Protein Sci
DOI10.1002/pro. 4098

Step-by-step design of proteins for small molecule interaction: a review on recent milestones

José M. Pereira[1,2], Maria Vieira[1], Sérgio M. Santos [1]

[1]CICECO & Departamento de Química, Universidade de Aveiro, Portugal

[2]Corresponding author: Campus Universitário de Santiago, 3810-193, Aveiro Portugal; (+351) 933 880 067; jose.manuel.pereira@ua.pt

Total number of pages: 33 (+8 w/ literature section)

Total number of tables: 1

Total number of figures: 7

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pro. 4098 © 2021 The Protein Society
Received: Mar 25, 2021;Revised: Apr 21, 2021;Accepted: Apr 23, 2021

Abstract:

Protein design is the field of synthetic biology that aims at developing *de-novo* custom made proteins and peptides for specific applications. Despite exploring an ambitious goal, recent computational advances in both hardware and software technologies have paved the way to high-throughput screening and detailed design of novel folds and improved functionalities. Modern advances in the field of protein design for small molecule targeting are described in this review, organized in a step-by-step fashion: from the conception of a new or upgraded active binding site, to scaffold design, sequence optimization and experimental expression of the custom protein. In each step, contemporary examples are described, and state-of-the art software is briefly explored.

Keywords:

Protein design; Computational chemistry; De-novo design; Protein synthesis

Statement:

Rational protein design can unlock the potential of new functions of proteins that nature has not explored yet. By designing proteins specific for a given task, science may be able to overcome some of the major hurdles that humanity faces today. With an impact in the most various areas of research, such as medical applications, industrial catalysis, soil remediation, among others, protein design is a growing topic of interest for the future.

1. Introduction

Synthetic biology is defined as the interdisciplinary scientific field that makes use of concepts from both biotechnology and engineering to design and construct new or improved biological systems.[1] The main difference setting this field apart from others such as genetic engineering is the focus to standardize, model and predict the biological behaviours, feeding back the gathered information to the design and improvement of new systems.[2] Part of this endeavour is centred around the development of new biological pathways and, therefore, newly modified proteins.[3] Proteins are naturally the workhorse of biological systems, and current estimations place the number of human proteins at ≈20 000 entities, excluding alternative splicing and post-translational modifications.[4] Traditionally, scientists made use of this large catalogue of engineerable protein parts to knock-out existing cellular pathways in detriment of newly upgraded ones, and, more recently, modern tools for genomic editing such as the CRISPR-Cas9 system opened the path to the creation of synthetic protein species by mutation and mixing of existing ones.[5] By tinkering with the old, scientists have paved the way for the creation of the new. Although labour-intensive and time-consuming, the development of improved proteins has already proven successful: some examples of commercially available engineered proteins are synthetic insulin homologues, as well as some engineered amylases and lipases used in food industry.[6] The new or improved proteins can be employed in numerous applications including, but not limited to, medical diagnosis and therapeutics, biosensors, research tools, environmental remediation and industrial enzymes.[13] Given this set of applications, a major topic of interest is the design of proteins for small molecule recognition (such as drugs or environmental pollutants). However, the list of readily available functional proteins is, despite extensive, limited, and naturally

obtainable proteins with specific sets of properties may not always exist. It is believed that the natural selection pressure blocks the complete exploration of folds and sequences, as proteins are natively picked for functions within the cell context.[7] This suggests the existence of huge untapped potential with uncharted functions and properties not explored by nature, but now made available to us by the implementation of de-novo protein design studies.[8]

Protein design is, in specific, the field of synthetic biology that explores this potential by rationally producing sequences of aminoacids “à la carte” that will fold into unnatural folds with novel activities or behaviours.[9] As Jane S. Richardson wrote in 1989:[10] *“No more need we be content with admiring the elegant and varied proteins produced by living things for doing their chemical, structural and organizational work – now we can tinker with those proteins however we please, and can even start from scratch and try to invent new ones.”*

Traditionally, this goal was achieved via experimental directed evolution, where a potential gene was subject to iterative rounds of mutagenesis and screening for amplification of the desired function.[11] However, more recently, one of the main strategies for the design of synthetic proteins harnesses the potential of computer simulations to design sequences and folds, including the incorporation of new functions, with high throughput from the ground up *in silico* before running experimental synthesis.[12] Each of these approaches comes with its own advantages and disadvantages, and, more often than not, the two work synergistically to achieve the best possible solution for a defined problem. Besides the introduction of new and/or improved functions, another drive for the implementation of computational tools for the study of protein folding is the ever widening gap between the known sequences and their corresponding 3D structure.[14] Addressing this gap will unveil unknown functions

of proteins in the cell, as well as expand our understanding of the underlying mechanisms behind protein folding mechanics. Before delving into more practical aspects of computational protein design, a brief exploration of the historical evolution of the topic is summarized.

Historically, computer-aided approaches started to gain traction in the 90's, as the ever-increasing computational resources necessary for the simulations were made available,[15] with improvements such as the introduction of dead-end elimination (DEE)[16] and genetic algorithms.[17] The works of Dahiyat *et al*.[18] and Harbury *et al.*[19] constitute two of the most complete works during this initial phase, among others.[20–22] In 1994 the first Critical Assessment of protein Structure Prediction (CASP) experiment was held worldwide and has been on-going every 2 years since then: the objective is to summarize and improve recent advances in the protein folding prediction problem.[23]

As the available protein databases increased in volume and detail, novel protein design strategies became possible. In 1999, Simons et al., from the Baker group, presented some of the first applications of homology-based methods for protein folding prediction,[24] resulting in the first release of the Rosetta software, one among the popularly used software since.[25] Research efforts have since shifted from fold prediction towards design of novel folds: in 2003, the first-ever unnatural fold was presented by Kuhlman et al., from the Baker group, using an improved version of the Rosetta software.[26]

More recently, high throughput screening for functionality became the major focus of research,[27,28] with developments such as the incorporation of artificial neural networks[29,30] and advancements in bioinformatics and big data sciences.[31,32] These allowed for the implementation of new and/or improved functions in the novel protein designs for the first time.[33–35]

Protein design methods can be classified based on the targets of the experiment. On one hand, several studies have tackled the modelling of protein-protein interactions and have been extensively reviewed before.[36,37] On the other hand, protein design for small molecule recognition intends to measure and/or capture low weight metabolites, such as drugs, toxins and environmental pollutants. Designed proteins are especially adequate for small-molecule recognition as they have shown extreme environmental resistance and have high specificity, often with femtomolar affinity.[38] This highly active field of research has specific applications in various fields such as biotechnology and medicine, sensors, therapeutic drug delivery systems and enzymes.[39] The design of proteins for small molecule interactions has, despite its successes, some hurdles to overcome, namely on the definition of protein-ligand free energy scoring functions with high accuracy (with correct electronic polarization effects, precise insertion of explicit water molecules in displacement effects and veracious binding site flexibility and dynamics during the docking process).[40]

The basic strategy underlying protein design for small molecule targeting can be summarized in 3 steps or tasks, as illustrated in Figure 2: firstly, a binding pocket or active site can be theorized for the desired application; secondly, a scaffold structure can be proposed in order to place the necessary amino acids in the correct 3D space; and finally, a sequence can be sampled that, theoretically, folds into the desired conformation.[41] This simplistic approach, despite incomplete, covers most of the efforts done in the past few years and described in this review.

The focus of this review will be the enumeration and discussion of several state-of-the-art approaches for each of the previously defined steps or tasks in computational protein design for interaction with small molecules, with focus on exemplary work from the last 6 years (2014-2020).

2. Scoring functions

Underlying all computational protein design efforts lies the ability to simulate the interactions between molecular species and to measure the *fitness* of a given protein conformation. In this context, *fitness* should be understood as the suitability of a given set of atom coordinates to a given goal. This goal can be anything from increasing the stability of a given conformation to augmenting the solubility of a peptide or even to stabilize the specific recognition of a target molecule. Therefore, scoring functions evaluate the current state of a system, allowing the comparison with other states, thus sampling the conformational space in search of the optimal solution.

Several options have been proposed in the past, and were traditionally subdivided in physical and statistical functions.[43] On one hand, physical functions evaluate the interaction between particles in the form of molecular-mechanics force fields. In this case, the scoring functions are often referred to as energy functions, since the calculated value reflects the energy contributions of various components, such as bond, angle, dihedral, Coulomb, Lennard-Jones, hydrogen bonding and van der Walls energies.[44] On the other hand, statistical functions rely on existing cheminformatics data to access the likelihood of a structure/sequence being plausible, based on previous observations in nature, and have often been referred to as fitness functions. Component terms of these functions include Ramachandran backbone angle preferences, side-chain rotamer preferences, long-distance contact probabilities, secondary structure probabilities, sequence-homology, among others.[45]

More recently, however, state-of-the-art energy functions incorporate both physical and statistical components in a hybrid approach, showing improved prediction power in software solutions such as the Rosetta software.[26] The default energy function of Rosetta software has evolved alongside with the code itself: In 2015,

O'Meara et al have presented the ElecHBv2 energy function (also known as Talaris2014), which encoded native distributions of hydrogen bonds and electrostatics interactions with relatively high precision,[46] following the OptE optimization methodology.[47] In 2016, Park et al released the Opt_Nov15 improvement of Talaris2014, optimizing the existing parameters in various benchmark tests such as decoy discrimination, homology modelling, molecular docking and sequence prediction.[48] Such improvements were latter condensed into the Rosetta Energy Function 2015 (also known as REF15), containing Lennard-Jones attraction/repulsion, implicit solvation, electrostatic, hydrogen bonds, disulphide bonds, Ramachandran probability and side-chain rotamer probability energy components, among others. Furthermore, this energy function also makes explicit the local environment of the protein during the simulation, accounting for pH variations and membrane microenvironment energy components. REF15 is considered the default score function of the Rosetta software since July 2017.[49]

Other contemporary energy functions make use of state-of-the-art machine learning strategies to assess the fitness of a proposed structure[50] or binding free-energies in protein-ligand systems.[51] An example of a physical machine learning model is TorchANI, an artificial model trained on a large dataset of DFT calculation results.[52] Using this approach, TorchANI has shown DFT-level accuracy in the calculation of a system energy, with much lower computational costs. However, when it comes to scoring functions, most efforts in machine learning application have been centred in statistical approaches. Among these, the AlphaFold algorithm, developed by the DeepMind team from Google, has probably gained the most renown after achieving impressive accuracy in the latest CASP 13 experiment. AlphaFold is a machine learning model trained on a large database of known sequence-structure pairs in order

to predict inter residue distances, who are then minimized by gradient descent to the basal energy state.[53] This triggered a shift from binary distance prediction research (whether two residues were in contact – usually less than 8Å) to an actual distance value prediction between residues. RaptorX, developed by Xu *et al*, has further shown the usefulness of predicting contact distances between residues, achieving high accuracies even with few homolog structures in the training set (less than 60 homologs).[54] Given these advances, the PDNET package, developed by Adhikari *et al*, was released: a fully open-source framework for deep learning of protein inter residue distances with high accuracy.[55] More recently, the Rosetta software team has built on top of these results, adding the ability to train deep convolution networks to predict inter residue orientations, in addition to distances.[56] The resulting package, dubbed trRosetta, has already been instrumental in generating 3D models of SARS-CoV-2 virus proteins with few or no homolog structures[57–59] and even vaccine constructs against this same virus.[60]

In terms of computational resources, some scoring functions can easily become too cumbersome, especially for bigger proteins, as the number of degrees of freedom increases. Therefore, although highly applicable in protein folding prediction problems, such models have yet to be extensively implemented in design efforts, where the need to thoroughly sample the conformation space is even more prominent. In such cases, coarse-grained models can be used to simplify intra-chain interactions to a few particles per residue, allowing simulations of bigger systems, longer time scales and much higher sampling of the conformational space, at the expense of reduced accuracy.[61] A state-of-the-art example is the Upside model, that reduces protein complexity to 1 single directional bead, whose position and orientation is defined by a set of over 10,000 jointly optimized parameters trained on approximately 500,000

residues. In this example, sidechains are considered as belief-propagated structures, where the precise positions of the atoms have little value and therefore do not follow typical rotamer discrimination. Despite this level of abstraction, Upside showed correct rotamer prediction values in over 90% of the test cases, with prediction times between 16 to 300 times faster.[62]

3. Designing a protein binding pocket

Several approaches have been suggested for structure-based binding pocket design, namely by stabilization of pre-defined binding shells, blind docking-based pocket search[63] and statistical methods.[64]

3.1) Stabilization of binding shells

For stabilization of pre-defined binding shells, a ligand-surrounding potential frame is defined a priori, containing information regarding the ligand non-covalent interactions, such as electrostatics, dipole moments, π- π stacking and hydrophobic effects. The design of the binding pocket intends, therefore, to stabilize these interactions through the rational placement of complementary aminoacids in the 3D space surrounding the ligand. The definition of the binding shell surrounding the ligand can be obtained from inspecting naturally found binders to that target and identifying the coordinated residues.[65]

One example for the implementation of this recipe is the digoxigenin (DIG) recognition by a protein designed in Tinberg *et al* work. In that project, the authors studied the binding shell from anti-DIG antibody 26-10 (PDB 1IGJ) and screened a set of 401 protein scaffolds for binding pockets that would mimic the identified hydrogen-bonding network, van der Waals interactions and hydrophobic packing of the ligand. The resulting set of candidates were then submitted to rounds of optimization by

mutation of specific residues both in the first and second shells of interaction, resulting in increased binding affinity and specificity towards DIG.[39,66]

Another example of this approach is the work of de los Santos *et al.* In this example, the QacR protein from *Staphylococcus aureus* was the target for rational protein design. This protein regulates the transcription of QacA, a multidrug efflux pump, and is allosterically regulated by a multitude of environmental agents. The objective was to modify the QacR protein to respond to vanillin, a by-product of lignin degradation and a known phenolic growth inhibitor. The new binding shell specific for vanillin was idealized by comparison with proteins from PDB that bind to similar molecules. The energy function was then biased, giving more weight to interactions that seem to stabilize vanillin or similar molecules in other natural proteins, namely π-π stacking and hydrogen bonds. Conformational sampling of both the ligand and binding pocket rotamers was performed by Monte Carlo simulated annealing, and several residues were allowed to mutate in the process. The candidate sequence was expressed and characterized in vitro, showing high levels of interaction with the target ligand and even extended cell growth that was normally inhibited in the presence of vanillin.[67]

In another strategy, the binding shell can be idealized *a priori* from physical characterization of the target molecule. As an example of this approach, Guffy *et al* have searched a large set of protein scaffolds using Fpocket (an open source protein cavity detection algorithm based on Voronoi tessellation)[68] to identify possible binding sites for zinc ions. The coordinating residues were optimized by applying multiple plausible rotamers and measuring the scoring function that satisfied the pre-defined binding shell of zinc ions in terms of distances and angles. The refined set of candidates then suffered 10 rounds of further optimization by allowing changes in the backbone conformation and second shell sidechain rotamers. Using this approach, the

authors were able to design protein sequences with binding affinities (K_D) of 1.1 μM, whose structure was stabilized when bound to the target zinc ions.[69]

3.2) Pre-existing pocket search

On the blind docking-based pocket search approach, the existing list of structures with curated binding pockets is blindly searched using docking protocols. The focus is therefore shifted from rationally stabilizing the binding shell surrounding the ligand in a small set of protein candidates to massively searching databases for compatible and pre-existing pockets, for the target ligand, even though further rounds of optimization by random mutation are usually employed. In 2018, Bhagavat et al released PocketDB, a massive online database with 249,096 curated binding pocket structures and their known/probable specific ligands, by screening the PDB with several different pocket identification algorithms. This database constitutes the most recent update to the known "pocketome".[70] A refined set of candidate binding pockets can therefore be identified and optimized by rotation/translation/mutation of the singled out aminoacids.

Using this approach, Fujieda *el at* have searched the protein databank for non-metalloenzymes that share three dimensional motifs with metalloenzymes, following the hypothesis that, given enzymatic promiscuity, current enzymes may have been ancient metalloenzymes that have since evolved through a different path. These reminiscent metal-binding pockets may still be slightly active or re-activated given simple mutations. One of the promising hits of this search campaign was the 6-phosphogluconolactase (6-PGLac) which, when in presence of Cu^{2+} cations, showed catalytic activity for the oxidation o-dianisidine (K_{cat} 78 ± 4 x 10^3 s^{-1}). By mutating the active site residues, single-point mutants were found with further increased catalytic activity (K_{cat} 490 ± 30 x10^3 s^{-1}).[71]

Despite its usefulness, binding pocket databases are still poorly catalogued. A more general approach tries to identify actives sites in a typical protein database and use the generated information as a starting point for binding pocket search or binding pocket stabilization. When the binding pocket of the protein is still undefined, statistical methods can be employed to find and characterize the residues of significance in an unknown active centre. Two statistical methods for active site identification and stabilization are explored.

Firstly, Statistical Coupling Analysis (SCA) is a method of covariation analysis that tries to identify pairs of independent and highly conserved residue positions responsible for a given motif or phenotype in the proteins structure, such as a binding pocket.[72–78] In this method, a given distribution of aminoacids at each position *i* is compared to the distributions in the same sites *i* in all sequences of a Multiple Sequence Alignment (MSA). An observed aminoacid distribution at site *i* that deviates from the average (i.e. only a single type of aminoacid is present at that position) illustrates a conserved site. Furthermore, if the deviation from the average on site *i* is accompanied by a deviation from average on site *j*, both residues are considered to be paired. These three-dimensional structures are often referred to as “sectors”,[79] and generally can hint at the existence of an active site in the structure (binding pockets are usually highly conserved regions of a protein). The objective is then to mutate and optimize the nature and conformation of the found residues to introduce new binding pockets. As an example, Banda-Vázquez *et al* managed to identify the aminoacids involved in the second shell stabilization of the binding pocket of LAO periplasmic binding protein of *Salmonella typhimurium* (which naturally binds L-amino acids lysine, arginine, ornithine and histidine) and successfully mutate them in order to bind L-Glutamine instead.[80]

A second statistical method with promising applications in protein design is the Direct Coupling Analysis (DCA) method. In contrast with SCA, where the empirical correlations between paired sites may result from indirect couplings (as, for example, mediated by one or more intermediate residues), DCA intends to separate the different types of correlation signals, highlighting the direct correlations only (also known as contacts) between a pair of aminoacids.[81–83] This method is, therefore, mostly employed in statistical energy function definitions, as a tool for structure prediction.[84] This application can, however, be expanded into protein design, as exemplified by the work of Peng *et al.*, where the ketoisovalerate decarboxylase enzyme (Kivd) was optimized for thermostability. In this work, DCA was employed to suggest mutations that increase the number of contacts with high correlation in homologous structures, outside the binding site, resulting in an increase in T_{50} temperature of up to 3.9ºC.[85]

4. Designing a protein scaffold

In some of the previously discussed cases the end product is an array of restricted aminoacids positions and orientations that bind specifically to the desired ligand. In order to place the newly defined binding/active site in the correct 3D space, or just improve the stability of a mutated existing binding site, a rational scaffold might need to be idealized, allowing a pre-set fold. In the past years, several design approaches have been presented in order to efficiently tackle this task.[86]

4.1) Repurpose of an existing structure

Recycled motifs are frequently found in nature, as estimations place the number of distinct tertiary structures just over a thousand.[87] This strategy aims to graft a newly defined active site in an existing stable structure.

An example of this approach is the work of Zhou *et al*, where a triple α-helix protein of unknown function from *Methanobacterium thermoautotrophicum* was modified to selectively bind to uranyl (UO_2^{2+}), by stabilizing the coordination-shell surrounding the ligand ion. In this example, the coordination shell of uranyl was obtained by comparison of structures naturally containing this ligand, and multiple scaffolds were tested for the stabilization of the defined potential by a scoring function that accounted for the effects of oxygen coordination and hydrogen bonding. A set of around 5 000 hits was further screened by identifying stability issues, steric clashes or inaccessibility to the binding site, reducing the population of promising scaffolds to only 4 structures. The selected structure (PDB 2PMR) showed binding affinity in the range of ~100 nM, which was further optimized by mutation of 3 aminoacids in the binding pocket (Leu13Asp, His64Glu, Leu67Thr), increasing the affinity of the super uranyl-binding protein to 7.4 ± 2.0 fM at pH 8.9. Such affinity values open the pathway for uranium recovery from sea water, which is estimated to contain 1 000 times more reserves than inland resources.[88]

In a simplistic approach to this same strategy, Moroz *et al* have designed a catalyst for ester hydrolysis by inducing a single mutation in Calmodulin, a 74-residue nonenzymatic protein with high stability. After docking studies to the target molecule (p-nitrophenyl-(2-phenyl)-propanoate), the potential binding sites were identified and subjected to single-point mutations by introducing histidine residues. The new conformation stability and binding energy were evaluated as part of a scoring function and promising candidates were expressed in *E. coli* for further characterization and catalytic efficiency assessment. Overall, this strategy, despite simplistic, resulted in catalytic efficiencies of up to 6600 ± 600 M^{-1} min^{-1}.[89]

In another example, Taylor *et al* have redesigned LacL allosteric binding site to be regulated by different effectors (fucose, lactitol and sucralose) beyond its natural inducers (allolactose and IPTG). In this case, the strategy included blind docking to the known allosteric site in multiple ligand conformations and consequent rounds of mutagenesis and fitness evaluation, resulting in strong responses to the new inducers without disrupting allosteric mechanisms.[90]

A final example is the work of Wijma *et al.* where the authors aimed at modifying the existing active site in limonene-1,2-epoxide hydrolase (PDB 1NWW) in order to be enantioselective to either S- or R-enantiomers of limonene-1,2-epoxide. To achieve this objective, the active site suffered mutagenesis rounds and the resulting structures were evaluated by molecular dynamics simulations. Snapshots of the ligand-containing binding site were collected at several points of the simulation and the near attack conformation (NAC) of the ligand was evaluated and classified as being in an S- or R-state, allowing the classification of the proposed active sites as stabilizers of either S- or R- enantiomers. The promising candidates from *in silico* screening were then expressed and the enantiomer selectivity was evaluated by GC. As a result, several designed sequences emerged from this work as enantioselective enzymes. Whereas the wild-type enzyme showed an enantiomeric excess of 23.5% towards the R-enantiomer, the variants pro-RR-8 (with 5 mutations) and pro-SS-16 (with 8 mutations) showed enantiomeric excess of 85.8% towards R-enantiomers and 90.2% towards S-enantiomers, respectively, showcasing the success of this study.[91]

4.2) Fragment-based scaffold design

Another approach is inspired by homologous recombination and exploits the secondary structure nature of proteins to recombine protein fragments in a modular

fashion. In this approach, naturally occurring parts of one or multiple structures, such as secondary structures, are pooled and then combined to create new conformations. This allows for high design flexibility, as most of these modules are self-stabilizing to some extent.[92] Traditionally, blind homologue recombination was explored experimentally. However, the newly designed proteins were often poorly expressed, with high tendency for aggregation or lacking functional activity. Applying computational methods to the fragment-based protein design approach allows a rational method to select fragments from a large pool and a method to optimize the novel protein sequence *a priori*, greatly augmenting the accuracy of this approach.[93]

As an application example, Brunette *et al* explored various combinations of fragments to create -helix-loop-helix-loop- repeat proteins.[94] Rational picking of the components allows for a fine tune of the conformation adopted by the full protein, resulting in models highly distinct from known natural folds, even with protein sequences having over 200 aminoacids.

In another example, Eisenbeis *et al* have grafted fragments from two distinct repetitive proteins in a chimera-fold, and then optimized the interface region by mutation of selected residues to maintain the affinity levels to phosphorylated ligands.[95]

Similarly to the methods described, Jacobs *et al* used a fragment-based approach, dubbed SEWING, to design and synthesize unnatural asymmetrical proteins, with high stability and high structural accuracy in comparison to the proposed computational models.[96]

Correia *et al* have also employed a similar strategy for the combination of poorly immunogenic peptide epitopes in non-viral scaffolds in order to increase its in vivo efficiency.[97] In one such experiment, possible scaffold fragments were obtained from multiple structures that showed promising complementary to the low-specificity epitope

4E10 for HIV-1 gp41 glycoprotein, identified from human IgG Fab antibody. The two fragments were grafted, optimized, expressed and experimentally characterized in immunological evaluation essays, showing promising results by demonstrating vaccine-induced neutralizing activity.[98]

To join two fragments, a loop section may need to be idealized that folds along the designed way. Lin *et al* have recently developed and summarized a set of rules for precise loop design.[99] Similarly, Agah *et al* focused on rethreading only the loop regions of dihydrofolate reductase from *E. coli*, where the loops were removed and reconnected to different secondary structures on the same original structure, originating a novel fold, albeit devoid of functionality.[100]

4.3) De-novo scaffold design

Finally, a third alternative sees the complete redesign of the protein backbone from scratch into a new unnatural conformation, in a bottom-up approach, which is referred to as fragment-free approach. This method tends to limit the natural complexity in evolved proteins, while opening the doors to completely new and unexplored structures.

Some examples of this strategy use repetitive scaffolds where the functional active site is then introduced. In one such case, Thomson *et al* have developed a highly stable and mutable *de novo* heptameric α-helical barrel scaffold, by combining residues in heptad repeats. In these repetitive structures, of the form *abcdefg*, the *a* and *d* position are highly hydrophobic residues (usually leucine or isoleucine), projecting the sidechains to the interior cavity of the coiled-coil. In this example, by developing the aminoacid sequence from scratch, fine tune of the protein structure could be achieved (the diameter of the inner channel in the coiled-coil conformation was set to be 8Å).

Overall, the proposed approach showed good performance with RMSD values between X-ray and computational model of 1.94 Å (1.17 Å when considering the Cα backbone atoms only).[101]

This structure was further improved by Burton *et al* with the implementation of a functionally active centre, of the form Cys-His-Glu, targeting the catalytic hydrolysis of esters (p-nitrophenyl acetate, in this case). Although promising, the initial catalytic efficiency values achieved by this design (3.7 ± 0.6 M^{-1} s^{-1}) are still ≈1000 times less efficient than the naturally occurring enzyme α-chymotrypsin.[102]

Using a similar approach, Olson *et al* designed a *de novo* four-helix bundle protein with catalase function, albeit with significantly reduced kinetics and efficiency when compared to naturally found proteins.[103]

In another example, MacDonald *et al* have recently developed specific beta-hairpin extensions to beta-solenoid repeat proteins that are expressed in vitro at the sub-Ångstrom level of precision.[104]

Several de-novo scaffold applications have also been reported following the maquette approach, where a simple topology is computationally optimized with specific aminoacids in order to host functionally active co-factors, such as porphyrins, chlorins, quinones and metal clusters, among others.[105] In this approach, the functionality is given by the presence of the naturally found co-factor, and not a covalently linked active site. C45 has been presented by the Anderson group as a workhorse for functionality integration, showcasing catalytic activity without a specific active centre.[106,107]

5. Designing a protein sequence

In some cases, the formalization of a scaffold that allows the rational placement of selected amino acids in specific conformations is not definitive, meaning that the sequence that folds into the designed structure is still completely unknown or can be further optimized. This is known as the inverse protein folding problem: what aminoacid sequence folds to the single lowest free energy conformation?[108] Regardless of the computational method employed, this problem persists as a bottleneck in synthetic protein development. This review will focus on strategies employed to find the best possible set of sequences for a pre-defined structure. Several sampling motors have been presented in past years and can be classified based on the plasticity of the target structure, that is, whether the backbone is mobile during sequence sampling.

5.1) Fixed backbone approximations

Fixed backbone approximations, also known as single-state design (SSD) or Global Minimum Energy Conformation (GMEC) based algorithms, are the traditional approach to sequence design and deal with a unique immobile structural backbone, where the identity/orientation of the aminoacids is exploited in order to decrease the overall energy of the system.[109] The identity of the aminoacids at each position is traditionally explored by deterministic algorithms, such as the dead-end elimination approach, or by stochastic algorithms, such as Monte Carlo simulated annealing. However, in recent years, fixed backbone approximations have been revisited when employed in high-throughput studies making use of machine-learning (ML) algorithms for fast screening of the universe of possibilities for the identity of each aminoacid in the designed protein.[110] These ML algorithms, such as the Support Vector Machines (SVM), Random Forest (RF) or Deep Neural Networks (DNN) have been employed in

computational pipelines for the quick identification of Mutation Sensitivity (MuSe) maps,[111] aiding in the design of proteins by identifying mutations that are supported by the current scaffold without major structural impact.[112] Various software solutions for structural stability predictions after single and multiple mutation points are available, such as ELASPIC[113] and ProMaya.[114] The orientation of the sidechains is a sub-problem of this approach, since the size of the exploitable conformational space is huge. This is usually simplified by the employment of rotamer libraries, where the possible angles for dihedrals in sidechains are restricted to a finite number of possibilities. Several state-of-the art software solutions have been presented in the past few years that efficiently tackle this problem, such as the SCWRL4[115] and Proteus[116] packages.

5.2) Multistate design

Multistate design (MSD) approaches, on the other hand, acknowledge that the crystal structure of a protein is not fully representative of its natural active shape in solution, as backbone motion is known to be functionally important. Moreover, when trying to sample all aminoacid identities/conformations, steric clashes can happen, causing the rejection of sequences that could be more stable given slight adjustments in the structure. MSD therefore performs a sequence optimization (by classical SSD) on multiple similar structures in parallel (with RMSD below 1 Å), and evaluates the fitness of the entire ensemble as a weighted average of all individual members of the experiment (also known as microstates).[117] Different strategies have been proposed in order to generate the target ensemble, such as backrub movements,[118] molecular dynamics or the PertMin algorithm.[119]

MSD can be further classified into two modes: positive and negative. In the positive mode, the reference structure for energy comparison is the initial crystal structure, with the objective being to maximize the stability of the sequence for the designed scaffold. However, it has been shown that this not always correlate with experimental stability, as the designed sequence also stabilizes off-target structures. In negative mode, this is taken into account by incorporating an additional layer of calculations, where off-target ensembles are also sampled and the weighted average of the entire ensemble acts as a comparison point for the stability assessment. This allows the algorithm to reject sequences that stabilize both the desired structure and partially unfolded states. Off-target ensembles are characterized by: RMSD values from crystal structure greater than 1 Å; only 50-90% of the initial secondary structure intact and surface-area comparable to the crystal structure.[120]

MSD approaches have also been further applied to stabilize transition states, essentially conferring motion to designed structures, in a design strategy dubbed meta-MSD. In this approach, geometry-based analysis of the rotamer-optimized structures allows for the classification of the microstates in either major, minor or transition states. The sequence optimization is then performed seeking the stabilization of all three states simultaneously, thus producing dynamic proteins with a low transition energy barrier between the two minor and major states.[121]

5.3) Flexible backbone approaches

Flexible backbone approaches share similarities with MSD strategies. However, in this case, the backbone configurations are not limited to a finite set on structures in an ensemble and is finely tuned using a set of movements instead (such as backrub, fragment reinsertion, dihedral and crankshaft movements, among others) or by

molecular dynamics, at the same time as different sequences are sampled. Although powerful, this approach is highly computationally expensive.[122] Software solutions that perform sidechain optimization using continuous rotamer searches include the OSPREY[123] and BKK*[124] packages, among others. More recently, strategies such as the proposed FlexiBal-GP attempt to incorporate information from multiple protein structures in data-driven machine learning movements in order to create complex movements that more closely mimic natural protein global motion.[125]

6. Synthesizing the designed protein

The last step of a protein design study is the correct synthesis of the computationally defined sequences. Traditionally, using cell-based approaches, this involved the incorporation of the protein-coding gene into the DNA of a host organism, selection of recombinant organisms and later purification of the culture medium or cell lysate [126]. This approach, however, has several setbacks, such as protein misfolding and aggregation due to cytoplasmatic homeostasis. Several studies have tackled this issued by rationally optimizing sequences to increase solubility and expression levels in cell-based synthesis models.[127] Modern synthesis of designed proteins overcomes this problem by employing cell-free protein synthesis (CFPS) and solid-phase peptide synthesis (SPPS) techniques.

The first approach of the CFPS methodology emerged in 1961, within the experiments carried out by Nirenberg and Matthaei.[128] This approach to protein and peptide synthesis attempts to assemble a minimalistic cell model, in a reaction medium containing all the biological machinery necessary for protein synthesis (usually obtained from purified cell lysates).[129] Besides that, an energy source (such as phosphoenol pyruvate), cofactors, buffers, salts, nucleotides and a supply of

aminoacids must be added to the CFPS mixture, so that the *in vitro* system closely resembles the cytoplasmic environment.[130] Once the DNA with the target genes is added, transcription and translation processes are simultaneously (coupled system) or sequentially performed (uncoupled/linked system) and the desired protein is produced.[131] The selection of CFPS reaction system depends, among other factors, on the source of cellular extract. It has been observed that one-pot reactions may result in suboptimal yields for eukaryotic platforms. In uncoupled systems, since transcription and translation reactions are operated separately (the already purified mRNA is transferred to another physical system), different conditions can be settled to achieve optimal protein yields, regardless of platform.[132] The type of translation reaction format also influences the protein yield. If a substrate-rich feeding buffer is continuously supplied to the mixture (continuous setting), the desired protein is filtered out at a constant rate and high protein concentrations are achieved;[133] if the protein synthesis is performed within a single tube (batch format), lower yields are expected.[132] In any case, since the CFPS methodology is performed in an open system format, active peptide yields can be maximized by selectively supplying the mixture with positive effectors, such as tRNA, chaperones and micelles. Analogously, negative effectors can be functionally inactivated.[130] Overall, this approach has several advantages, such as faster turnover times, possibility of implementation of unnatural aminoacids, tighter control of the medium properties and higher productivity, due to the inexistence of concentration induced toxicity effects on living cells.[134]

In 1963, Merrifield *et al*. developed the revolutionizing solid-phase peptide synthesis (SPPS) approach, later earning him the Nobel Prize in Chemistry.[135] In this methodology, N-end capped aminoacids are initially covalently bound to a supporting framework, with the help of linkers: specific organic moieties, such as trityl and

aminomethyl-based molecules. These linkers allow the reversible linkage between the peptide chain and the solid phase. Once the first aminoacid is attached, its carboxylic acid end is protected by the linker and side reactions are avoided. Therefore, unlike the natural protein synthesis, the solid phase peptide elongation is performed from the C- to the N-terminus, in rounds of chemical deprotection and aminoacid addition. The newly synthesized peptide is then detached from the solid support (uncoupled from the linker) and purified.[136] Some aminoacids (e.g.: arginine, histidine, aspartic and glutamic acids) have potentially reactive sidechains, which must be capped by specific sidechain protecting groups.[137] Since they generate carbocations and other reactive species when chemically deprotected during the cleavage treatment, water and other scavenger reagents are used to remove the undesired by-products. In sum, this approach shows high efficiency and throughput, allowing high protein yields and purity. However, SPPS is restricted to relatively short peptides owing to aggregation of longer chains.[138] Non-natural aminoacids can be also added[139] and exotic conformations (such as circular proteins) can be produced with relative ease.[140]

Altogether, cell-free and solid-phase peptide synthesis are emerging platforms with a huge role in future biomanufacturing, making possible the physical assembly of rationally designed proteins.

7. Conclusion and future perspectives

Recent developments in protein design have paved the way to the next generation of synthetic biology studies. Given the technical advances in computational power and recent new high throughput methodologies, both *in vitro* and *in silico*, the design of proteins with new unnatural conformations and expanded functionality are now ordinary procedures in medical and biochemical labs.[141] The design of tailored proteins

and peptides may prove to be the solution to a variety of contemporary problems: peptides, for example, are promising compounds for both antibiotic and anticarcinogenic usage.[142] Computational methods have shown incredible results as precise tools for screening and testing of new designs prior to experimental synthesis. Such contributions are only expected to increase in both quantity and quality with the implementation of modern computational tools such as new programming languages, neural networks and machine learning algorithms, especially considering the massive global effort in cataloguing and maintaining highly accurate and curated databases. Certain software packages, such as the Rosetta software, have forever changed the playing field in protein design, allowing access to powerful algorithms and methodologies to relatively inexperienced scientists. Long gone are the days of unreadable code in obscure languages. The undeniable success of Rosetta software and AlphaFold packages are, at least partially, due to the openness of the community in sharing resources in mediated forums, such as the CASP program, and therefore we foresee not only the increased participation of the scientific community in more open access sharing of software but also the emergence of modern and improved platforms specifically tailored for protein design.

References

1. Scott D, Berry D, Calvert J Synthetic biology. In: Routledge Handbook of Genomics, Health and Society. Elsevier; 2018. pp. 300–307.

2. Farny NG (2018) A vision for teaching the values of synthetic biology. Trends Biotechnol 36:1097–1100.

3. Gainza-Cirauqui P, Correia BE (2018) Computational protein design — the next generation tool to expand synthetic biology applications. Curr Opin Biotechnol 52:145–152.

4. Ponomarenko EA, Poverennaya E V, Ilgisonis E V, Pyatnitskiy MA, Kopylov AT, Zgoda VG, Lisitsa A V, Archakov AI (2016) The size of the human proteome: The width and depth. Int J Anal Chem 2016:7436849.

5. Chen X, Gonçalves MAFV (2018) DNA, RNA, and protein tools for editing the genetic information in human cells. iScience 6:247–263.

6. Dhanjal JK, Malik V, Radhakrishnan N, Sigar M, Kumari A, Sundar D. Computational Protein Engineering Approaches for Effective Design of New Molecules. In: Reference Module in Life Sciences. Academic Press; 2018. pp. 631–643.

7. Ljubetič A, Gradišar H, Jerala R (2017) Advances in design of protein folds and assemblies. Curr Opin Chem Biol 40:65–71.

8. Huang PS, Boyken SE, Baker D (2016) The coming of age of de novo protein design. Nature 537:320–327.

9. Makhlynets OV, Korendovych IV (2016) Minimalist design of allosterically regulated protein catalysts. Methods Enzymol 580:191–202.

10. Richardson JS, Richardson DC (1989) The de novo design of protein structures. Trends Biochem Sci 14:304–309.

11. Cobb RE, Chao R, Zhao H (2013) Directed evolution: Past, present, and future. AIChE J 59:1432–1440.

12. Ludwiczak J, Jarmula A, Dunin-Horkawicz S (2018) Combining Rosetta with molecular dynamics (MD): A benchmark of the MD-based ensemble protein design. J Struct Biol 203:54–61.

13. Allison B, Combs S, DeLuca S, Lemmon G, Mizoue L, Meiler J (2014) Computational design of protein-small molecule interfaces. J Struct Biol 185:193–202.

14. Lee J, Freddolino PL, Zhang Y Ab initio protein structure prediction. In: From Protein Structure to Function with Bioinformatics: Second Edition. Dordrecht: Springer Netherlands; 2017. pp. 3–35.

15. Desjarlais JR, Handel TM (1995) New strategies in protein design. Curr Opin Biotechnol 6:460–466.

16. Desmet J, De Maeyer M, Hazes B, Lasters I (1992) The dead-end elimination theorem and its use in protein side-chain positioning. Nature 356:539–542.

17. Holland JH Adaptation in natural and artificial systems: an introductory analysis. MIT Press; 1975.

18. Dahiyat BI, Mayo SL (1997) De novo protein design: fully automated sequence selection. Science 278:82–87.

19. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS (1998) High-resolution protein design with backbone freedom. Science 282:1462–1467.

20. Ponder JW, Richards FM (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 193:775–791.

21. Gordon DB, Marshall SA, Mayo SL (1999) Energy functions for protein design. Curr Opin Struct Biol 9:509–513.

22. Sung SS (1995) Folding simulations of alanine-based peptides with lysine residues. Biophys J 68:826–834.

23. Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. Proteins 23:ii–iv.

24. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225.

25. Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 37:171–176.

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

27. Peláez-Cid A-A, Herrera-González A-M, Salazar-Villanueva M, Bautista-Hernández A (2016) Elimination of textile dyes using activated carbons prepared from vegetable residues and their characterization. J Environ Manage 181:269–278.

28. Gainza P, Nisonoff HM, Donald BR (2016) Algorithms for protein design. Curr Opin Struct Biol 39:16–26.

29. Wang J, Cao H, Zhang JZH, Qi Y (2018) Computational protein design with deep learning neural networks. Sci Rep 8:6349.

30. Ghasemi F, Mehridehnavi A, Pérez-Garrido A, Pérez-Sánchez H (2018) Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks.

Drug Discov Today 23:1784–1790.

31. Tinberg CE, Khare SD (2017) Computational design of ligand binding proteins. Methods Mol Biol 1529:363–373.

32. Glusman G, Rose PW, Prlić A, Dougherty J, Duarte JM, Hoffman AS, Barton GJ, Bendixen E, Bergquist T, Bock C, et al. (2017) Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. Genome Med 9:113.

33. Sterner R, Merkl R, Raushel FM (2008) Computational design of enzymes. Chem Biol 15:421–423.

34. Heine A, DeSantis G, Luz JG, Mitchell M, Wong CH, Witson IA (2001) Observation of covalent intermediates in an enzyme mechanism at atomic resolution. Science 294:369–374.

35. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, et al. (2008) Kemp elimination catalysts by computational enzyme design. Nature 453:190–195.

36. Chen TS, Keating AE (2012) Designing specific protein-protein interactions using computation, experimental library screening, or integrated methods. Protein Sci 21:949–963.

37. Schreiber G, Fleishman SJ (2013) Computational design of protein–protein interactions. Curr Opin Struct Biol 23:903–910.

38. Examinatin G, Baker D, Stayton-bioengineering P, Thomas-bioengineering W (2012) Computational Design of Small Molecule Binding Proteins.

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

40. Śledź P, Caflisch A (2018) Protein structure-based drug design: from docking to molecular dynamics. Curr Opin Struct Biol 48:93–102.

41. Kumar A, Ranbhor R, Patel K, Ramakrishnan V, Durani S (2017) Automated protein design: Landmarks and operational principles. Prog Biophys Mol Biol 125:24–35.

42. Rajagopalan S, Wang C, Yu K, Kuzin AP, Richter F, Lew S, Miklos AE, Matthews ML, Seetharaman J, Su M, et al. (2014) Design of activated serine-containing catalytic triads with atomic-level accuracy. Nat Chem Biol 10:386–391.

43. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov M V., Renfrew PD, Mulligan VK, Kappel K, et al. (2017) The Rosetta all-atom energy function for macromolecular modeling and design. J Chem Theory Comput 13:3031–3048.

44. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. Proteins 35:133–152.

45. Shapovalov M V, Dunbrack RL (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19:844–858.

46. O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, Dimaio F, Bradley P, Kortemme T, Baker D, Snoeyink J, et al. (2015) Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput 11:609–622.

47. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, et al. (2013) Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol 523:109–143.

48. Park H, Bradley P, Greisen P, Liu Y, Mulligan VK, Kim DE, Baker D, Dimaio F (2016) Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. J Chem Theory Comput 12:6201–6212.

49. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov M V., Renfrew PD, Mulligan VK, Kappel K, et al. (2017) The Rosetta all-atom energy function for macromolecular modeling and design. J Chem Theory Comput 13:3031–3048.

50. Paladino A, Marchetti F, Rinaldi S, Colombo G (2017) Protein design: from computer models to artificial intelligence. Wiley Interdiscip Rev Comput Mol Sci 7:e1318.

51. Gomes J, Ramsundar B, Feinberg EN, Pande VS (2017) Atomic convolutional networks for predicting protein-ligand binding affinity. IEEE Trans Image Process 14:1360–1371.

52. Gao X, Ramezanghorbani F, Isayev O, Smith JS, Roitberg AE (2020) TorchANI: A free and open source PyTorch-based deep learning implementation of the ANI neural network potentials. J Chem Inf Model 60:3408–3415.

53. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, et al. (2020) Improved protein structure prediction using potentials from deep learning. Nature 577:706–710.

54. Xu J (2019) Distance-based protein folding powered by deep learning. Proc Natl Acad Sci USA 116:16856–16865.

55. Adhikari B (2020) A fully open-source framework for deep learning protein real-valued distances. Sci Rep 10:1–10.

56. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D (2020) Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad

Sci USA 117:1496–1503.

57. Gordon DE, Hiatt J, Bouhaddou M, Rezelj V V., Ulferts S, Braberg H, Jureka AS, Obernier K, Guo JZ, Batra J, et al. (2020) Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. Science 22:eabe9403.

58. Banerjee AK, Blanco MR, Bruce EA, Honson DD, Chen LM, Chow A, Bhat P, Ollikainen N, Quinodoz SA, Loney C, et al. (2020) SARS-CoV-2 disrupts splicing, translation, and protein trafficking to suppress host defenses. Cell 183:1–15.

59. Li Z, Hirst JD (2020) Computed optical spectra of SARS-CoV-2 proteins. Chem Phys Lett 758:137935.

60. Kar T, Narsaria U, Basak S, Deb D, Castiglione F, Mueller DM, Srivastava AP (2020) A candidate multi-epitope vaccine against SARS-CoV-2. Sci Rep 10:1–24.

61. Marze NA, Roy Burman SS, Sheffler W, Gray JJ (2018) Efficient flexible backbone protein-protein docking for challenging targets. Bioinformatics 34:3461–3469.

62. Jumper JM, Freed KF, Sosnick TR (2016) Rapid calculation of side chain packing and free energy with applications to protein molecular dynamics. PLoS Comp Biol 14:e1006342.

63. Yang W, Lai L (2017) Computational design of ligand-binding proteins. Curr Opin Struct Biol 45:67–73.

64. Starr TN, Thornton JW (2017) Exploring protein sequence–function landscapes. Nat Biotechnol 35:125–126.

65. Lucas JE, Kortemme T (2020) New computational protein design methods for de novo small molecule binding sites. PLoS Comput Biol 16:e1008178.

66. Ghirlanda G (2013) Computational biology: A recipe for ligand-binding proteins. Nature 501:177–178.

67. De Los Santos ELC, Meyerowitz JT, Mayo SL, Murray RM (2016) Engineering transcriptional regulator effector specificity using computational design and in vitro rapid prototyping: Developing a vanillin sensor. ACS Synth Biol 5:287–295.

68. Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics 10:168.

69. Guffy SL, Der BS, Kuhlman B (2016) Probing the minimal determinants of zinc binding with computational protein design. Protein Eng Des Sel 29:327–338.

70. Bhagavat R, Sankar S, Srinivasan N, Chandra N (2018) An augmented pocketome: Detection and analysis of small-molecule binding pockets in proteins of known 3D structure. Structure 26:499–512.

71. Fujieda N, Schätti J, Stuttfeld E, Ohkubo K, Maier T, Fukuzumi S, Ward TR (2015)

Enzyme repurposing of a hydrolase as an emergent peroxidase upon metal binding. Chem Sci 6:4060–4065.

72. Rivoire O, Reynolds KA, Ranganathan R (2016) Evolution-based functional decomposition of proteins. PLoS Comput Biol 12:1004817.

73. Süel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 10:59–69.

74. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286:295–299.

75. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R (2012) The spatial architecture of protein function and adaptation. Nature 491:138–142.

76. Raman AS, White KI, Ranganathan R (2016) Origins of allostery and evolvability in proteins: A case study. Cell 166:468–480.

77. Lee J, Goodey NM (2011) Catalytic contributions from remote regions of enzyme structure. Chem Rev 111:7595–7624.

78. Yang W, Lai L (2017) Computational design of ligand-binding proteins. Curr Opin Struct Biol 45:67–73.

79. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: Evolutionary units of three-dimensional structure. Cell 138:774–786.

80. Banda-Vázquez J, Shanmugaratnam S, Rodríguez-Sotres R, Torres-Larios A, Höcker B, Sosa-Peinado A (2018) Redesign of LAOBP to bind novel l-amino acid ligands. Protein Sci 27:957–968.

81. Morcos F, Hwa T, Onuchic JN, Weigt M (2014) Direct coupling analysis for protein contact prediction. Methods Mol Biol 1137:55–70.

82. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 108:E1293-E1301.

83. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 106:67–72.

84. Zerihun MB, Schug A (2017) Biomolecular coevolution and its applications: Going from structure prediction toward signaling, epistasis, and function. Biochem Soc Trans 45:1253–1261.

85. Peng M, Maier M, Esch J, Schug A, Rabe KS (2019) Direct coupling analysis improves the identification of beneficial amino acid mutations for the functional

thermostabilization of a delicate decarboxylase. Biol Chem 400:1519–1527.

86. Lechner H, Ferruz N, Höcker B (2018) Strategies for designing non-natural enzymes and binders. Curr Opin Chem Biol 47:67–76.

87. Churchfield LA, George A, Tezcan FA (2017) Repurposing proteins for new bioinorganic functions. Essays Biochem 61:245–258.

88. Zhou L, Bosscher M, Zhang C, Özçubukçu S, Zhang L, Zhang W, Li CJ, Liu J, Jensen MP, Lai L, et al. (2014) A protein engineered to bind uranyl selectively and with femtomolar affinity. Nat Chem 6:236–241.

89. Moroz YS, Dunston TT, Makhlynets O V., Moroz O V., Wu Y, Yoon JH, Olsen AB, McLaughlin JM, Mack KL, Gosavi PM, et al. (2015) New tricks for old proteins: Single mutations in a nonenzymatic protein give rise to various enzymatic activities. J Am Chem Soc 137:14905–14911.

90. Taylor ND, Garruss AS, Moretti R, Chan S, Arbing MA, Cascio D, Rogers JK, Isaacs FJ, Kosuri S, Baker D, et al. (2016) Engineering an allosteric transcription factor to respond to new ligands. Nat Methods 13:177–183.

91. Wijma HJ, Floor RJ, Bjelic S, Marrink SJ, Baker D, Janssen DB (2015) Enantioselective enzymes by computational design and in silico screening. Angew Chem Int Ed Engl 54:3726–3730.

92. Mackenzie CO, Grigoryan G (2017) Protein structural motifs in prediction and design. Curr Opin Struct Biol 44:161–167.

93. Khersonsky O, Fleishman SJ (2016) Why reinvent the wheel? Building new proteins based on ready-made parts. Protein Sci 25:1179–1187.

94. Brunette TJ, Parmeggiani F, Huang PS, Bhabha G, Ekiert DC, Tsutakawa SE, Hura GL, Tainer JA, Baker D (2015) Exploring the repeat protein universe through computational protein design. Nature 528:580–584.

95. Eisenbeis S, Proffitt W, Coles M, Truffault V, Shanmugaratnam S, Meiler J, Höcker B (2012) Potential of fragment recombination for rational design of proteins. J Am Chem Soc 134:4019–4022.

96. Jacobs TM, Williams B, Williams T, Xu X, Eletsky A, Federizon JF, Szyperski T, Kuhlman B (2016) Design of structurally distinct proteins using strategies inspired by evolution. Science 352:687–690.

97. Correia BE, Ban YEA, Holmes MA, Xu H, Ellingson K, Kraft Z, Carrico C, Boni E, Sather DN, Zenobia C, et al. (2010) Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. Structure 18:1116–1126.

98. Correia BE, Bates JT, Loomis RJ, Baneyx G, Carrico C, Jardine JG, Rupert P, Correnti C, Kalyuzhniy O, Vittal V, et al. (2014) Proof of principle for epitope-focused

vaccine design. Nature 507:201–206.

99. Lin Y-R, Koga N, Tatsumi-Koga R, Liu G, Clouser AF, Montelione GT, Baker D (2015) Control over overall shape and size in de novo designed proteins. Proc Natl Acad Sci USA 112:E5478–E5485.

100. Agah S, Poulos S, Yu A, Kucharska I, Faham S (2016) Protein rethreading: A novel approach to protein design. Sci Rep 6:26847.

101. Thomson AR, Wood CW, Burton AJ, Bartlett GJ, Sessions RB, Brady RL, Woolfson DN (2014) Computational design of water-soluble α-helical barrels. Science 346:485–488.

102. Burton AJ, Thomson AR, Dawson WM, Brady RL, Woolfson DN (2016) Installing hydrolytic activity into a completely de novo protein framework. Nat Chem 8:837–844.

103. Olson TL, Espiritu E, Edwardraja S, Canarie E, Flores M, Williams JAC, Ghirlanda G, Allen JP (2017) Biochemical and spectroscopic characterization of dinuclear Mn-sites in artificial four-helix bundle proteins. Biochim Biophys Acta Bioenerg 1858:945–954.

104. MacDonald JT, Kabasakal B V, Godding D, Kraatz S, Henderson L, Barber J, Freemont PS, Murray JW (2016) Synthetic beta-solenoid proteins with the fragment-free computational design of a beta-hairpin extension. Proc Natl Acad Sci 113:10346–10351.

105. Ennist NM, Mancini JA, Auman DB, Bialas C, Iwanicki MJ, Esipova T V., Discher BM, Moser CC, Dutton PL (2017) Maquette strategy for creation of light- and redox-active proteins. In: Photosynthesis and Bioenergetics. World Scientific, pp. 1–33.

106. Watkins DW, Jenkins JMX, Grayson KJ, Wood N, Steventon JW, Le Vay KK, Goodwin MI, Mullen AS, Bailey HJ, Crump MP, et al. (2017) Construction and in vivo assembly of a catalytically proficient and hyperthermostable de novo enzyme. Nat Commun 8:358.

107. Grayson KJ, Anderson JR (2018) The ascent of man(made oxidoreductases). Curr Opin Struct Biol 51:149–155.

108. Riazanov A, Karasikov M, Grudinin S (2016) Inverse protein folding problem via quadratic programming. In: Information Technology and Systems, pp. 561–568.

109. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 97:10383–10388.

110. Jia L, Yarlagadda R, Reed CC (2015) Structure based thermostability prediction models for protein single point mutations with machine learning tools. PLoS One 10:e0138022.

111. Farhoodi R, Shelbourne M, Hsieh R, Haspel N, Hutchinson B, Jagodzinski F Predicting the Effect of Point Mutations on Protein Structural Stability. In: Proceedings

of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17. New York, New York, USA: ACM Press; 2017. pp. 247–252.

112. Siderius M, Jagodzinski F (2018) Mutation sensitivity maps: Identifying residue substitutions that impact protein structure via a rigidity analysis in silico mutation approach. J Comput Biol 25:89–102.

113. Witvliet DK, Strokach A, Giraldo-Forero AF, Teyra J, Colak R, Kim PM (2016) ELASPIC web-server: Proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. Bioinformatics 32:1589–1591.

114. Wainreb G, Wolf L, Ashkenazy H, Dehouck Y, Ben-Tal N (2011) Protein stability: A single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. Bioinformatics 27:3286–3292.

115. Krivov GG, Shapovalov M V., Dunbrack RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77:778–795.

116. Gaillard T, Panel N, Simonson T (2016) Protein side chain conformation predictions with an MMGBSA energy function. Proteins 84:803–819.

117. Davey JA, Chica RA (2012) Multistate approaches in computational protein design. Protein Sci 21:1241–1252.

118. Davis IW, Arendall WB, Richardson DC, Richardson JS (2006) The backrub motion: How protein backbone shrugs when a sidechain dances. Structure 14:265–274.

119. Davey JA, Chica RA (2014) Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. Proteins 82:771–784.

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

121. Davey JA, Damry AM, Goto NK, Chica RA (2017) Rational design of proteins that exchange on functional timescales. Nat Chem Biol 13:1280–1285.

122. MacDonald JT, Freemont PS (2016) Computational protein design with backbone plasticity. Biochem Soc Trans 44:1523–1529.

123. Ojewole A, Lowegard A, Gainza DP, Reeve SM, Georgiev I, Anderson AC, Donald BR (2017) OSPREY predicts resistance mutations using positive and negative computational protein design. Methods Mol Biol 1529:291–306.

124. Ojewole AA, Jou JD, Fowler VG, Donald BR BBK* (Branch and bound over K*): A provable and efficient ensemble-based algorithm to optimize stability and binding affinity over large sequence spaces. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

Vol. 10229 LNCS. Springer, Cham; 2017. pp. 157–172.

125. Sun MGF, Kim PM (2017) Data driven flexible backbone protein design. PLoS Comput Biol 13:e1005722.

126. Yeo J, Jung GS, Martín-Martínez FJ, Ling S, Gu GX, Qin Z, Buehler MJ (2018) Materials-by-design: Computation, synthesis, and characterization from atoms to structures. Phys Scr 93:053003.

127. Sormanni P, Aprile FA, Vendruscolo M (2015) The CamSol method of rational design of protein mutants with enhanced solubility. J Mol Biol 427:478–490.

128. Nirenberg MW, Matthaei JH (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci USA 47:1588–1602.

129. Jia H, Heymann M, Bernhard F, Schwille P, Kai L (2017) Cell-free protein synthesis in micro compartments: building a minimal cell from biobricks. N Biotechnol 39:199–205.

130. Jin X, Hong SH (2018) Cell-free protein synthesis for producing 'difficult-to-express' proteins. Biochem Eng J 138:156–164.

131. Georgi V, Georgi L, Blechert M, Bergmeister M, Zwanzig M, Wüstenhagen DA, Bier FF, Jung E, Kubick S (2016) On-chip automation of cell-free protein synthesis: New opportunities due to a novel reaction mode. Lab Chip 16:269–281.

132. Gregorio NE, Levine MZ, Oza JP (2019) A user's guide to cell-free protein synthesis. Methods Protoc 2:24.

133. Spirin AS, Baranov VI, Ryabova LA, Ovodov SY, Alakhov YB (1988) A continuous cell-free translation system capable of producing polypeptides in high yield. Science 242:1162–1164.

134. Dopp BJL, Tamiev DD, Reuel NF (2018) Cell-free supplement mixtures: Elucidating the history and biochemical utility of additives used to support in vitro protein synthesis in E. coli extract. Biotechnol Adv 37:246–258.

135. Merrifield RB (1963) Solid phase peptide synthesis. I. The synthesis of a tetrapeptide. J Am Chem Soc 85:2149–2154.

136. Petrou C, Sarigiannis Y (2017) Peptide synthesis: Methods, trends, and challenges. In: Peptide Applications in Biomedicine, Biotechnology and Bioengineering. Woodhead Publishing; pp. 1–21.

137. Caporale A, Doti N, Monti A, Sandomenico A, Ruvo M (2018) Automatic procedures for the synthesis of difficult peptides using oxyma as activating reagent: A comparative study on the use of bases and on different deprotection and agitation conditions. Peptides 102:38–46.

138. Paradís-Bas M, Tulla-Puche J, Albericio F (2016) The road to the synthesis of “difficult peptides.” Chem Soc Rev 45:631–654.

139. O’Donnell MJ, Zhou C, Scott WL (1996) Solid-phase unnatural peptide synthesis (UPS). J Am Chem Soc 118:6070–6071.

140. Angell YM, Thomas TL, Flentke GR, Rich DH (1995) Solid-phase synthesis of cyclosporin peptides. J Am Chem Soc 117:7279–7280.

141. Gainza-Cirauqui P, Correia BE (2018) Computational protein design — the next generation tool to expand synthetic biology applications. Curr Opin Biotechnol 52:145–152.

142. Torres MDT, Sothiselvam S, Lu TK, de la Fuente-Nunez C (2019) Peptide design principles for antimicrobial applications. J Mol Biol 431:3547-3567.

FIGURE LEGENDS

Figure 1. Growth of the protein design field of study, by number of publications. The data shown refers to the number of publications listed in ScienceDirect repository when querying for “Protein Design”, from 1998 until 2020.

Figure 2. Schematic overview of the main improvements in the field of computational protein design. A – 1995 – Sung *et al* studied the secondary structure folding of simple alanine-based peptides. B – 1997 – Several improvements in forcefield definition and the implementation of more modern DEE and genetic algorithms allowed Dahiyat *et al* to design a de-novo sequence for a simple ββα-motif with relative accuracy (PDB 1FSD). C – 2003 – The first applications of homology-based algorithms started to appear, such as the Rosetta software, allowing Kuhlman *et al* to design a previously inexistent fold dubbed Top7 (PDB 1QYS). D – 2008 - Röthlisberger *et al* designed an inexistent enzyme for the catalysis of the Kemp elimination reaction (PDB 2RKX).

Figure 3. Simplistic approach to protein design. Using Rajagolapan *et al* work as an example [42]; A – Design of a novel binding site: the interaction site (in green) was rationally idealized in order to stabilize fluorphosphonate-alkalyne probes (in pink); B – Scaffold design or adaptation: The designed binding site was computationally grafted to a set of 800 pre-existing protein scaffolds using RosettaMatch algorithm, of which 380 were accepted with relative confidence; C – Identification or mutation of an aminoacid sequence that folds into the desired conformation: Three rounds of sequence optimization were performed for each of the candidate structures, minimizing

the energy of the system. From the initial set, 85 candidates showed relatively good stability and were chosen for experimental characterization.

Figure 4. Schematic representation of the protein folding funnel according to different scoring function types. Data shown corresponds to the inverted topology map of Mt. Everest, for exemplification purposes. A- All-atom scoring functions take all atoms in the system into consideration, allowing for a greater accuracy level. B – Coarse-grain models reduce the degrees of freedom of the system by considering a smaller number of interacting particles, at the expense of reduced accuracy.

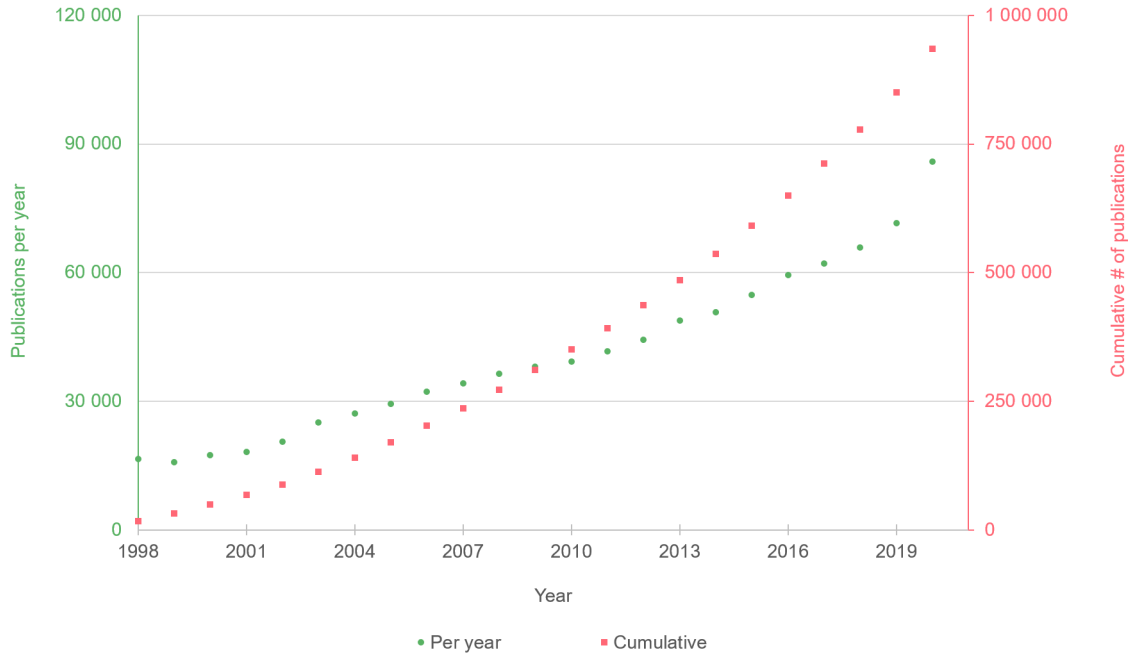
Figure 5. Binding pocket design approaches. Proteins depicted (in green) serve as examples of caffeine (in red) binding pockets. A – Stabilization of a binding shell: potential non-covalent forces are identified in the ligand, and complementary aminoacids are rationally placed in a theoretical binding pocket as to stabilize these forces. a – hydrogen bonds; b – aromatic interaction (π-π stacking); c – electrostatic interactions. B – Blind pocket search: large databases of proteins/binding pockets are blindly searched for complementarity towards the target ligand. Promising hits can be further optimized.

Figure 6. Protein scaffold design approaches. A – Repurpose of an existing structure. A protein of unknown function from *Methanobacterium thermoautrophicum* (PDB 2PMR) was repurposed by mutating residues Leu13, His64 and Leu67 to Asp, Glu and Thr, respectively, augmenting the binding affinity of the designed protein towards uranyl ions. B – Fragment-based approach. A triple α-helix fragment from ribosome recycling factor (*Vibrio parahaemolyticus*, PDB 1IS1) was grafted to the 4E10 Peptide, an epitope for HIV-1 gp41 glycoprotein (PDB 3IXT), resulting in a designed non-viral

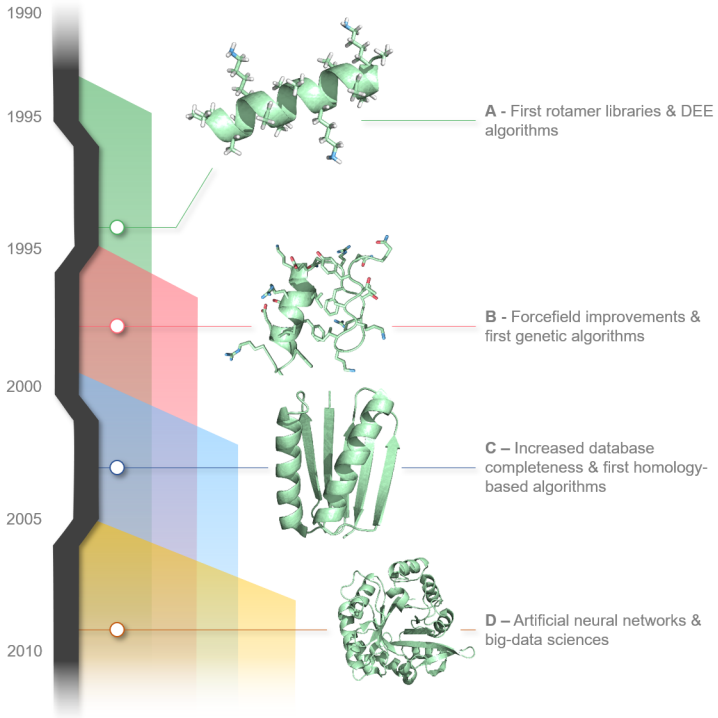
protein capable of vaccine-inducing neutralizing activity. C – *De novo* design of protein scaffolds. An α-helix monomer was designed by rationally placing hydrophobic and hydrophilic residues in opposite sides of the theorical scaffold. The designed protein is stabilized in an heptameric coiled-coil. Further mutations in Ile18, Leu22 and Ile25 to Cys, His and Glu, respectively, induce catalytical hydrolysis efficiencies towards p-nitrophenyl acetate.

Figure 7. Protein sequence design approaches. Schematically organized from left to right with increasing computational cost and complexity. A – Fixed backbone approach: The identity/orientation of one or more aminoacids of the initial protein/desired scaffold crystal structure is mutated, and the new sequence energy is compared with the starting state (black dashed line). In this example, the sequence *b* (in green) shows lower energy/higher stability than sequence *a* (in red). B – Positive multistate design: Multiple microstates with on-target variations of the starting structure are evaluated simultaneously (thin lines), and the energy value used for comparison with the initial state is a weighted average of the whole ensemble (thick lines). In this example, even though some microstates of the sequence *b* ensemble show lower energy than some microstates of sequence *a* ensemble (similarly to the fixed backbone approach), overall sequence *a* shows greater stability. C – Negative multistate approach: Complementary to the positive multistate approach, a set of on-target microstates are evaluated simultaneously to determine the energy of the sampled sequence. However, in this case, the comparison point is not the initial crystal structure, but an ensemble of off-target structures (darker shades). In this example, even though sequence *a* seems to better stabilize the desired structure, it also lowers the energy on undesired off-target structures, while sequence *b* destabilizes them. This reflects in higher expression

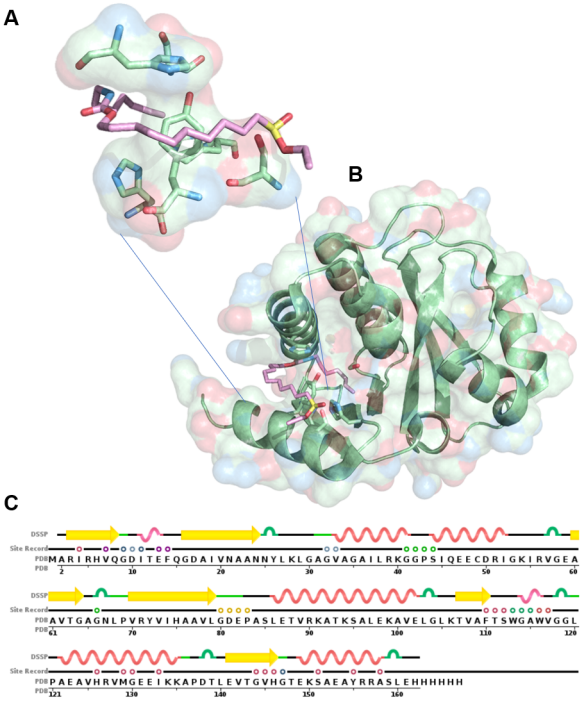
probability in experimental essays. D – Flexible backbone approaches continuously sample both the conformational and sequence space simultaneously. In this simplified example, during the produced trajectory, a sequence *c* was found that was more stable than both sequences *a* and *b* and that folds to an on-target structure.



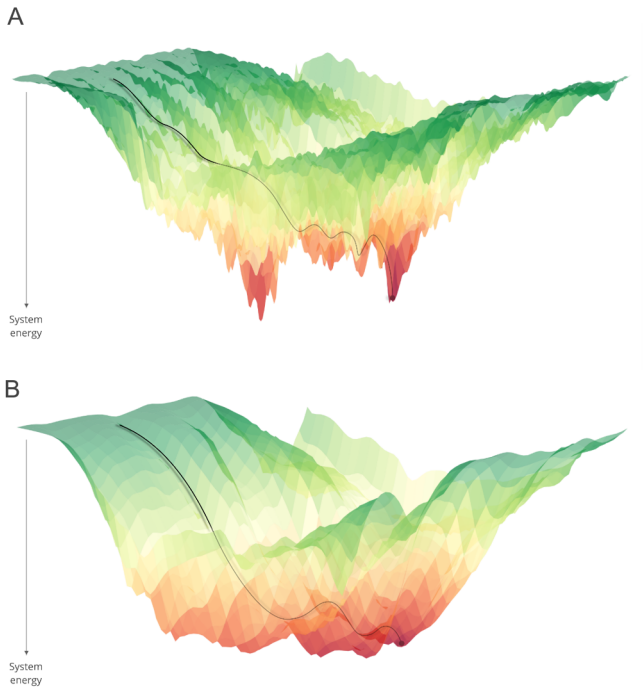
PRO_4098_figure-1.tif



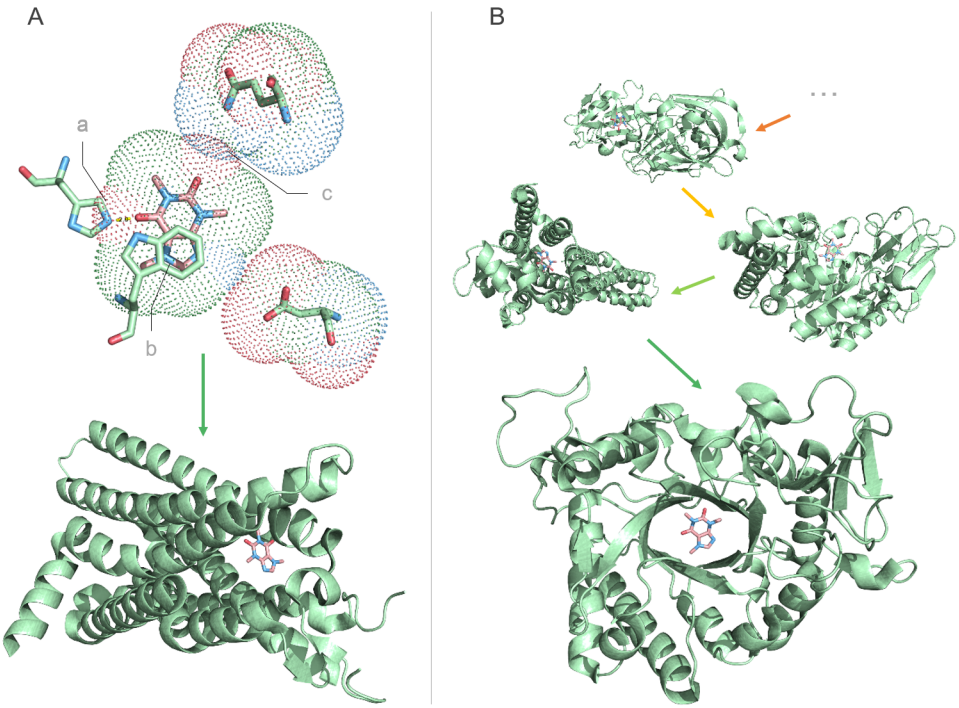
PRO_4098_figure-2.tif



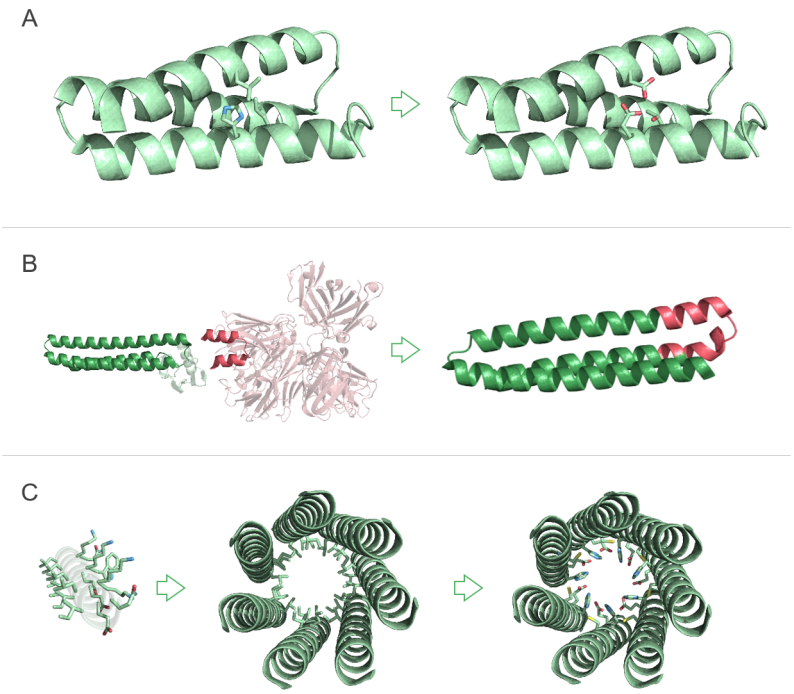
PRO_4098_figure-3.tif



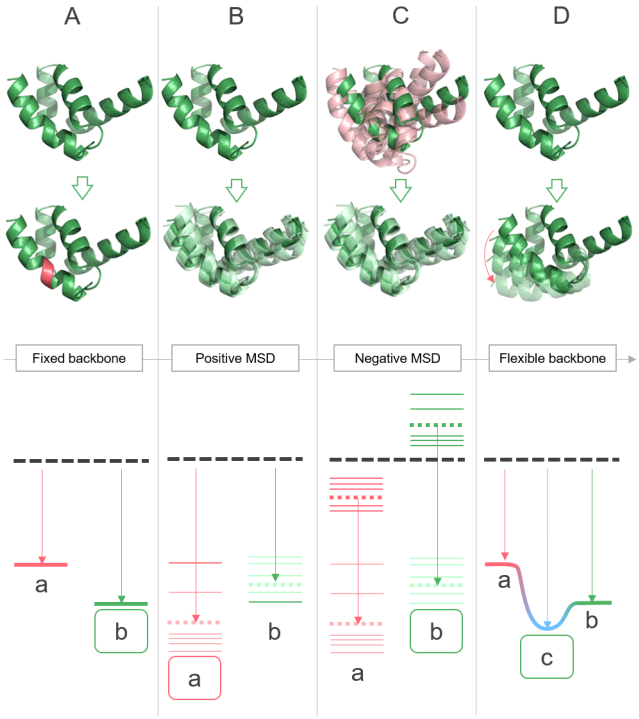
PRO_4098_figure-4.tif



PRO_4098_figure-5.tif



PRO_4098_figure-6.tif



PRO_4098_figure-7.tif

	Design approach					
Main author	Scaffold	Binding site	Sequence	Expression	Mutagenesis	Ref.
Rajagolapan *et al*. (2014)	SR	BSS	SSD	Cell-based	Yes	[42]
Tinberg *et al*. (2013)	SR	BSS	FB	Cell-based	Yes	[39]
De Los Santos *et al*. (2016)	SR	BSS	MSD	Cell-free	No	[67]
Guffy *et al*. (2016)	SR	BSS	SSD	Cell-based	No	[69]
Zhou *et al*. (2014)	SR	BSS	-	Call-based	Yes	[88]
Fujieda *et al*. (2015)	SR	PS	SSD	Cell-based	Yes	[71]
Banda-Vásquez *et al*. (2018)	SR	PS[a]	SSD	Cell-based	No	[80]
Moroz *et al*. (2015)	SR	Targeted	SSD	Cell-based	Yes	[89]
Taylor *et al*. (2016)	SR	Targeted	FB	Cell-based	Yes	[90]
Wijma *et al*. (2015)	SR	Targeted	FB	Cell-based	No	[91]
Correia *et al*. (2014)	FR	Targeted	FB	Cell-based	No	[98]
Eisenbeis *et al*. (2012)	FR	-	SSD	Cell-based	No	[95]
Jacobs *et al*. (2016)	FR	-	SSD	Cell-based	No	[96]
Brunette *et al*. (2015)	FR	-	MSD	Cell-based	No	[94]
Agah *et al*. (2016)	FR	-	-	Cell-based	No	[100]
Thomson *et al*. (2014)	DND	-	SSD	SPPS	No	[101]
MacDonald *et al*. (2016)	DND	-	SSD	Cell-based	No	[104]
Burton *et al*. (2016)	DND	BSS	SSD	SPPS	No	[102]
Olson *et al*. (2017)	DND	BSS	-	Cell-based	No	[103]
Watkins *et al*. (2017)	DND[b]	-	-	Cell-based	Yes	[106]

[a] – Pocket search was carried out with statistical coupling analysis (SCA)

[b] – De-novo design of a maquette scaffold for heme-group incorporation

Legend:

SR – Structure repurpose

FR – Fragment recombination

DND – De-novo design

BSS – Blind shell stabilization

PS – Pocket search

SSD – Single state design

MSD – Multiple state design

FB – Flexible backbone

SPPS – Solid-phase peptide synthesis