied in a recent successful enzyme redesign [73]. The authors redesigned the *E. coli* BCAT enzyme to accept two substrates, using a total of 300 backbone templates. Experimental validation revealed four variants with a 200-fold increase in catalytic efficiency.

Multistate design requires only modest changes to MC code. In contrast, another group reported difficult and impressive algorithmic work to provably obtain the GMEC in the multibackbone case [74]. This is an example where physics-based sampling of Boltzmann ensembles can provide top sequences with tight confidence intervals, using much simpler algorithms than GMEC proofs. Note also that the GMEC is only determined up to the typical errors in the energy function.

The second strategy for backbone flexibility is to include backbone deformation moves directly in the MC scheme. Hops between predefined backbones require a sophisticated and expensive, hybrid MC move [14]. Local moves, such as "backrub" motions, do not require such specific machinery, although they do destroy the possibility of a fixed energy matrix. Loshbaugh and Kortemme quantified the ability of different backrub move sets to recapitulate observed protein sequence profiles in a set of 21 test proteins [71]. They also considered Kinetic Closure moves, which are slightly less local. The flexible backbone methods performed better than iterations of design-then-relax and better than fixed backbone design in almost all cases.

A third strategy uses MD simulations to sample backbone and side chain degrees of freedom. This strategy is picking up speed. MD was initially introduced for two problems related to CPD: (1) alchemical free energy simulations to asses the effect of side chain mutations on ligand binding and (2) constant-pH MD to compare side chain protonation states. The first development used a pseudo-coordinate lambda to weight a particular side chain. Lambda was propagated through MD, along with the other system coordinates, a method known as "lambda dynamics" [75]. Constant pH simulations were proposed around the same time [76]. Lambda dynamics was applied to the redesign of up to five sites in lysozyme, with 2–5 allowed side chain types each, for a total of up to 240 allowed sequences. Unlike other CPD calculations, these simulations used explicit solvent and periodic boundary conditions. Folding free energy changes could be estimated from side chain type populations, because equilibrium MD naturally explores a Boltzmann distribution. Mean unsigned errors for 32 single mutations were just 1.2 kcal/mol. More recently, adaptive landscape flattening was used to estimate protein-ligand binding free energies; a GPU implementation was developed, and an automatic workflow was set up [77].

Acid/base calculations are related to CPD: they have the same structure as design calculations, with (de)protonation playing the role of a mutation. Thus, CPD software can often do pK_a calculations, and these provide a benchmark for the physical model employed. In addition, during CPD, it is common to allow multiple protonation states of certain side chain types, like His, Cys, Asp, or Glu. Michael et al. implemented an MC/MD method in the CPD software Proteus, and tested it using acid/base calculations on five proteins [78]. Typically, one side chain was allowed to change its protonation state and another was allowed to change its type. MC moves where a protonation state or a side chain type were changed alternated with short MD segments where the whole protein was flexible. A GB implicit solvent was used. With the MC/MD method and its flexible backbone treatment, the mean pK_a error was reduced by almost half. In addition, an adaptive landscape flattening method was applied in the space of sequences and protonation states, which improved efficiency. The implementation is still too slow for routine applications. Other developments in constant-pH MD include a very fast, GPU implementation in Amber [79] and a replica-exchange MD implementation combined with Poisson-Boltzmann implicit solvent [80].

6 Conclusions

Physical chemistry's first main job is to provide CPD with predictive energy functions [59–61]. The first main development covered here was the application of polarizable energy functions (Drude, Amoeba) to two CPD sub-problems: side chain repacking and pK_a's [57, 58]. It is tricky to mix, in the same model, explicit electron rearrangements with an implicit treatment of, say, backbone motions. With continuum electrostatics, it is more common to average out the faster motions (electron rearrangements) and treat the slower ones (backbone) explicitly. While the applications above gave improved performance, it remains to be seen just how self-consistent and how predictive such models will be in CPD. The second main development covered above was the continuing improvement of implicit solvents models, which are one of the factors that limit accuracy in biomolecular simulations.

The second area where physics contributes is sampling. Physics-based sampling that follows a Boltzmann distribution is an important search strategy. As we extend trajectories, confidence intervals become tighter, until quantities of interest, like the GMEC, are inferred with an uncertainty below that of the energy function. With Boltzmann sampling, new design strategies have become possible. We saw above that with the introduction of new, adaptive landscape flattening algorithms, free energies can be obtained, such as the binding free energy of an enzyme's transition state, at a modest computational cost. Designed protein variants that lower the activation free energy or increase catalytic efficiency are exponentially enriched during the MC or MD trajectory, and the optimal variants are readily identified.

The third role of physics is to help describe motions. A main development was the use of MD for conformational sampling. This method remains expensive but has the potential to dramatically improve the physical model.

Despite its many successes, CPD still has a low success rate and there is considerable room for improvement. The advances above are complementary to widely-used empirical ingredients, and should help make CPD much more predictive. Although we mostly focussed on methodology, several important applications were also convered, including examples of whole protein design, enzyme design, and design of membrane proteins. Finally, high throughput experimental assays are incredibly important in CPD and, as they continue to develop, the whole CPD field will advance.

References

[1] G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, D. Baker, Global analysis of protein folding using massively parallel design, synthesis, and testing, Science 357 (2017) 168–175.

[2] H. Q. Nguyen, J. Roy, B. Harink, N. P. Damle, N. R. Latorraca, B. C. Baxter, K. Brower, S. A. Longwell, T. Kortemme, K. S. Thorn, M. S. Cyert, P. M. Fordyce, Quantitative mapping of protein-peptide affinity landscapes using spectrally encoded beads, Elife 8 (2019) e40499–e40526.

[3] A. Nisthal, C. Y. Wang, M. L. Ary, S. L. Mayo, Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis, Proc. Natl. Acad. Sci. USA 116 (2019)

16367–16377.
⋆⋆ Using a novel automated method, the authors generated experimental stability data for almost all point mutations of a protein G domain (935 out of 1064) then used it to benchmark several models.

[4] J. Jumper, R. Evans, A. Pritzel1, T. Green, et al, Highly accurate protein structure prediction with AlphaFold, Nature in press (2021). doi:10.1038/s41586-021-03819-2.

[5] B. M. Bonk, J. W. Weis, B. Tidor, Machine learning identifies chemical characteristics that promote enzyme catalysis, J. Am. Chem. Soc. 141 (2019) 4108–4118.
⋆ Machine learning was used to interpret MD simulations of the KARI enzyme and identify the structural and dynamical features that promote catalysis.

[6] W. Gao, S. P. Mahajan, J. Sulam, J. J. Gray, Deep learning in protein structural modeling and design, Patterns 1 (2020) 1–23.

[7] Y. Bouchiba, J. Cortes, T. Schiex, S. Barbe, Molecular flexibility in computational protein design: an algorithmic perspective, Prot. Eng. Des. Sel. 34 (2021) 1–7.

[8] H. Lechner, N. Ferruz, B. Hoecker, Strategies for designing non-natural enzymes and binders, Curr. Opin. Chem. Biol. 47 (2018) 67–76.

[9] V. V. Welborn, T. Head-Gordon, Computational design of synthetic enzymes, Chem. Rev. 119 (2019) 6613–6630.

[10] J. E. Lucas, T. Kortemme, New computational protein design methods for *de novo* small molecule binding sites, PLoS Comp. Biol. 16 (2020) e1008178–e1008204.

[11] Y. L. Vishweshwaraiah, J. Chen, N. V. Dokholyan, Engineering an allosteric control of protein function, J. Phys. Chem. B 125 (2021) 1806–1814.

[12] J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, et al., Macromolecular modeling and design in Rosetta: recent methods and frameworks, Nat. Meth. 17 (2020) 665–680.

[13] M. A. Hallen, J. W. Martin, A. Ojewole, J. D. Jou, A. U. Lowegard, M. S. Frenkel, P. Gainza, H. M. Nisonoff, A. Mukund, S. Wang, G. T. Holt, D. Zhou, E. Dowd, B. R. Donald, OSPREY 3.0: Open-Source Protein Redesign for You, with powerful new features, J. Comput. Chem. 39 (2018) 2494–2507.

[14] D. Mignon, K. Druart, E. Michael, V. Opuu, S. Polydorides, F. Villa, T. Gaillard, T. Gaillard, N. Panel, G. Archontis, T. Simonson, Physics-based computational protein design: an update, J. Phys. Chem. A 124 (2020) 10637–10648.

[15] I. Peran, A. S. Holehouse, I. S. Carrico, R. V. Pappu, O. Bilsel, D. P. Raleigh, Unfolded states under folding conditions accommodate sequence-specific conformational preferences with random coil-like dimensions, Proc. Natl. Acad. Sci. USA 116 (2019) 12301–12310.

[16] M. Mugnao, D. Thirumalai, Molecular transfer model for pH effects on intrinsically disordered proteins: theory and applications, J. Chem. Theory Comput. 17 (2021) 1944–1954.

[17] V. M. de Oliveira, D. L. Z. Caetano, F. B. da Silva, P. R. Mouro, A. B. de Oliveira Jr., S. J. de Carvalho, V. B. P. Leite, pH and charged mutations modulate Cold Shock protein folding and stability: a constant pH Monte Carlo study, J. Chem. Theory Comput. 16 (2020) 765–772.

[18] Y. Zhao, R. Cortes-Huerto, K. Kremer, J. F. Rudzinski, Investigating the conformational ensembles of intrinsically disordered proteins with a simple physics-based model, J. Phys. Chem. B 124 (2020) 4097–4113.

[19] B. B. Kragelund, K. Skriver (Eds.), Intrinsically Disordered Proteins: Methods and Protocols, Springer Verlag, New York, 2021.

[20] W. Jespers, G. V. Isaksen, T. A. Andberg, S. Vasile, A. van Veen, J. Åqvist, B. O. Brandsdal, H. Gutierrez-de-Teran, QresFEP: an automated protocol for free energy calculations of protein mutations in Q, J. Chem. Theory Comput. 15 (2019) 5461–5473.

[21] J. Duan, D. Lupyan, L. Wang, Improving the accuracy of protein thermostability predictions for single point mutations, Biophys. J. 98 (2021) 2309–2316.

[22] V. G. annd Servaas Michielssens, D. Seeliger, B. de Groot, Accurate and rigorous prediction of the changes in protein free energies in a large-scale mutation scan, Ang. Chemie 55 (2016) 7364–7368.

[23] L. Cao, I. Goreshnik, B. Coventry, J. B. Case, L. Miller, L. Kozodoy, R. E. Chen, L. Carter, L. Walls, Y.-J. Park, L. Stewart, M. Diamond, D. Veesler, D. Baker, De novo design of picomolar SARS-Cov-2 miniprotein inhibitors, Science 370 (2020) 426–431.

[24] G. Sormani, Z. Harteveld, S. Rosset, B. Correia, A. Laio, A Rosetta-based protein design protocol converging to natural sequences, J. Chem. Phys. 154 (2021) 074114.

[25] V. Opuu, D. Mignon, T. Simonson, Modeling the unfolded state for protein design, Computational Peptide Science: Methods and Protocols, Methods Molec. Biol. 9999 (2021) Chapter 19.

[26] V. Opuu, Y. J. Sun, T. Hou, N. Panel, E. J. Fuentes, T. Simonson, A physics-based energy function allows the computational redesign of a pdz domain, Scientific Reports 10 (2020) 11150.
⋆ This work reports the first successful whole protein redesign using a physics-based scoring function for the folded protein state, in combination with a knowledge-based energy for the unfolded state.

[27] C. Norn, B. I. M. Wicky, D. Juergens, S. Liu, D. Kim, D. Tischer, B. Koepnick, I. Anishchenko, F. Players, D. Baker, S. Ovchinnikov, Protein sequence design by conformational landscape optimization, Proc. Natl. Acad. Sci. USA 118 (2021) e2017228118.

[28] J. Chen, N. P. Schafer, P. G. Wolynes, C. Clementi, Localizing frustration in proteins using all-atom energy functions, J. Phys. Chem. B 123 (2019) 4497–4504.

[29] J. Marchi, E. A. Galpern, R. Espada, D. U. Ferreiro, A. M. Walczak, T. Mora, Size and structure of the sequence space of repeat proteins, PLoS Comp. Biol. 15 (2019) e1007282.

[30] E. Facco, A. Pagnani, E. T. Russo, A. Laio, The intrinsic dimension of protein sequence evolution, PLoS Comp. Biol. 15 (2019) e1006767.

[31] X. Ding, Z. Zou, C. L. Brooks III, Deciphering protein evolution and fitness landscapes with latent space models, Nature Comm. 10 (2019) 5644.

[32] C. Blanco, E. Janzen, A. Pressman, R. Saha, I. A. Chen, Molecular fitness landscapes from high-coverage sequence profiling, Ann. Rev. Biophys. 48 (2019) 1–18.

[33] B. Basanta, M. J. Bick, A. K. Bera, C. Norn, C. M. Chow, L. P. Carter, I. Goreshnik, F. DiMaio, D. Baker, An enumerative algorithm for de novo design of proteins with diverse pocket structures, Proc. Natl. Acad. Sci. USA 117 (2020) 22135–22145.

[34] A. Bhattacherjee, S. Wallin, Exploring protein-peptide binding specificity through computational peptide screening, PLoS Comp. Biol. 7 (2013) e1003277.

[35] R. L. Hayes, K. A. Armacost, J. Z. Vilseck, C. L. Brooks III, Adaptive landscape flattening accelerates sampling of alchemical space in multisite lambda dynamics, J. Phys. Chem. B 121 (2017) 3626–3635.

[36] F. Villa, N. Panel, X. Chen, T. Simonson, Adaptive landscape flattening in amino acid sequence space for the computational design of protein:peptide binding, J. Chem. Phys. 149 (2018) 072302.

[37] X. Ding, Y. Wu, Y. Wang, J. Z. Vilseck, C. L. Brooks III, Accelerated CDOCKER with GPUs, parallel simulated annealing, and Fast Fourier transforms, J. Chem. Theory Comput. 16 (2020) 3910–3919.

[38] M. Fan, J. Wang, H. Jiang, Y. Feng, M. Mahdavi, K. Madduri, M. T. Kandemir, N. V. Dokholyan, GPU-accelerated flexible molecular docking, J. Phys. Chem. B 125 (2021) 1049–1060.

[39] H. Park, G. Zhou, M. Baek, D. Baker, F. DiMaio, Force field optimization guided by small molecule crystal lattice data enables consistent sub-Angstrom protein–ligand docking, J. Chem. Theory Comput. 17 (2021) 2000–2010.

[40] J. F. Gilabert, C. Grebner, D. Soler, D. Lecina, M. Municoy, O. G. Carmona, R. Soliva, M. J. Packer, S. J. Hughes, C. Tyrchan, A. Hogner, V. Guallar, PELE-MSM: A Monte Carlo based protocol for the estimation of absolute binding free energies, J. Chem. Theory Comput. 15 (2019) 6243–6253.

[41] S. Sasmal, S. C. Gill, N. M. Lim, D. L. Mobley, Sampling conformational changes of bound ligands using Nonequilibrium Candidate Monte Carlo and molecular dynamics, J. Chem. Theory Comput. 16 (2020) 1854–1865.

[42] S. C. Gill, D. L. Mobley, Reversibly sampling conformations and binding modes using molecular darting, J. Chem. Theory Comput. 17 (2021) 302–314.

[43] T. Sitthiyotha, S. Chunsrivirot, Computational design of 25-mer peptide binders of SARS-CoV-2, J. Phys. Chem. B 124 (2020) 10930–10942.

[44] E. P. Barros, J. M. Schiffer, A. Vorobieva, J. Dou, D. Baker, R. E. Amaro, Improving the efficiency of ligand-binding protein design with molecular dynamics simulations, J. Chem. Theory Comput. 15 (2019) 5703–5715.

[45] E. A. Rifai, V. Ferrario, J. Pleiss, D. P. Geerke, Combined linear interaction energy and alchemical solvation free-energy approach for protein-binding affinity computation, J. Chem. Theory Comput. 16 (2020) 1300–1310.

[46] N. Panel, F. Villa, V. Opuu, D. Mignon, T. Simonson, Computational design of PDZ-peptide binding, PDZ mediated interactions: methods and protocols, Methods Molec. Biol. 2256 (2021) Chapter 14, pgs. 239–258.

[47] P. Gainza, H. M. Nisonoff, B. R. Donald, Algorithms for protein design, Curr. Opin. Struct. Biol. 39 (2016) 16–26.

[48] A. U. Lowegard, M. S. Frenkel, G. T. Holt, J. D. Jou, A. A. Ojewole, B. R. Donald, Novel, provable algorithms for efficient ensemble-based computational protein design and their application to the redesign of the c-Raf-RBD:KRas protein-protein interface, PLoS Comp. Biol. 16 (2020) e1007447.

⋆⋆ Improved K* algorithms were applied to protein-protein binding and partition function estimation. The rank order of 41 mutant affinities was in good agreement with experiment. The method was used to redesign two positions in the complex and a mutant discovered with improved binding.

[49] S. Polydorides, G. Archontis, Computational optimization of the SARS-CoV-2 receptor-binding-motif affinity for human ACE2, Biophys. J. in press (2021) doi:10.1016/j.bpj.2021.02.049.

[50] K. Oki, F. S. Lee, S. L. Mayo, Attempts to develop an enzyme converting DHIV to KIV, Prot. Eng. Des. Sel. 32 (2019) 261–270.

[51] V. V. Welborn, L. R. Pestana, T. Head-Gordon, Computational optimization of electric fields for better catalysis design, Nat. Catal. 1 (2018) 649–655.

[52] W. Beker, W. A. Sokalski, Bottom-up nonempirical approach to reducing search space in enzyme design guided by catalytic fields, J. Chem. Theory Comput. 16 (2020) 3420–3429.

[53] G. Jindal, K. Slanska, V. Kolev, J. Damborsky, Z. Prokop, A. Warshel, Exploring the challenges of computational enzyme design by rebuilding the active site of a dehalogenase, Proc. Natl. Acad. Sci. USA 116 (2019) 389–394.

⋆ The authors performed design to maximize the catalytic activity of the enzyme DhlA, by first designing mutations that reduced activity, then introducing mutations that restored and increased it.

[54] V. Opuu, G. Nigro, T. Gaillard, Y. Mechulam, E. Schmitt, T. Simonson, Adaptive landscape flattening allows the design of both enzyme:substrate binding and catalytic power, PLoS Comp. Biol. 16 (2020) e1007600.

⋆⋆ This study reports the first example of enzyme redesign using transition state stabilization as the design target, with the help of adaptive landscape flattening Monte Carlo. For methionyl-tRNA synthetase, experimental catalytic efficiencies were reproduced and new active mutants obtained.

[55] S. T. Smith, J. Meiler, Assessing multiple score functions in Rosetta for drug discovery, PLoS One 15 (2020) e0240450.

[56] B. Coventry, D. Baker, Protein sequence optimization with a pairwise decomposable penalty for buried unsatisfied hydrogen bonds, PLoS Comp. Biol. 17 (2021) e1008061.

[57] A. Aleksandrov, B. Roux, A. D. MacKerell Jr., pKa calculations with the polarizable Drude force field and Poisson-Boltzmann solvation model, J. Chem. Theory Comput. 16 (2020) 4655–4668.

[58] M. R. Tollefson, J. M. Litman, G. Qi, C. E. O'Connell, M. J. Wipfler, R. J. Marini, V. Bernabe, Hernan, W. T. A. Tollefson, T. A. Braun, T. L. Casavant, R. J. H. Smith, M. J. Schnieders, Structural insights into hearing loss genetics from polarizable protein repacking, Biophys. J. 117 (2019) 602–612.

[59] Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J.-P. Piquemal, P. Ren, Polarizable force fields for biomolecular simulations: recent advances and applications, Ann. Rev. Biophys. 48 (2019) 371–94.

[60] R. A. Corrigan, G. Qi, A. C. Thiel, J. R. Lynn, B. D. Walker, T. L. Casavant, L. Lagardere, J.-P. Piquemal, J. W. Ponder, P. Ren, M. J. Schnieders, Implicit solvents for the polarizable atomic multipole AMOEBA force field, J. Chem. Theory Comput. 17 (2021) 2323–2341.
⋆⋆ 3 implicit solvent models were developed and parameterized for use with the Amoeba polarizable force field. They combined an analytic generalized Kirkwood approximation with nonpolar part was described using novel cavitation and dispersion terms. Extensive testing against experiments was done.

[61] A. V. Onufriev, D. A. Case, Generalized Born implicit solvent models for biomolecules, Ann. Rev. Biophys. 48 (2019) 275–296.

[62] X. Gong, M. Chiricotto, X. Liu, E. Nordquist, M. Feig, C. L. Brooks III, Accelerating the Generalized Born with molecular volume and solvent accessible surface area implicit solvent model using graphics processing units, J. Comput. Chem. 41 (2020) 830–838.

[63] N. Forouzesh, A. Mukhopadhyay, L. T. Watson, A. V. Onufriev, Multidimensional global optimization and robustness analysis in the context of protein-ligand binding, J. Chem. Theory Comput. 16 (2020) 4669–4684.

[64] R. F. Alford, P. J. Fleming, K. G. Fleming, J. J. Gray, Protein structure prediction and design in a biologically realistic implicit membrane, Biophys. J. 118 (2020) 2042–2055.

[65] J. Y. Weinstein, A. Elazar, S. J. Fleishman, A lipophilicity-based energy function for membrane-protein modelling and design, PLoS Comp. Biol. 15 (2019) e1007318.

[66] R. E. Pavlovicz, H. Park, F. DiMaio, Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking discrimination, PLoS Comp. Biol. 16 (2020) e1008103.

[67] M. Municoy, S. Roda, D. Soler, A. Soutullo, V. Guallar, AquaPELE: a Monte Carlo-based algorithm to sample the effects of buried water molecules in proteins, J. Chem. Theory Comput. 16 (2020) 7655–7670.

[68] I. Y. Ben-Shalom, Z. Lin, B. K. Radak, C. Lin, W. Sherman, M. K. Gilson, Accounting for the central role of interfacial water in protein-ligand binding free energy calculations, J. Chem. Theory Comput. 16 (2020) 7883–7894.

[69] A. Broom, R. V. Rakotoharisoa, M. C. Thompson, N. Zarifi, E. Nguyen, N. Mukhametzhanov, L. Liu, J. S. Fraser, R. A. Chica, Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico, Nat. Comm. 11 (2020) 4808.

[70] J. B. Maguire, H. K. Haddox, D. Strickland, S. F. Halabiya, B. Coventry, J. R. Griffin, S. V. S. R. K. Pulavarti, M. Cummins, D. F. Thieker, E. Klavins, T. Szyperski, F. DiMaio, D. Baker, B. Kuhlman, Perturbing the energy landscape for improved packing during computational protein design, Proteins 89 (2021) 436–449.

[71] A. L. Loshbaugh, T. Kortemme, Comparison of Rosetta flexible-backbone computational protein design methods on binding interactions, Proteins 88 (2019) 206–226.
$\star$ A large benchmark quantified the ability of flexible-backbone design methods with Rosetta to reproduce observed protein sequence profiles in a set of 21 test proteins. Performance increases with different methods and Monte Carlo move sets were quantified.

[72] M. F. Sauer, A. M. Sevy, J. E. Crowe Jr., J. Meiler, Multi-state design of flexible proteins predicts sequences optimal for conformational change, PLoS Comp. Biol. 16 (2020) e1007339.

[73] A. D. St-Jacques, M. E. Eyahpaise, R. A. Chica, Computational design of multisubstrate enzyme specificity, ACS Catal. 9 (2019) 5480–5485.
⋆ A multi-backbone protocol was used to design a BCAT enzyme with specificity for two substrates. 300 backbone templates were used, and four variants had a 200-fold catalytic efficiency increase.

[74] J. Vucinic, D. Simoncini, M. Ruffini, S. Barbe, T. Schiex, Positive multistate protein design, Bioinf. 36 (2020) 122–130.

[75] X. Kong, C. L. Brooks, Lambda-dynamics: a new approach to free energy calculations, J. Chem. Phys. 105 (1996) 2414–2423.

[76] A. M. Baptista, P. J. Martel, S. B. Petersen, Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration, Proteins 27 (1997) 523–544.

[77] E. P. Raman, T. J. Paul, R. L. Hayes, C. L. Brooks III, Automated, accurate, and scalable relative protein-ligand binding free energy calculations using lambda dynamics, J. Chem. Theory Comput. 16 (2020) 7895–7914.
⋆⋆ Adaptive landscape flattening was incorporated into FEP calculations using multisite lambda dynamics on GPUs. Close agreement with experiment was obtained for 31 test systems.

[78] E. Michael, S. Polydorides, T. Simonson, G. Archontis, Hybrid MC/MD for protein design, J. Chem. Phys. 153 (2020) 054113.

[79] R. C. Harris, J. K. Shen, GPU-accelerated implementation of continuous constant pH molecular dynamics in Amber: pK_a predictions with single-pH simulations, J. Chem. Inform. Model. 59 (2019) 4821–4832.

[80] D. Vila-Viccosa, P. B. P. S. Reis, A. M. Baptista, C. Oostenbrink, M. Machuqueiro, A pH Replica Exchange scheme in the stochastic titration constant pH MD method, J. Chem. Theory Comput. 15 (2019) 3108–3116.

Graphical abstract