Computational design of proteins and enzymes

Vaitea Opuu

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAX081

Thèse de doctorat

Computational design of proteins and enzymes

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Biologie

Thèse présentée et soutenue à Palaiseau, le 29/10/2020, par

VAITEA OPUU

Composition du Jury :

Sebastian Will Professeur, École Polytechnique (UMR 7161)	Président
Anne-Claude Camproux Professeur, Université Paris-Diderot (UMR-S 973)	Rapporteur
Marc Delarue Directeur de recherche, Institut Pasteur (UMR 3528)	Rapporteur
Martin Weigt Professeur, Sorbonne Université (UMR 7238)	Examinateur
Thomas Simonson Professeur, École Polytechnique (UMR 7654)	Directeur de thèse

Contents

Remerciements

Je souhaite en premier lieu remercier mon directeur scientifique et mentor *Thomas Simonson* sans qui l'expérience de la thèse n'aurait pas été aussi enrichissante et productive. Un directeur avisé, disponible et rigoureux, ce sont là quelques unes des qualités de mon directeur qui nous ont permis d'atteindre le niveau de production détaillé dans le présent manuscrit. Plus qu'une aventure scientifique, la thèse a été une expérience de vie. J'ai par ailleurs eu la chance d'être initié à l'aïkido scientifique, une discipline que l'étudiant de thèse se doit de maîtriser. Pour ces expériences uniques, merci!

Je remercie les rapporteurs *Anne-Claude Camproux* et *Marc Delarue* ainsi que les examinateurs *Martin Weigt* et *Sebastian Will* pour l'évaluation ce travail malgré les conditions sanitaires actuelles.

Je remercie nos nombreux collaborateurs expérimentaux. Merci à *Yves* et *Emmanuelle* pour les travaux MetRS. Merci à *Giuliano* et *Christine* qui ont produit les résultats expérimentaux MetRS présentés dans ce manuscrit. Je remercie nos collaborateurs de l'Université de l'Iowa du groupe d'*Enersto J. Fuentes* pour les tests expérimentaux PDZ. Je remercie notre collaborateur *Gilles Travé* pour les tests expérimentaux en cours pour les PDZ chevauchants.

Je remercie mes compagnons de thèse *Francesco* et *Nicolas* avec qui nous avons partagé autant de discussions scientifiques que récréatives. Vivement nos prochaines pauses cafés. Je remercie *Maxime I.* pour nos folles aventures et pauses cafés. Je remercie aussi les membres actuels du groupe de bioinformatique, *Paula* (thank you for the english courses, it turned out quite useful at the end), *Ivan*, *Xingyu*, *Valentin*. Je remercie Thomas G. et David pour leurs conseils, aide et discussions qui m'ont permis de mieux appréhender le CPD. Je remercie aussi les membres passés *Rojo*, *Alexandrine*, *Hélène*, *Théo*, *Marie-Pierre*. Je remercie *Clara*, *Marc D.* et *Pierre P.* pour nos conversations aux repas au cours de ces dernières années. Merci aussi aux autres membres avec qui j'ai pu discuter: *Ramy*, *Irène*, *Myriam*, *Pierre-damien*, *Guillaume*,

Gabrielle, *Marc G.*, *Nathalie U.*. Je remercie aussi le secrétariat du laboratoire, *Mélanie* et *Lydia*, pour leur aide.

Je souhaite aussi remercier les personnes qui ont pavé mon chemin jusqu'ici. Je remercie *Anne Lopes* grâce à qui j'ai dérivé dans le domaine de la structure protéique. Je remercie *Tamatoa Bambridge* qui m'a fait découvrir la recherche ainsi que pour son aide déterminante.

Je remercie aussi les amis de Tahiti grâce à qui j'ai pu garder le contact avec l'île: *Hawaiki*, *Tama*, *Manu*...

Finalement, je remercie la famille. Je remercie mes parents, *Murna* et *Tony*, pour leur soutien m'ayant mené jusqu'au bout du monde et va me pousser encore plus loin. Je remercie aussi mes frères *Dany* et *Rainui*. Je remercie aussi ma belle famille *Béatrice*, *Brigitte*, *Jean-Louis*, *Jean-Charles* qui m'ont accueilli. Je tiens à remercier *Marie*, sans qui rien de tout ceci n'aurait eu lieu, pour ton soutien (mis à rude épreuve lors des corrections de ce manuscrit, je l'imagine).

MĀURUURU ROA!

Résumé

Le but de l'ingénierie numérique de protéines (ou *Computational Protein Design* CPD) est de construire des systèmes moléculaires capables d'accomplir des fonctions biologiques. Pour concevoir ce type de systèmes, nous utilisons comme modèles des systèmes optimisés naturellement, les protéines. Ainsi, il est possible de construire ces machines moléculaires en utilisant la machinerie de traduction biologique. L'approche générale consiste à utiliser le paradigme reliant la structure tridimensionnelle d'une protéine à sa fonction. Ces approches ont déjà prouvé leur efficacité, par exemple avec la production d'enzymes capables de digérer du plastique ([Tournier et al., 2020]). Néanmoins, les techniques et les principes fondamentaux liés à la conception de tels systèmes n'en sont qu'à leur début. En effet, une récente étude à grande échelle montre que le taux de réussite de ce type d'approches est de 6% ([Rocklin et al., 2017]).

Dans ce travail de thèse, nous avons étudié plusieurs aspects de l'ingénierie de protéines. Nous avons d'abord entièrement redessiné un domaine PDZ impliqué dans de nombreuses voies métaboliques. Nous utilisons une approche *physics-based* basée sur la mécanique moléculaire, un modèle de solvant implicite et un échantillonnage Monte Carlo ([Mignon et al., 2020]). Parmi plusieurs milliers de variants prédits pour adopter le repliement PDZ, trois ont été sélectionnés et montrent expérimentalement un repliement correct. Deux ont une affinité détectable pour les ligands peptidiques naturels. Ce travail permet d'étayer l'utilisation des principes fondamentaux quand la stratégie actuelle tend à s'appuyer sur des descriptions statistiques. Cette étude a ainsi montré le premier succès de l'utilisation d'une fonction *physics-based* pour une application de cette taille ([Opuu et al., 2020b]).

Nous avons ensuite étudié l'aspect catalytique au travers du redessin du site actif d'une enzyme impliquée dans le mécanisme de traduction génétique, la Méthionyl-ARNt synthétase (MetRS) ([Opuu et al., 2020a]). Ce travail s'inscrit dans un projet d'expansion du code génétique à des acides aminés non naturels. Dans ce travail, nous avons modifié la Methionyl-ARNt

synthétase pour modifier son activité de catalyse avec la Méthionine. Nous avons utilisé un nouveau paradigme de dessin basé sur l'état de transition (TS) de la réaction de catalyse. Nous avons ainsi démontré l'efficacité de cette approche pour le redessin d'enzymes, tout d'abord pour le ligand naturel. D'autre part, nous avons retrouvé des résultats expérimentaux pour la catalyse d'un ligand non-naturel, l'azidonorleucine (AnL).

Puis, nous avons étudié la possibilité de modifier la MetRS pour étendre son activité aux acides aminés β, afin d'étendre le code génétique. Ces acides aminés non-naturels permettraient d'enrichir le répertoire structural des protéines. 20 variants MetRS obtenus à partir de prédictions d'affinité MetRS/β-Met ont été testés. 5 variants actifs ont été détectés mais aucun n'augmente l'activité. Toutefois, trois ont amélioré la sélectivité en faveur de la β-Met par des facteurs de 2 à 8. Pour explorer l'espace de mutation de manière systématique, nous avons implémenté une méthode de sélection de positions d'intérêt et production de variants pour β-Met et β-Val. Une vingtaine de prédictions sont en cours de tests expérimentaux.

Pour finir, nous avons étudié la possibilité de créer des paires de domaines PDZ avec une contrainte de codage chevauchant. Le codage chevauchant est une stratégie de codage exploitée par tous les domaines de la Vie mais plus particulièrement par les virus. D'un point de vue biotechnologique, cette stratégie de codage permet de mettre en place des stratégies de bio-confinement nécessaires pour l'exploitation de génomes modifiés. Dans ce travail, nous avons utilisé un algorithme de programmation dynamique développé récemment permettant de concevoir des paires de gènes codées de manière chevauchante sur la base de protéines modèles. Près de 2000 paires de séquences PDZ chevauchantes ont été calculées. 3 paires ont été choisies pour des validations numériques par dynamique moléculaire tout-atomes. Une paire a été validée par 3 microsecondes de dynamique moléculaire. Des tests expérimentaux sont en cours.

Chapter 1

Introduction

The purpose of Computational Protein Design (CPD) is to build molecular systems capable of biological functions. To design such systems, we use as models naturally optimized components, proteins. Thus, it is possible to produce these molecular components using the biological translation machinery. The general approach is to use the paradigm connecting the three-dimensional structure of a protein to its biological function. CPD has already shown some successes. For example, an enzyme capable of digesting plastics was recently reported ([Tournier et al., 2020]). Nevertheless, techniques and fundamental principles related to the design of such machines are only in the early stage. Indeed, a recent large study shows that the success rate of these approaches for whole protein redesign is around 6% ([Rocklin et al., 2017]).

In this work, we studied several aspects of CPD, from numerical predictions to experimental testing. First, we completely redesigned a protein domain involved in protein-protein interactions. For this study, we used techniques based on physics to sample protein variants according to their stability. We call this approach physics-based or ϕ-CPD ([Mignon et al., 2020]). We selected three variants for experimental validation. All three variants showed evidence of adopting the correct fold. Moreover, two variants showed a binding affinity for natural peptide ligands. This work supports the use of physical principles, in contrast to the current strategy which relies heavily on statistical information. This study represents the first successful use of a physics-based energy function for a whole protein redesign of this size ([Opuu et al., 2020b]).

Next, we will present work on the engineering of an enzyme involved in the translation machinery, methionyl tRNA synthetase (MetRS) ([Opuu et al., 2020a]). In this work, we redesigned MetRS for its activation reaction. We used a recently developed method that allows

the sampling of variants rigorously on their binding free energy. We were able to select variants that were then shown to be active in experiments. We also performed design calculations for the activation of azidonorleucine, a Methionine analog, where we recovered experimentally known variants. Then, we extended the design method to the binding of a transition state ligand. Therefore, we were able to sample MetRS variants according to their catalytic power for the Met substrate.

In a third chapter, we will present the redesign of MetRS for the activation of two β amino acids. First, we searched for variants where only three positions were allowed to vary. Three variants were shown experimentally to have a slightly improved selectivity for β-Met compared to α-Met. The improvement factors were 2-8. However, the β-Met activity was not improved. To go further, we explored additional positions in the active site. We introduced a new method to select positions for design according to their binding potential. We computationally redesigned eight positions in search of β-Met and β-Val activity. Some variants from these predictions are in an experimental testing phase (Y. Mechulam, E. Schmitt, personal communication).

Finally, we report the design of two pairs of PDZ domains with a fully overlapping coding scheme at the DNA level. Overlapping coding is a strategy used especially by viruses. From a biotechnological perspective, this coding can help in the bio-containment of modified genomes. In this work, we used an algorithm we developed earlier for designing pairs of overlapping genes, based on protein sequences. First, we designed almost two thousand pairs of PDZ proteins encoded in an overlapping fashion. Then, three of the designs were selected for molecular dynamic simulation testing. Two were found stable during simulations of at least 500 ns. One pair was validated by simulations of 3 μs. Experimental testing is underway (G. Travé, personal communication).

In this chapter, we recall some biological and structural aspects of the PDZ protein family. We detail the specificity of PDZ domains and some experimental and computational studies. Then, we present the aminoacyl-tRNA synthetase (AARS) structures and some associated studies. Finally, we present technical and theoretical aspects of CPD with special attention to ϕ-CPD using the Proteus software ([Mignon et al., 2020]).

1.1 PDZ domains

PDZ domains (postsynaptic protein-95 Disk wide Zonula occludens-1) are recognition domains involved in protein-protein interactions. This is a ubiquitous family that can be found several times in the same protein. These domains bind specifically their protein partners. The engineering of such domains may enable the engineering of biological networks.

1.1.1 Three-dimensional structure of a PDZ domain

PDZ domains are small globular domains of about 90 amino acids. They are composed of five to six β sheets numbered β_1-β_6 and two α helices (α_1 and α_2). PDZ domains are usually paired or grouped in modules. The PDZ domains recognize a short sequence (or motif) of four to seven terminal amino acids. This recognition occurs between the β_2 and β_3 strands (figure 1.1). The specificity of the PDZ domains allows biochemical messages in signaling pathways. Therefore, one can disrupt these signaling pathways, *e.g.* by using molecules that inhibit the PDZ domain recognition. This approach was used by ([Thorsen et al., 2009]). They identified a molecule that inhibits specifically a PDZ domain that interacts with the protein C kinase 1 (PICK1). The affinity of the identified molecule is equivalent to the natural ligand. Finally, molecular docking methods and mutational analyzes allowed the identification of the binding conformation.

Synthesis and experimental testing can be expensive, and high-throughput approaches may not find a specific ligand ([Chen et al., 2007]). A second approach is using artificial peptides. [Amacher et al., 2014] identified peptides with an improved affinity for the PDZ domain of the protein CFTR-Associated Ligand. However, it seems difficult to produce inhibitory peptides without systematic methods. Also, artificial peptides are difficult to maintain *in vivo* since proteases may degrade them. One approach is to use unnatural amino acids, such as β amino acids.

1.1.2 Computational studies of PDZ domains

Experiments are difficult, resource consuming and, may not provide enough details. For PDZ domains, several computational approaches have complemented the experiments, such as molecular dynamics.

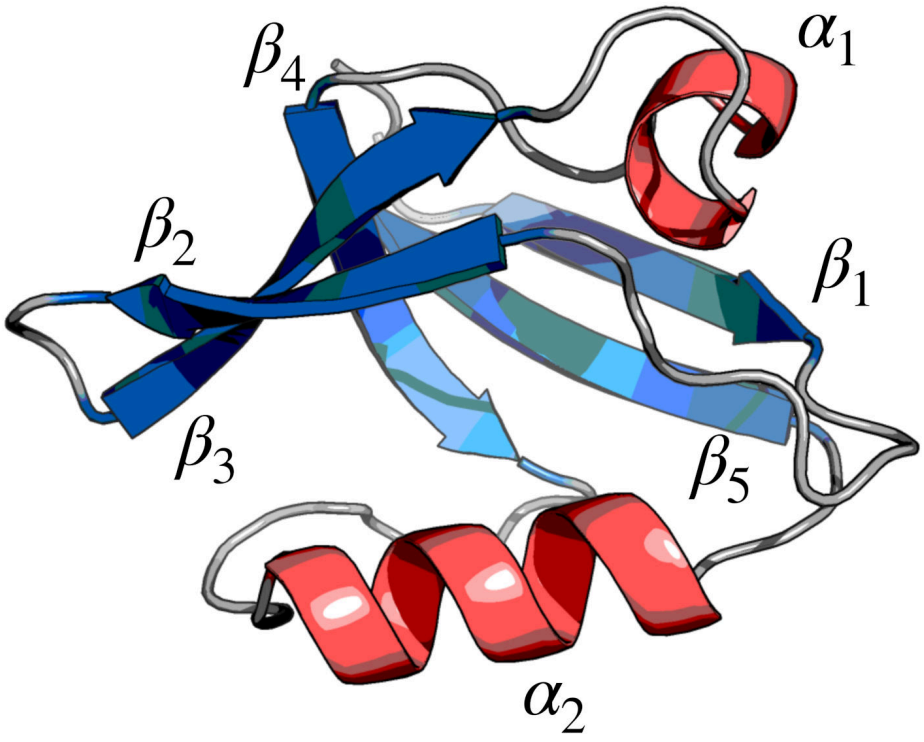


Figure 1.1: **Three-dimensional structure of the Cask PDZ domain.**

A study by [Blöchliger et al., 2015] unveiled the peptide recognition mechanism of PDZ domain partners. Ten independent simulations for a total of 57 μs established two phases in the binding process. The peptide was first recognized by non-specific long-range electrostatic interactions with side chains around the binding site. Positively charged residues around the binding site guided the peptide. In a second phase, the complex is held by hydrophobic interactions and the C-terminal side chain is buried in the binding site. This contrasts with the standard mechanism of protein folding where specific interactions are the key.

A second study by molecular dynamics identified quantitatively different contributions to affinity and specificity between PDZ and peptide motifs ([Basdevant et al., 2006]). To identify these contributions, 12 different PDZ domains were studied with MD trajectories from 20 to 25 ns. These simulations showed that the electrostatic contribution is minor compared to the nonpolar contribution and suggested that such contributions do not explain specific recognition. They concluded that a peptide can bind a given PDZ if it can provide a certain level of non-polar interactions, while the entropy contribution may explain the specificity. During the binding,

one observes a loss of degrees of freedom and thus a loss of entropy. [Basdevant et al., 2006] proposed that the variability observed for the entropic contribution could explain the specificity of the interactions PDZ/peptide.

A study conducted by [Smith and Kortemme, 2010] showed the effectiveness of a prediction method based on the PDZ peptide complex. This method is based on a Monte Carlo algorithm from the Rosetta suite, allowing the sampling of peptide variants. The sampling of variants allows them to construct an affinity profile. Then, Smith & Kortemme used phage display data from 17 human PDZ domains to describe their peptide preferences in the form of a profile. This approach allowed them to recover 70 to 80% of the most common amino acids experimentally found.

Another approach recently developed allowed the design of peptides based on their affinity to PDZ domains ([Bhattacherjee and Wallin, 2013, Villa et al., 2018]). It used an importance sampling method called Adaptive Landscape Flattening (ALF). Explicit modeling of bound and unbound states are used. With statistical physics concepts, Villa *et al* introduced a method capable of sampling peptide variants directly and rigorously on their affinity. This approach allowed the evaluation *in silico* of around 75,000 peptides for the affinity to the Tiam1 PDZ domain. Affinity free energies estimated in this study were in fair agreement with available experimental data.

While the usual strategy for pathway engineering is the use of inhibitors, it is possible to create new PDZ domains. The ability to create new PDZ domains and new signaling pathways is an interesting strategy. [Mignon et al., 2017] studied the complete redesign of the Tiam1 PDZ domain. They used a physics-based approach which is a key point for the understanding of fundamental protein design principles. Mignon *et al* sampled several thousand variants. Ten representative designs were selected and tested by molecular dynamics simulations up to 1.2 μs. This study paved the way for the use of a rigorous physical model to understand and design proteins.

1.1.3 The PDZ domain of the protein Cask

We studied here the PDZ domain of the protein Calcium calmodulin-dependent serine protein kinase (Cask). The Cask protein is involved in neuronal development and regulation of genes ([Hsueh et al., 2000]). A study of mice mortality demonstrated the role of Cask in neuronal

development ([Hsueh, 2009]). The Cask protein has several protein-protein interaction domains with one PDZ domain (Figure 1.2).

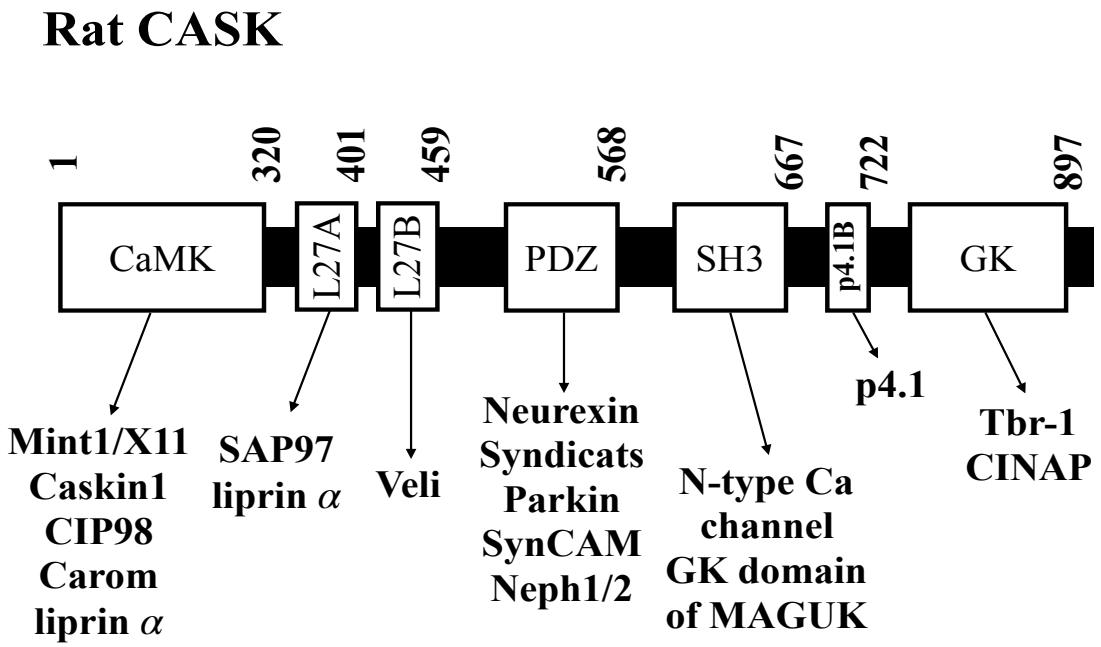


Figure 1.2: **Schematic representation of the Cask domains and their partners (from [Hsueh, 2009])** Each domain is represented by a rectangle. The numbers indicate the positions of amino acids in the protein sequence. The partners of each domain are listed.

The three-dimensional structure of the Cask PDZ domain with the peptide Neurexin-1 (6nid) highlights the binding interactions (Figure 1.3). The binding site is located between the β_3 sheet and the α_2 helix. The C-terminal position is completely buried, ARG 517 interacts with K_{-6} via a salt bridge. Other polar contacts illustrated in figure 1.3 involve Y_{-1} and the backbone atoms in the binding site.

1.2 Aminoacyl-tRNA synthetases

Protein synthesis is a biological process involving several families of macromolecules. In this thesis, we will study the family of aminoacyl tRNA synthetases (aaRS). aaRSs link genetic information stored in genomes and the production machine called the ribosome. Each aaRS

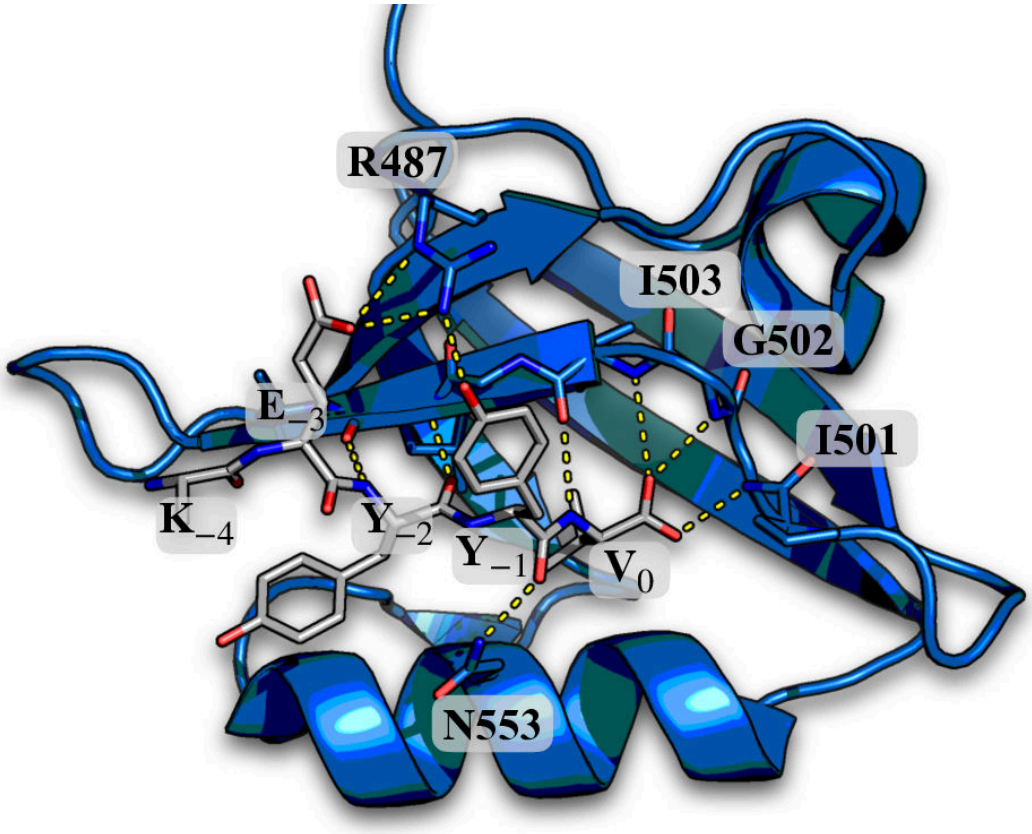


Figure 1.3: **Three-dimensional structure of Cask PDZ domain with Neurexin-1 peptide (pdb code 6nid).** In blue, the PDZ domain. Gray, peptide residues numbered is reversed order (0 for the last position, then -1 to position -5). Yellow, polar interactions between the peptide and binding site.

catalyzes the two-step aminoacylation reaction:

$$\begin{aligned} aa + ATP &\rightleftharpoons aaAMP + PP_i \\ aaAMP + PP_i + ARNt &\rightleftharpoons aa\text{-}ARNt + AMP \end{aligned} \tag{1.1}$$

This first reaction creates an aminoacyl adenylate ($aaAMP$) and releases a pyrophosphate (PP_i). The second reaction connects the tRNA to the amino acid ().

Each aaRS specifically binds an amino acid and tRNA. This specific relationship between aaRS, amino acids, and tRNA is essential for the accuracy of protein translation. Specificity results from the side chain composition of the active site and the backbone geometry (figure 1.4). However, errors may occur. To manage these errors, some aaRSs have an editing site.

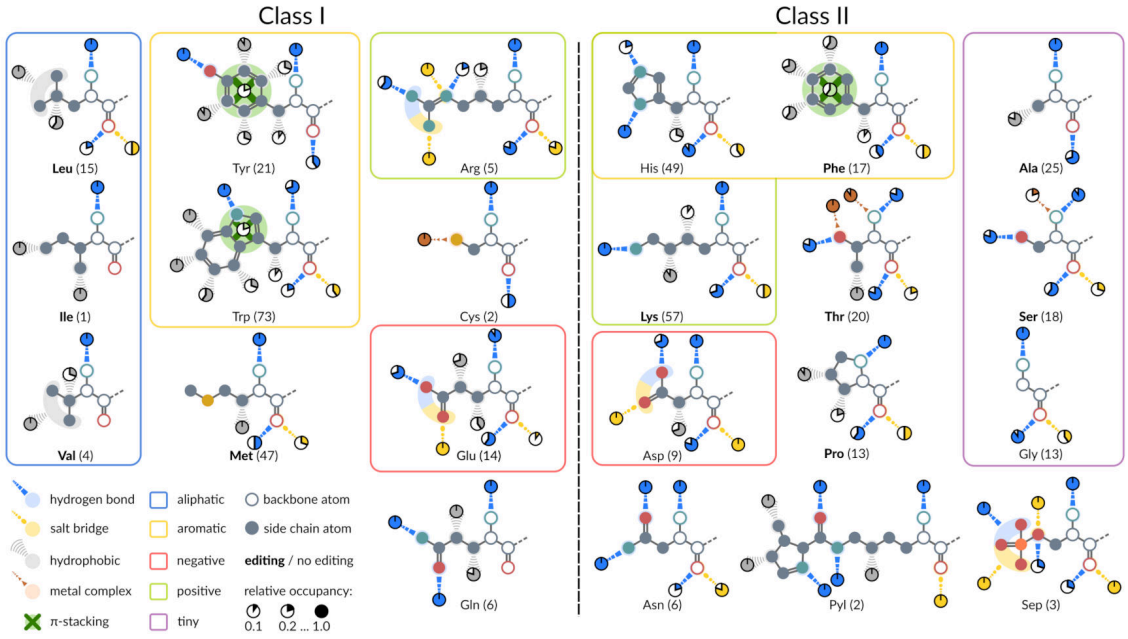


Figure 1.4: **Both aaRS classes for the 20 canonical amino acids, pyrolysine and, phosphoserine (from [Kaiser et al., 2020])** Each aaRS is associated with its cognate amino acid. Ligands are grouped by physicochemical properties. Interaction properties were determined with PLIP tool ([Salentin et al., 2015]) whose types are represented by an annotated color code. The quality of the interaction is represented by pie charts.

With increasing antibiotic resistance, bacterial inhibitors against aaRSs have been developed ([Vondenhoff and Aerschot, 2012]). Although aaRS structures are highly conserved in different areas of life, some differences can be used by specific molecules. A common strategy is to create molecules similar to aaRS substrates.

In addition, some parasites involved in diseases like malaria (*Plasmodium falciparum*) lack effective vaccines ([Organisation, 2015]). In the case of malaria, the World Health Organization reported about 438,000 deaths for 200 million infections in 2015 ([Organisation, 2015]). Inhibitors of the protein translation system were used as anti-parasite compounds. For example, inhibitors that bind specifically Alanyl-tRNA synthetase and threonyl tRNA synthetase were discovered ([Khan et al., 2011]).

AaRSs are also a target of interest for applications in biotechnology. Indeed, engineering of such enzymes is used to extend the genetic code by the incorporation of unnatural amino acids. The incorporation of new amino acids in proteins provides new or improved properties such as improved resistance to proteases with β amino acids ([Daura et al., 2001]). The additional carbon in the main chain decreases protease recognition. Here, we will first recall the structural properties of aaRS. Then, a few applications of aaRS engineering will be shown. Then, we will show some examples of genetic code expansion.

1.2.1 Three-dimensional structures of aminoacyl tRNA synthetases

Structural studies have shown that class I aaRSs contain a Rossman fold domain while class II aaRSs contain an ensemble of β sheets instead ([Ibba and Söll, 2000]) (Figure 1.5). The Rossman fold contains two well-known motifs involved in the production of the aminoacyl-adenylate. The KMSKS motif is contained in a mobile loop called the activation loop (figure 1.5). When the activation loop changes its conformation, this motif stabilizes the ATP moiety in the active site ([First and Fersht, 1995]). The second motif is HIGH whose role is linked to the stabilization of the transition state for aaAMP formation ([Schmitt et al., 1995]). Although the association error rate between amino acid and tRNA is low, a system of correction is needed. Class I has a domain called *Connective polypeptide 1* (CP1) which in some cases is an editing domain able to correct errors ([Ling et al., 2009]). Except for Tyrosyl-tRNA synthetase and tryptophanyl-tRNA synthetase, class I aaRSs are monomeric and are grouped into three subclasses. The first, denoted Ia, handles hydrophobic amino acids. The second, denoted Ib, handles residues with a long side chain. The third, denoted Ic, handles aromatics.

Class II has a different structural organization (Figure 1.6). The first point is the absence of the Rossman fold. Indeed, the catalytic site consists of seven β sheets associated with α helices, as described in the first class II structures (Seryl-tRNA synthetase and aspartyl-

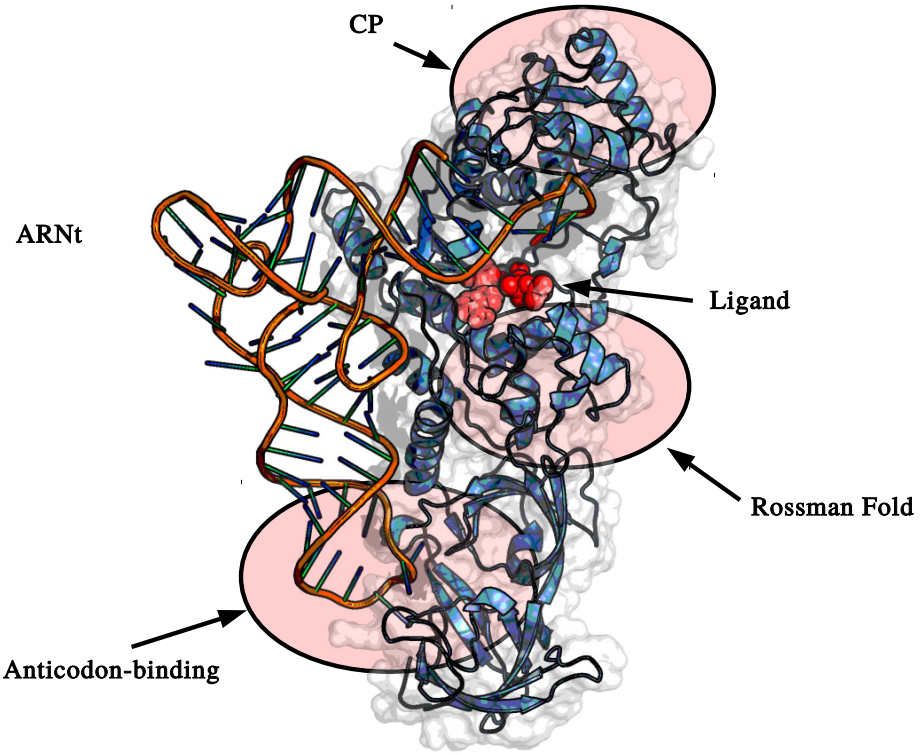


Figure 1.5: **Three-dimensional structures of class I, the example of glutaminyl-tRNA synthetase with glutamine and tRNA.** The ligand is represented by red spheres. The domain of interest are highlighted in red. The non-active tRNA is shown in orange.

tRNA synthetase) ([Kowal et al., 2001, Cusack et al., 1990]). ATP binds to class II in a bent conformation, and to class I in a more extended conformation (figure 1.7). Class II aaRSs can be grouped into three subgroups. The first, denoted IIa, groups aaRSs whose C-terminal domains are homologous. The C-terminal region is involved in recognizing tRNA. The second subclass, denoted IIb, has an N-terminal domain organized in β barrels, involved in binding tRNA. The third, denoted IIc, includes aaRSs whose aaRS oligomeric structure is not preserved.

1.2.2 aaRS engineering for genetic code extension

One application of interest is the extension of the genetic code to unnatural amino acids. Thus, nearly 70 new unnatural amino acids (UAA) have been added to the genetic code of *Escherichia coli* ([Liu and Schultz, 2010]). To expand the genetic code, it is necessary to design aaRS/tRNA couples that does not disturb the existing translation system. Usually, the amber codon is used

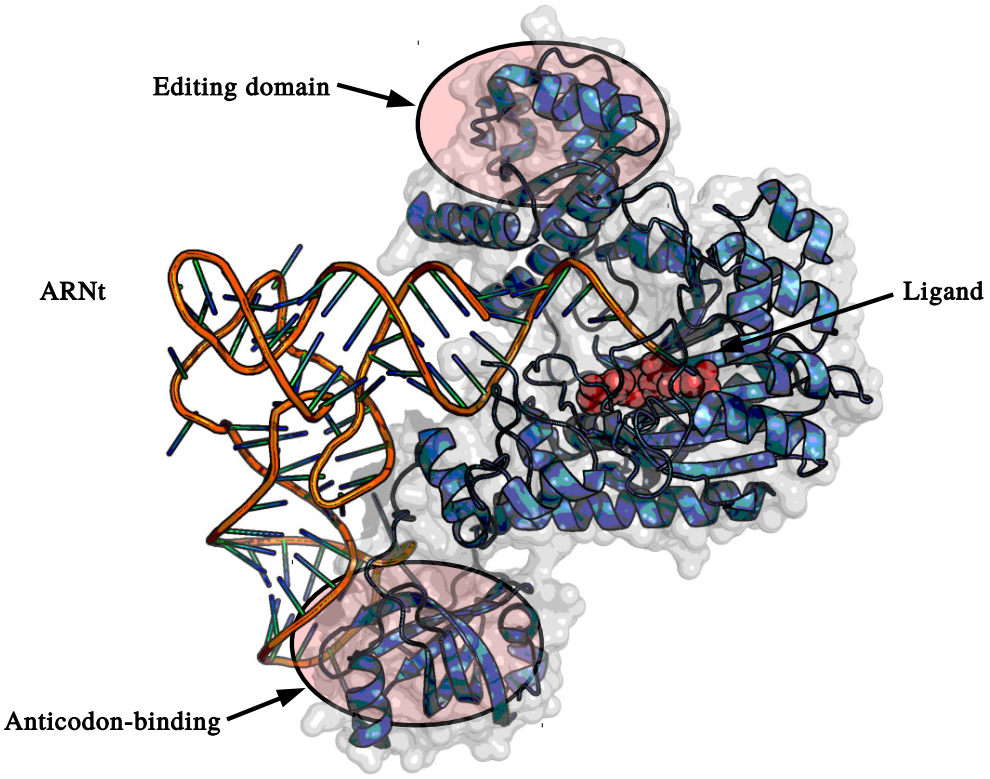


Figure 1.6: **Three-dimensional structure of Class II, the example of aspartyl-tRNA synthetases in complex with aspartyl-adenylate and tRNA.** The ligand is represented by red spheres. The domains of interest are highlighted in red. The non-active tRNA is shown in orange.

to encode the UAA.

Rajbhandary et al. built two specific aaRS/tRNA couples. The first is the glutaminyl-tRNA synthetase of *E. coli* paired with the human suppressor tRNA. The second is tyrosyl-tRNA synthetase (TyrRS) paired to a tRNA with the amber anticodon. However, the modified aaRS still process the natural tRNA, but in small quantities. Therefore, these two couples are not perfectly orthogonal to the expression system. Moreover, the expression plasmid carrying the gene for the modified aaRS cannot be maintained in *E. coli* ([Kowal et al., 2001]). This application illustrates the difficulty of introducing changes in the translation system.

Schultz et al. built a aaRS/tRNA couple which satisfies the principle of orthogonality. It uses *Methanococcus jannashii* TyrRS, which differs from *E. coli* TyrRS. *Methanococcus jannaschii* TyrRS has no editing domain and a minimalist anticodon binding domain. Thus, the aaRS will not eliminate the unnatural amino acid. Then, a amber anticodon is incorporated in

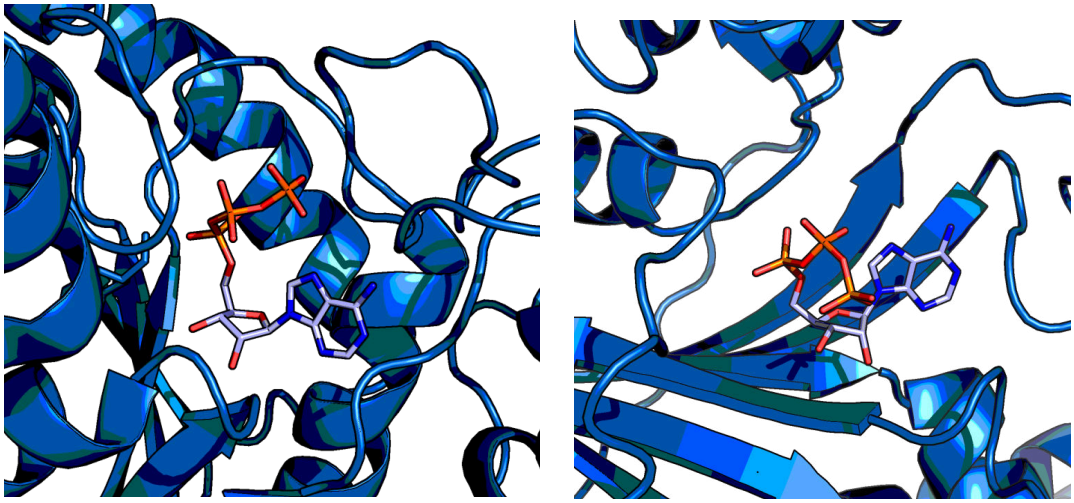


Figure 1.7: **Representation of ATP conformations bound to an aaRS in extended conformation (left, bound to a Class I, tyrosyl-tRNA synthetase, 1h3e) and compact (right, bound to a class II, histidyl-tRNA synthetase, 1kmn).**

the tRNA. The couple was optimized by a directed evolution method. First, positions to mutate were determined using crystallographic structures. Positive selection was applied to variants that activated the unnatural amino acid through its ability to suppress an amber mutation in a specific gene. Then, a negative selection was applied through a cytotoxic Barnase processing ([Xie and Schultz, 2006]).

1.2.3 Computational approaches to genetic code extension

Computational methods are important for the design of binding sites for bio-catalysts in industry. To illustrate these approaches, we present two applications to TyrRS. Baumann *et al* proposed a variant of TyrRS which binds the ortho-nitrobenzyl tyrosine (ONBY), a tyrosine homolog ([Baumann et al., 2019]). The active site of TyrRS contains around 30 residues. For the design of this variant, ten mutations were introduced in the active site. This set of mutations allows the selective activation of ONBY. To design this variant, the first numerical step was to create a library of variants predicted to bind ONBY. This step used the design procedure implemented in Rosetta3 ([Richter et al., 2011]). 143 variants were identified by the procedure. To refine these variants, 26 positions were allowed to mutate to obtain 3575 new variants. 49

variants with a satisfying Rosetta score were finally selected to form a profile. This set allowed the detection of a selective variant for ONBY.

Simonson *et al* proposed to change the stereospecificity of TyrRS from L-tyrosine (L-Tyr) to D-Tyrosine (D-Tyr) ([Simonson et al., 2016]). The subtle change of symmetry in ammonium is challenging. For this application, the Proteus software was used ([Simonson et al., 2013, Simonson, 2019]). The specificity of this approach is the use of a rigorous physical description of molecular interactions. Indeed, for such systems, a realistic description is necessary to discriminate the active variants. The design of variants was based on two types of approach: a high-throughput approach based on a heuristic algorithm for quick sampling ([Busch et al., 2008]). Then, alchemical free energy simulations were used to estimate free energy changes ([Simonson, 2001]). Mutations were introduced in a set of four positions (D81, Y175, N179, and N201). A variant whose preference has been improved for D-Tyr was discovered and validated experimentally. The agreement between experimental measurements and calculated ones shows that the use of physical models is a promising route for this type of application.

1.3 The Proteus framework

Here we present the main aspects of *CPD* using Proteus. First, we will describe the folded and unfolded models. Then, we will detail the energy function and the approach used for the exploration of sequences. Finally, we will illustrate the other techniques with two examples, involving Rosetta and Toulbar2.

1.3.1 Folded and unfolded models

Numerical applications related to the three-dimensional protein structure usually suffer from the curse of dimensionality. Indeed, a protein is a flexible object whose space of accessible configurations is large. This difficulty is added to the combinatorial mutation space. Thus, a simplification used here consists in considering a fixed backbone. Only the side chains remain flexible. A second simplification is the discrete set of conformations for the side chain called rotamers. Here, we use the Tuffery library ([Tuffery et al., 1997]).

Since the unfolded state is even more complex, one standard approach is to model it as an

extended peptide. In such a model, residues interact with the solvent but not with the other positions of the polypeptide. For a given sequence, the energy is the sum of the contributions of each amino acid independently.

1.3.2 The energy function

Proteus uses an energy function that can be divided into two contributions: protein-protein interactions and protein-solvent interactions. For the first contribution, we use the Amber force field ff99SB ([Tian et al., 2019]). The second contribution uses continuum electrostatics.

The interaction with the solvent in the folded state is divided into a polar contribution for the electrostatic effects and a non-polar contribution for the dispersion and hydrophobic effects. The polar contribution is computed with the *Generalized Born* (GB) model where the protein is modeled as a low dielectric medium embedded in a high dielectric medium:

$$E_{GB} = \frac{1}{2}(\frac{1}{\epsilon_W} - \frac{1}{\epsilon_P})\sum_{i,j} q_i q_j (r_{ij}^2 + b_i b_j exp[-r_{ij}^2/4b_i b_j])^{-1/2} \tag{1.2}$$

Here, ϵ_W and ϵ_P are the dielectric constants of the solvent and the protein, r_{ij} is the distance between atoms i and j, q_i and q_j represent the partial charges on the atoms i and j; b_i and b_j represent GB radii. GB models generally differ in how solvation radii are calculated. Proteus mainly uses the HCT method ([Hawkins et al., 1995]).

The GB term is a multi-body term which is very penalizing for the calculation speed. However, Simonson and Archontis ([Archontis and Simonson, 2005]) introduced the concept of *residue GB* which makes the GB method pairwise additive. This approach allows the GB calculation (figure 1.8) to be done in a reasonable time. This variant is called Fluctuating Dielectric Boundary (FDB) ([Villa et al., 2017]). A more approximate methode computes the GB radii of each side chain assuming it is in its native environment (NEA, [Simonson et al., 2013]).

For the non-polar contribution, Proteus offers two models. The first is a surface area model (SA):

$$E_{nonpolar}^{SA} = \sum_i \sigma_i A_i \tag{1.3}$$

The σ_i parameter reflects the preference of atom i to be buried or exposed. A_i is the solvent accessible surface area of the atom. Moreover, this model is also multi-body since more than

two atoms may overlap. Here we use a pairwise approximation ([Street and Mayo, 1998]).

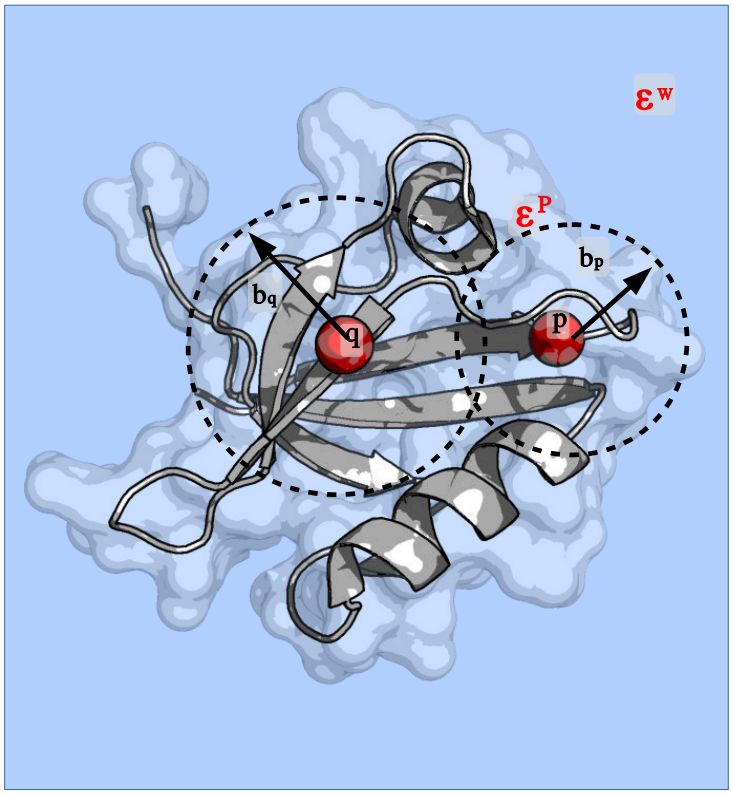


Figure 1.8: **Schematic representation of solvation radii for two charges p and q.** The two charges q and p are shown as red spheres. The solvation radii induced by the environment are represented by an arrow and a non-continuous circle. The protein is modeled by a dielectric medium of constant ϵ^p solvated in a dielectric medium of constant ϵ^W.

The second model proposed is the Lazaridis-Karplus model (LK) ([Lazaridis and Karplus, 1999, Michael et al., 2017]):

$$E^{LK}_{nonpolar} = \sum_i W_i = \sum_i (W_i^{ref} - \sum_{j \neq i} \int_{Vj} g_i(r_{ij}) dV) \approx \sum_i (W_i^{ref} - \sum_{j \neq i} g_i(r_{ij}) V_j) \qquad (1.4)$$

For each atom i, the contribution W_i to the solvation energy corresponds to the transfer of this atom from a fully solvated state to a partially buried one. W_i^{ref} is a parameter that corresponds to the free energy in the fully solvated state. Every other atom j contributes to a screening effect. The screening is here proportional to the volume of each j atom (V_j) and a Gaussian function (g_i) which depends on the distance between atoms i and j (r_{ij}).

1.3.3 Sampling of sequence space

Proteus uses an exploration based on the Monte Carlo approach. This method coupled with a physical energy function allows the sampling of sequences from the Boltzmann distribution. It is thus possible to deduce certain quantities of interest such as folding or binding free energies. In the context of a fixed backbone, implicit solvent and flexible side chains, consider a polypeptide sequence $S = S_1, S_2, ...S_p$ of p positions. Each position is associated with a rotamer denoted r_i. We denote $C(S) = r_1, r_2, ..., r_p$ a conformation of S. Sampling is defined by two types of motion: a change of rotamers ($r_i \rightarrow r'_i$) or type ($S_i \rightarrow S'_i$).

The polypeptide evolves with the energy function $E_M(S) = E_f(S) - E_U(S)$ where $E_M(S)$ is the energy of folding, $E_f(S)$ is the energy in the folded state, and E_U is the energy in the unfolded state. For a mutation $S_i \rightarrow S'_i$, we have the following energy change:

$$\Delta E_M = E_M(S) - E_M(S') = E_f(S') - E_f(S) - E_u(S') - E_u(S) \tag{1.5}$$

S' is the polypeptide with the S'_i mutation. A mutation in the folded state is thus accompanied by the reverse mutation in the unfolded state (figure 1.9). We obtain the following ratio of probabilities for S and S':

$$\frac{P(S)}{P(S')} = \frac{\sum_c exp(-\beta E_M(S,c))}{\sum_c exp(-\beta E_M(S',c))} = \frac{exp(-\beta \Delta G_{folding}(S))}{exp(-\beta \Delta G_{folding}(S'))} \tag{1.6}$$

We recognize the Boltzmann ratio; $\Delta G_{folding}(S)$ is the folding energy (respectively $\Delta G_{folding}(S')$). Therefore, variants are populated according to their relative folding energy, or stability.

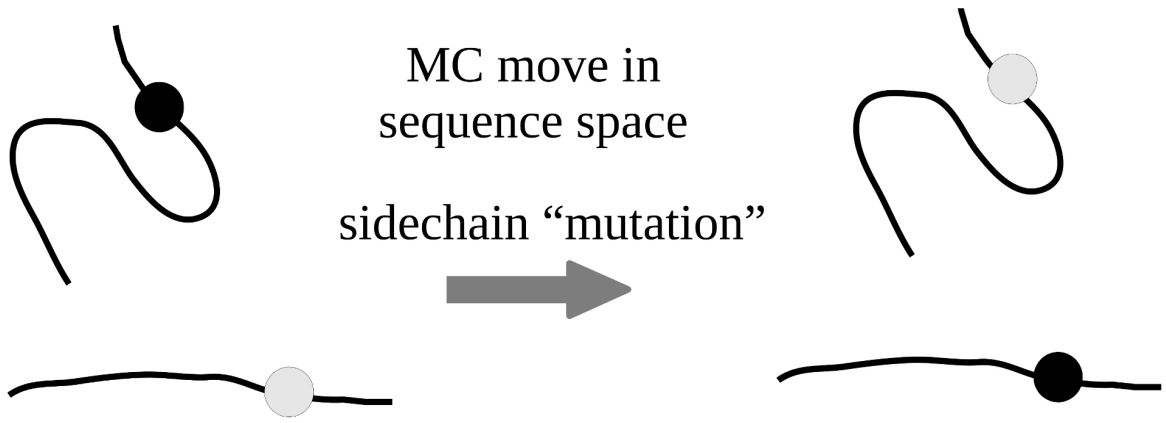


Figure 1.9: **Representation of a MC move.** A mutation in the folded state is accompanied by the reverse mutation in the unfolded state.

A second approach was proposed to design ligand binding. We allow the competition, not between the folded and unfolded states but between the bound and the unbound states. This required a new approach. The solution is to create an approximation of the free energy surface of sequences in the unbound state with an adaptive MC approach called Adaptive Landscape Flattening ([Villa et al., 2018]). This approximation, or bias, B is gradually incremented in an MC simulation of the unbound state. At the end of the simulation, the bias B approximates the free energy of sequences in the unbound state $E_u(S)$. In a second simulation, the bias is applied to the sampling of variants for the bound state. Now, sequences evolve with the energy function $E_a(S) = E_b(S) + B(S)$. Since B(S) approximates the free energy of sequences in the bound state, sequences are sampled according to Boltzmann distribution, but now controlled by the binding free energy. Details for this method are provided in chapter 4.

1.4 Other approaches

We can categorize existing methods according to two criteria: representation of the system ranging from strictly physical models to ones based on biological data. Another criterion is the exploration algorithm. One category seeks to produce one or more solutions (with or without guarantees) using stochastic and/or heuristic approaches such as Monte Carlo methods ([Mignon and Simonson, 2016]) or genetic algorithms. A second category seeks to find the optimal solution in the context of a given energy function.

1.4.1 A *knowledged-based* energy function paired with a stochastic search

We illustrate the stochastic approach and *knowledged-based* energy function with the design tool Rosetta fixbb ([Richter et al., 2011]). This approach can provide acceptable solutions in reasonable time. Also, the knowledged-based approach optimizes the energy function for a specific task. Like Proteus, Rosetta uses an all-atom representation. For the folded state, the backbone is also held fixed while side chains remain flexible. The flexibility of the rotamers is modeled with a library of rotamers depending on the local geometry of the backbone ([Shapovalov and Dunbrack, 2011]).

The energy function used to describe the interactions in the folded state is based on physical terms to which is assigned a weight. The energy function uses a modified Lennard Jones potential to model short-distance interactions. Electrostatic interactions are modeled by the Coulomb potential, modified to take account of a change in dielectric medium between buried regions and exposed regions. For the interaction with the solvent, Rosetta uses an implicit representation modeled by a Lazaridis-Karplus term ([Lazaridis and Karplus, 1999]). Although hydrogen bonds in standard force fields are described by the Coulomb potential, Rosetta uses a dedicated and optimized term to reproduce hydrogen networks of a set of 8,000 crystallographic structures. The bonded interactions are modeled by a statistical potential optimized on observed rotamers.

This energy function can compare a configuration to another for the same sequence but not different sequences. Thus, a type-dependent unfolded model is needed. Its parameters are empirically selected to reproduce sequences observed in crystal structures of high resolution. Thus, the fitness function used by Rosetta to search sequences is based on the folding energy of stability.

To search sequences, Rosetta uses a simulated annealing algorithm in which the temperature is reduced using a specific path during a Monte Carlo simulation. It is thus possible to obtain solutions that may be close to the global minimum. However, this approach cannot provide any guarantee.

The statistical optimization of the energy function improves the performance in real-world applications. In addition, the cost of the energy evaluation is similar to the energy calculation using ff99SB. Since the function is imperfect, finding the exact global minimum energy is not necessary. However, heavy use of empirical optimization makes this type of function non-transferable. In addition, the interpretation may be difficult. Finally, dependence on biological data limits its use for protein-ligand applications, since there are only little data.

1.4.2 Combinatorial exploration for an exact solution

To illustrate the exact approach, we use the example of ToulBar2 ([Traoré et al., 2013]). This type of approach focuses on the search for the optimal solution or Global Minimum Energy Conformation (GMEC). ToulBar2 uses an energy function based on the ff94SB force field ([Traoré et al., 2013]) calculated with OSPREY ([Hallen et al., 2018]). One of the exact search

algorithms proposed by Toulbar2 is based on a cost function network (CFN). A first element to note here is the need for a discrete representation of the protein. Indeed, this approach models the search as a linear programming problem. Also, the energy function must be decomposable in pairs to allow an exact search.

The CFN approach models the research problem by a triplet P = (X, D, C) where X is a set of variables representing the positions or the pairs of positions for a given polypeptide, whose possible values correspond to the D rotamers. A cost function, unary or binary, representing the energy contribution is associated with each variable. A depth-first branch and bound algorithm is used to identify the GMEC. The second benefit of this approach is the possibility of determining a set of sub-optimal solutions in a given energy interval above the GMEC. The enumeration of sub-optimal solutions in a range of values energy may ensure an effective sampling of the most favorable variants according to the model used. Nevertheless, the required decomposition into pairs of interaction prevents the use of more realistic physical models such as GB FDB for the solvent.

Chapter 2

A physics-based energy function allows the computational redesign of a PDZ domain

The following chapter uses the text from the article: ***A physics-based energy function allows the computational redesign of a PDZ domain***, Vaitea Opuu, Young Joo Sun, Titus Hou, Nicolas Panel, Ernesto J. Fuentes, Thomas Simonson (2020), *Scientific Reports* 10, 11150.

SCIENTIFIC REPORTS
nature research

OPEN

A physics-based energy function allows the computational redesign of a PDZ domain

Vaitea Opuu[1,3], Young Joo Sun[2,3], Titus Hou[2], Nicolas Panel[1], Ernesto J. Fuentes[2]✉ & Thomas Simonson[1]✉

Computational protein design (CPD) can address the inverse folding problem, exploring a large space of sequences and selecting ones predicted to fold. CPD was used previously to redesign several proteins, employing a knowledge-based energy function for both the folded and unfolded states. We show that a PDZ domain can be entirely redesigned using a "physics-based" energy for the folded state and a knowledge-based energy for the unfolded state. Thousands of sequences were generated by Monte Carlo simulation. Three were chosen for experimental testing, based on their low energies and several empirical criteria. All three could be overexpressed and had native-like circular dichroism spectra and 1D-NMR spectra typical of folded structures. Two had upshifted thermal denaturation curves when a peptide ligand was present, indicating binding and suggesting folding to a correct, PDZ structure. Evidently, the physical principles that govern folded proteins, with a dash of empirical post-filtering, can allow successful whole-protein redesign.

Protein sequences have been selected by evolution to fold into specific structures, stabilized by a subtle balance of interactions involving protein and solvent[1,2]. In contrast, random polymers of amino acids are very unlikely to adopt a specific, folded structure[3,4], and exhibit instead a more disordered structure[5]. A powerful approach to understand the evolution of proteins and the basis of folding is to perform computer simulations that mimic evolution. This can be done with computational protein design (CPD), which explores a large set of sequences and selects ones predicted to adopt a given fold[6–8]. A typical simulation imposes a specific geometry for the protein backbone, corresponding to the experimental conformation of a natural protein. Side chains are mutated randomly. Variants with a favorable predicted folding free energy are saved. The folded state energy function can be physics-based or knowledge-based[9–11] while the unfolded state energy is knowledge-based. The protein is considered "redesigned" if most of the protein side chains are allowed to mutate during the simulation.

The successful redesign of complete proteins was reported in 2003[7,12] and small miniproteins were redesigned even earlier[6,13]. Several other successes were obtained[14–17], including a study where 15000 miniproteins (40–43 amino acids) were redesigned[18]. 6% of the designs were shown to be successful; i.e., the designed miniproteins folded into the correct tertiary structure. The others either could not be overexpressed and purified, or did not fold as predicted. All of the applications to proteins described the folded structure with an energy function that was at least partly knowledge-based, or statistical. Statistical energy terms included terms derived from experimental amino acid propensities and evolutionary covariances[17], terms derived from inter-residue distance distributions in crystal structures[16], and terms derived from torsion angle and hydrogen-bond distance distributions in crystal structures[11,14,15]. All of the applications described the unfolded structure with a fully statistical, knowledge-based model.

Energy functions for the folded state can also be non-empirical, or physics-based, and taken from molecular mechanics[19]. There are then only two energy terms for nonbonded interactions between protein atoms, which correspond to the elementary Coulomb and Lennard-Jones effects. Their parameterization relies mainly on fitting quantum chemical calculations performed on small model compounds in the gas phase. The solvent is described implicitly, using varying levels of approximation[20]. The most rigorous model used so far is a dielectric continuum model[21]. This requires solving a differential equation, which is technically impractical in a protein

[1]Laboratoire de Biologie Structurale de la Cellule (CNRS UMR7654), Ecole Polytechnique, Institut Polytechnique de Paris, Palaiseau, France. [2]Department of Biochemistry, Carver College of Medicine, University of Iowa, Iowa City, USA. [3]These authors contributed equally: Vaitea Opuu and Young Joo Sun. ✉email: ernesto-fuentes@uiowa.edu; thomas.simonson@polytechnique.fr

design framework. Therefore, a Generalized Born (GB) approximation is more common. GB contains much of the same physics but provides a simpler, analytical energy expression[20]. GB models have been studied extensively in the context of protein design but also molecular dynamics, free energy simulations, acid/base calculations, ligand binding and protein folding[22–25]. They reproduce the behavior of the dielectric continuum model rather accurately. Therefore, an energy function that combines molecular mechanics for the protein with a GB solvent can be considered "physics-based", even though it is not entirely constructed from first principles. A molecular mechanics energy, combined with a very simple solvent model, was used to design two artificial proteins that each consisted of a four-helix bundle, where an elementary unit of 34 amino acids was replicated four times[26,27]. However, until now, there has not been a complete, experimentally-verified redesign of a natural protein using a physics-based energy function for the folded protein.

Here, we report the successful use of a physics-based energy function to completely redesign a PDZ domain of 83 amino acids. PDZ domains ("Postsynaptic density-95/Discs large/Zonula occludens-1") are globular domains that establish protein-protein interaction networks[28]. They interact specifically with target proteins, usually by recognizing a few amino acids at the target C-terminus. They have been extensively studied and used to elucidate principles of protein evolution and folding[29,30]. Our design started from the PDZ domain of the Calcium/calmodulin-dependent serine kinase (CASK) protein. It used the backbone conformation from a new, high-resolution X-ray structure of apo CASK reported here. Several other CASK X-ray structures are also known, with bound peptides. The CASK melting temperature is about 10 °C higher than that of the Tiam1 PDZ domain, which we attempted to redesign earlier[33]. This increased thermostability could aid in retrieving folded CASK designs. Design was performed by running long Monte Carlo (MC) simulations where most positions were allowed to mutate and all positions could explore a library or conformers, or rotamers. Positions occupied by glycine (seven) or proline (two) were not allowed to mutate. 13 positions that directly contact a peptide ligand in CASK:peptide complex structures (such as PDB 6NID) also kept their wild-type identity. All 61 of the other side chains (73.5% of the sequence) were allowed to mutate freely into any amino acid type except Gly or Pro, for a total of 3.7×10^{76} possible sequences. To describe the folded state, we used a physics-based energy function that combined the Amber molecular mechanics force field[31] and a GB solvent[32]. To describe the unfolded state, we used a knowledge-based energy function[33]. The Proteus software was used[34]. Three sampled sequences, or designs were chosen for experimental testing, based on their low energies and several empirical criteria. All three were shown to fold, with good evidence the folded structure was the target, native PDZ fold. In particular, secondary structure content was native-like and binding to one or two peptides that are known CASK ligands was demonstrated for two of the three designs. Therefore, the redesign is considered a success. Evidently, the physical principles that govern folded proteins, as captured by molecular mechanics and continuum electrostatics are sufficient to allow whole-protein design, at least when assisted by a moderate empirical post-filtering. This is encouraging, since these methods give physical insights, can be systematically improved, and are transferable to nucleic acids, sugars, noncanonical amino acids, and ligands of biotechnological interest.

Results

MC simulations were done using the CASK backbone conformation (Fig. 1). The method is detailed in Supplementary Material. 61 of 83 residues were allowed to mutate into all types except Gly and Pro. 13 residues known to be directly involved in peptide binding were not allowed to mutate (but could explore rotamers). The exploration did not use any bias towards natural sequences or any limit on the number of mutations. The 2,000 sequences with the lowest folding energies were kept for analysis. Below, we describe their computational characterization and the selection of three representative sequences for experimental characterization.

Computational characterization and sequence selection. The top 2,000 sequences spanned a folding energy interval of 1.5 kcal/mol. They were analyzed by the Superfamily fold recognition tool[35], which assigns sequences to SCOP[36] structural families. None of the top 2,000 Proteus sequences were assigned by Superfamily to any other fold in SCOP; all were recognized as belonging to the PDZ family. Blosum40 similarity scores between the designed sequences and natural sequences from the Pfam database were also computed (Fig. 2). The scores were high, and comparable to those of natural PDZ domains. The peaks in the Proteus histograms are narrow, indicating that the 2,000 lowest-energy sequences are similar to each other. Similarities to CASK are in Supp. Material (Fig. S1).

To narrow down the number of sequences for experimental testing, we excluded those with isoelectric points estimated to be close to the physiological pH, between 6.5 and 8.5, which might be subject to aggregation and difficult to express. This reduced the number of sequences from 2,000 to 1,268. Next, we used a criterion of negative design, by considering the confidence levels for the Superfamily assignments to the PDZ family, instead of another SCOP family. Of the 1,268 sequences left, we only retained those that had Superfamily match lengths above the mean value (over the 1,268) and E-values above the mean ($\log_{10} E < -31$). This left us with 692 sequences. We reduced the number further using four empirical criteria. (1) We excluded sequences with similarity scores versus Pfam below the mean (over the 692 remaining sequences). This eliminated a window of candidate sequences about 10 points wide, to the left of the mean, plus a few sequences in the lefthand tail of the distribution. We were left with 215 sequences. (2) We excluded sequences that had a cavity buried in the predicted 3D structure. (3) We required a total unsigned protein charge of less than 6. (4) We allowed no more than 15 mutations that drastically changed the amino acid type (defined by a Blosum62 similarity score between the two amino acid types of − 2 or less).

We were left with 16 candidate sequences, shown in Fig. 3. They were separated into four groups by visual inspection. Group 2 was eliminated based on its Arg494 residue, absent from CASK homologs. One candidate was selected from each of the other groups (highlighted in Fig. 3), with a preference for native or homologous

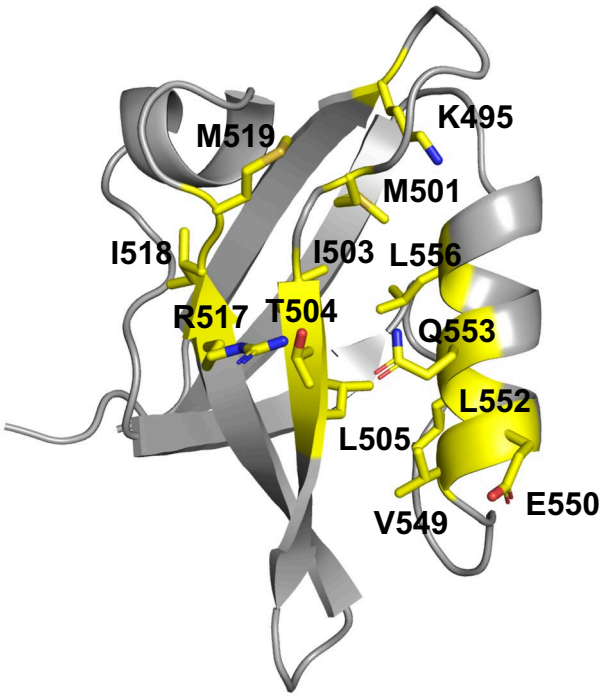


Figure 1. CASK 3D structure. The 13 amino acids in yellow are involved in ligand binding and were not allowed to mutate in the simulations.

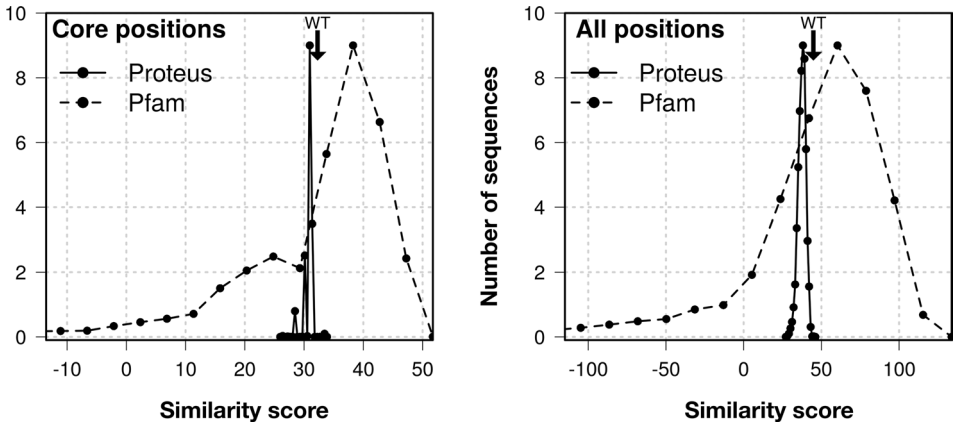


Figure 2. Blosum similarity scores compared to natural Pfam sequences. Black line: histogram of scores for the top 2,000 Proteus sequences, considering only 15 core positions (left) or all positions (right). Dashed line: scores for the Pfam sequences themselves. WT CASK score is indicated by an arrow.

residue types at positions 492 (candidate 1350), 494 (candidate 1555), and 548 (candidate 1669)—positions that are close to the peptide binding interface. The three candidates were simulated by molecular dynamics with explicit solvent for one microsecond each, and their stabilities and flexibilities appeared comparable to the wild-type (Supplementary Material, Figs. S2–S3). Therefore, the three sequences were retained for experimental testing. The number of mutations, compared to wild-type CASK, were 50 (candidate 1350), and 51 (candidates 1555 and 1669), representing just over 60% of the sequence.

Experimental characterization of selected sequences. *Earlier designs based on the Tiam1 template.* Computational redesign of Tiam1 was described earlier[33]. It used the Tiam1 PDZ domain structure (PDB code 4GVD; Supplementary Material, Fig. S4). The GB electrostatics model included an additional "Native Environment Approximation" (NEA)[37], where each atom experienced a constant dielectric environment that corresponded to the native sequence and conformation (see Computational Methods in Supplementary Mate-

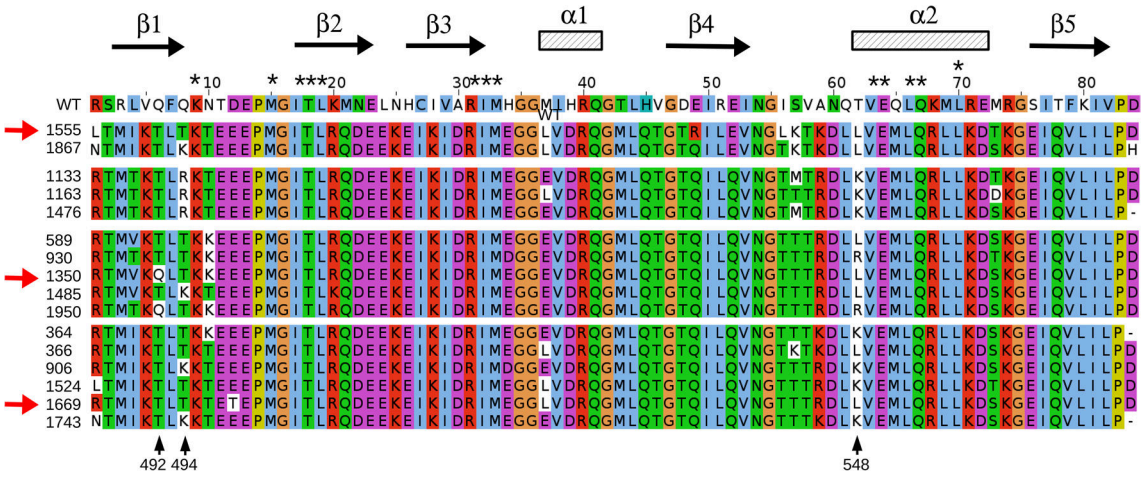


Figure 3. WT and the 16 final candidate designed sequences based on the CASK template (Clustal colors). The sequences tested experimentally are indicated by red arrows. Asterisks (above) indicate positions not allowed to mutate during the design, in addition to Gly, Pro.

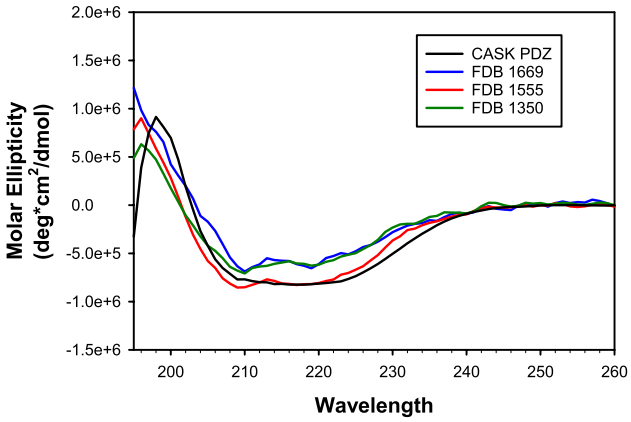


Figure 4. Circular dichroism spectra of a natural PDZ domain (CASK, black) and three selected designs based on the CASK template and the FDB electrostatic model. FDB-1350 (green), FDB-1555 (red), and FDB-1669 (blue) all have α helix and β strand signals similar to a native PDZ domain like CASK (black). The concentration of each protein ranged from 10 to 20 μM.

rial). This removed the many-body character of the GB model and made the calculations very efficient. Eight designs were expressed and purified. Their yields were low. CD gave spectra typical of random coil polymers, suggesting the proteins were misfolded (Supplementary Material, Fig. S5). 1D-NMR spectra of the amide region of the NEA designs had limited dispersion and broad resonances compared to the native Tiam1 PDZ domain, corroborating the CD data. An example is shown below; others are in Fig. S6. Differential scanning fluorimetry (DSF) in the presence of known Tiam1 ligands did not show any binding by the Tiam1 NEA designs, while the Tiam1 PDZ domain showed robust binding (Supplementary Material, Fig. S7). Together, these data indicate that the NEA-based designs of the Tiam1 PDZ domain could be overexpressed but adopted unfolded structures, unable to bind known Tiam1 peptide ligands.

Designs based on the CASK template. Next, we characterized the three designs selected above, which we refer to as FDB-1350, FDB-1555, and FDB-1669. They were obtained using as template a new apo CASK PDZ domain structure (PDB code 6NH9, reported here). The Tiam1 and CASK backbone conformations have a small rms deviation of 1.0 Å, despite a low sequence identity of 20.5%. CASK has a ~ 10 °C higher melting temperature, which could facilitate its redesign. The new calculations used a more rigorous GB electrostatics model (Supplementary Material), termed the "Fluctuating Dielectric Boundary" model (FDB)[38]. With this model, the dielectric environment of each atom was updated on-the-fly during the simulation, instead of being represented by a mean environment. The expression yields in *E. coli* were improved over the NEA Tiam1 designs, though not to the level typically seen with native PDZ domains. In contrast to the NEA Tiam1 designs, CD spectra were similar to native PDZ domains, suggesting these designs were structured (Fig. 4). 1D-proton NMR of the amide region showed good dispersion and sharp lines, consistent with a folded protein (Fig. 5B) and in contrast to the earlier,

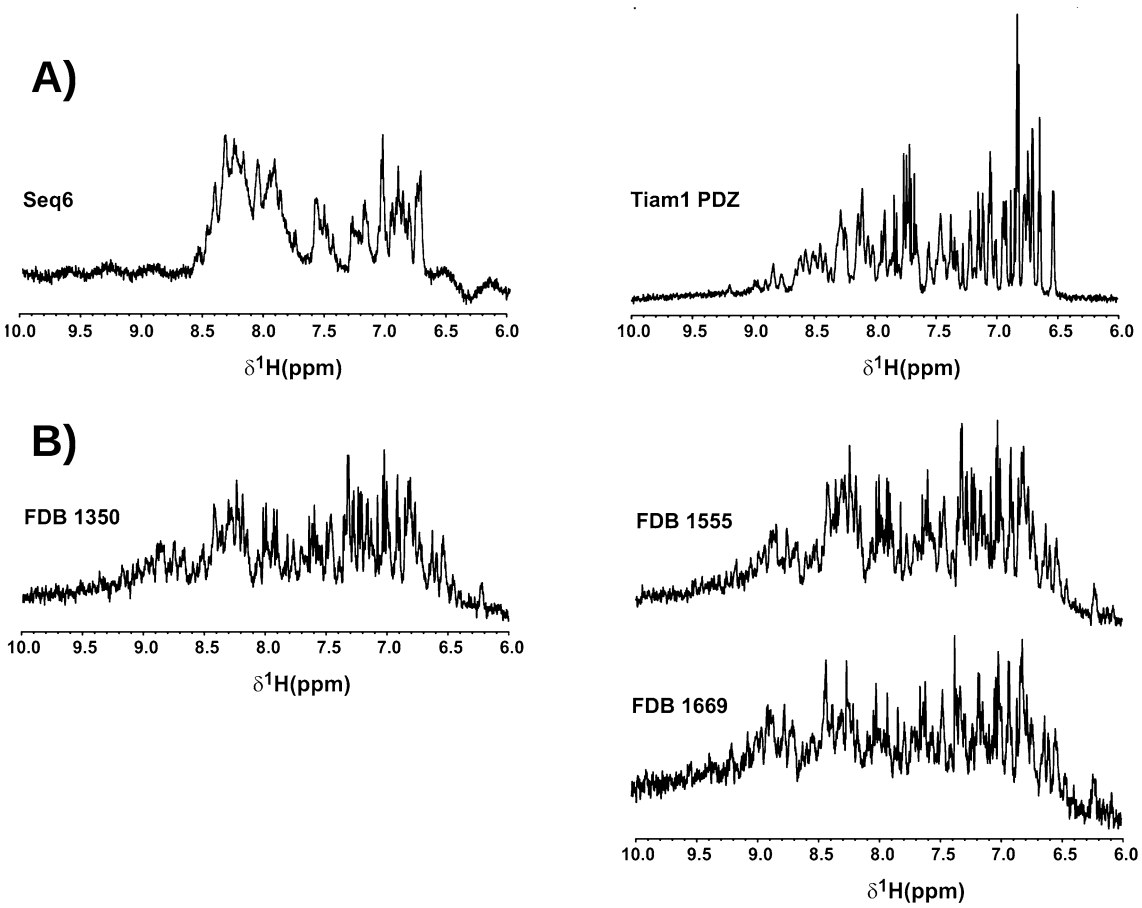


Figure 5. Proton NMR spectra of the natural Tiam1 PDZ domain and selected designs. (**A**) Left: a design obtained with the Tiam1 template and the NEA electrostatic model; right: Tiam1. (**B**) 3 designs obtained based on the CASK template and the FDB electrostatic model. The concentration of the designed proteins ranged from 14 to 22 μM; Tiam1 concentration was 150 μM.

Protein[a]	$T_{1/2}$ (°C) and $\delta T_{1/2} = T_{1/2}^{apo} - T_{1/2}$ (in parentheses)				Binding[b]
	Apo	SDC1	Caspr4	NRXN	
CASK PDZ	57.2 ± 0.2	58.4 ± 0.1	58.7 ± 0.1	58.1 ± 0.2	**SDC1, Caspr4**
		(+ 1.2)	(+ 1.5)	(+ 0.9)	NRXN
FDB-1350	49.8 ± 0.4	50.7 ± 0.2	50.4 ± 0.4	51.3 ± 0.2	SDC1
		(+ 0.9)	(+ 0.6)	(+ 1.5)	**NRXN**
FDB-1669	49.1 ± 0.1	49.6 ± 0.1	49.5 ± 0.0	50.1 ± 0.1	**NRXN**
		(+ 0.4)	(+ 0.4)	(+ 1.0)	
FDB-1555	49.9 ± 0.2	50.2 ± 0.1	50.3 ± 0.1	50.5 ± 0.6	–
		(+ 0.3)	(+ 0.5)	(+ 0.6)	

Table 1. DSF for wild-type CASK and three Proteus designs. [a]Protein concentration was ~ 25 μM (about 0.25 mg/ml). Peptide concentration was 300 μM. [b]When $\delta T_{1/2}$ was larger than sum of the standard deviation of apo and each peptide, we considered the peptides to have a significant change in $T_{1/2}$, indicating binding to the PDZ domain. ± indicates standard deviation of three biological replicates. Peptides in bold (right column) produced the largest changes.

Tiam1 redesigns (Figs. 5A and S6). The designed proteins' spectra had noisier baselines, due to a seven- to ten-fold lower concentration, compared to CASK.

We tested the ability of the designs to bind CASK ligands, using DSF experiments. The CASK PDZ domain showed binding to SDC1, Caspr4 and NRXN (Fig 6 and Table 1), as expected. Strikingly, two of the three CASK FDB designs characterized also showed binding to some of the peptides. Thus, FDB-1350 had a significant thermal shift in the presence of NRXN and SDC1. FDB-1669 showed a 1.0 °C change in $T_{1/2}$ in the presence of the

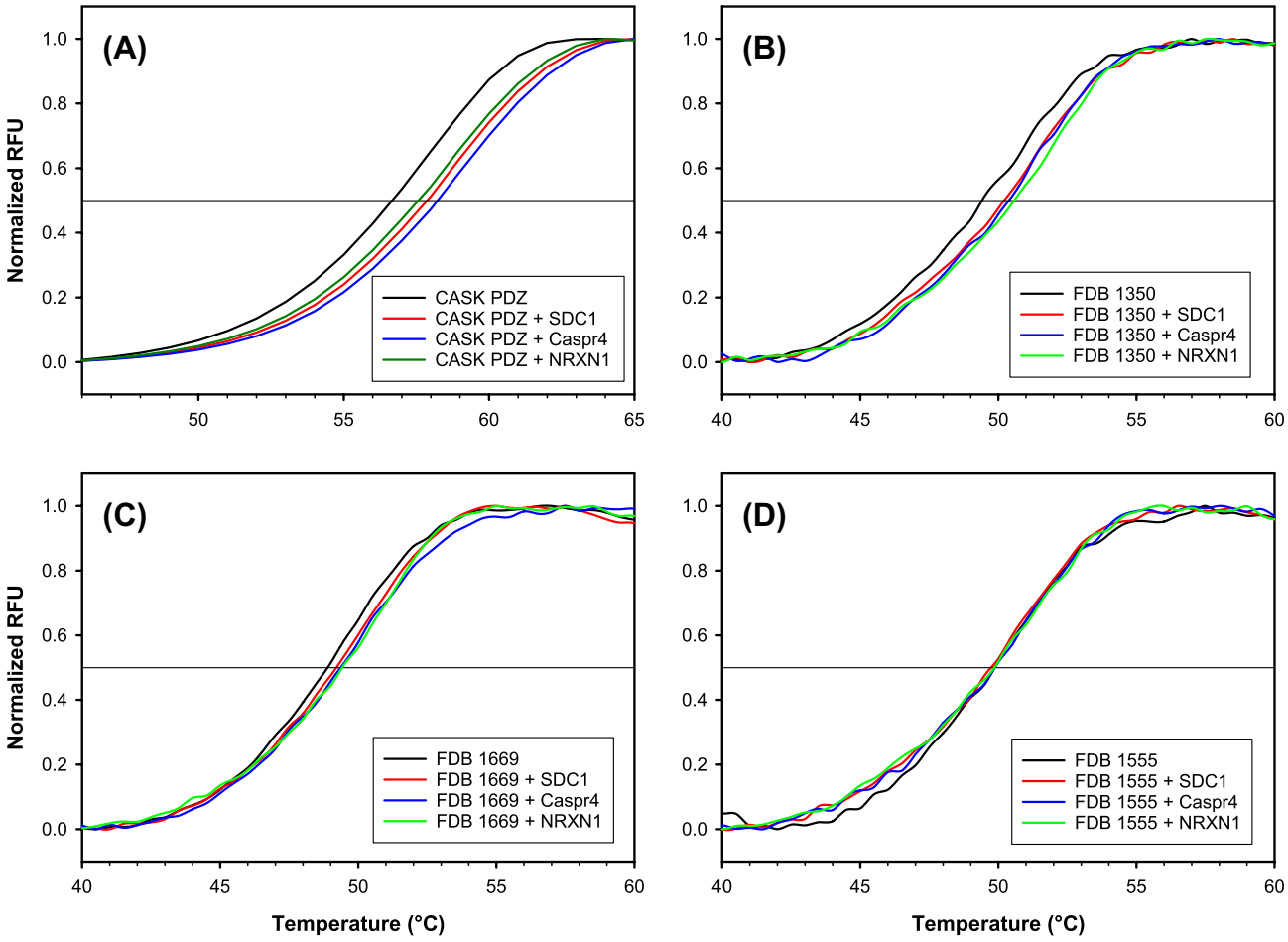


Figure 6. Differential scanning fluorimetry of (**A**) a natural PDZ domain (CASK) and (**B**–**D**) three selected designs based on the CASK template and the FDB electrostatic model. Signals in the absence and presence of the SDC1, Capr4 and NRXN peptides.

NRXN peptide. In contrast, FDB-1555 did not show significant thermal shifts in the presence of any peptide. From these data, we conclude that the three CASK FDB designs were folded and two were capable of interacting with peptide ligands. In principle, the CD and NMR spectra could be obtained with an alternative protein fold, distinct from the target PDZ fold. However, the structural data clearly indicate that the designs are well-ordered and have a secondary structure content similar to the CASK target. Importantly, the ordered character, the secondary structure content, the ability to bind CASK ligands, the structural stability during microsecond MD runs, and the Superfamily classification as a PDZ domain strongly suggest that the designed proteins adopt the target PDZ fold.

Discussion

Protein folding is thought to be induced by protein–solvent and solvent–solvent interactions[39], since folding buries nonpolar groups and allows waters to interact with polar amino acid side chains and other waters. In this picture, the protein dielectric properties play a role, with the low-dielectric interior pushing polar protein groups out towards high-dielectric solvent. The protein nonpolar surface also plays a role, with exposed surface leading to fewer water–water interactions[40]. Thus, it is common to discuss protein solvation in terms of nonpolar and electrostatic components, and most implicit solvent models rely on this separation[20]. Small proteins have been found to fold correctly in MD simulations with both explicit solvent and accurate implicit solvent models[22,41], which can all be considered physics-based. The inverse folding problem is even more complex, since it explores an enormous space of sequences, albeit with a reduced conformation set. Modeling the solvent is a key step to solve this problem, and a key ingredient of our procedure.

The first solvation component in our model is nonpolar and uses accessible surface areas and atomic surface tensions. Nonpolar solvation of a large collection of small molecules correlated well with surface area[42], supporting this treatment. The surface tension parametrization was updated recently, compared to our earlier Tiam1 designs[43]. Surface interactions in proteins are complex and have a many-body character[6,32], since three or more groups can have surfaces that all overlap. Our model explicitly includes backbone-side chain triple overlaps, while others are accounted for implicitly[43].

The largest solvation effects are electrostatic, and they also have a many-body character. Indeed, a side chain can desolvate an interacting pair, affecting the strength of their interaction. The electrostatic, Generalized Born component of our model captures this effect. However, for previous Tiam1 design calculations[33], we had used an approximation where each protein residue experienced a constant, native-like, dielectric environment. This removed the many-body character of electrostatic solvation. The Tiam1 designs were shown here to be largely unsuccessful: the proteins could be overexpressed but were only weakly structured. In contrast, preserving the many-body solvation was shown previously to give improved accuracy for side chain pK_a's[38] and increased similarity between CPD sequences and natural sequences of several PDZ proteins[38]. Therefore, for the CASK redesign, we applied the newer, many-body FDB model and obtained improved results. We did not test whether the improved, FDB model would have also produced valid designs with the Tiam1 backbone as the template.

Our calculations used a CASK X-ray structure reported here, determined at 1.85 Å resolution. In our design procedure, the protein backbone was held fixed in the X-ray conformation, while side chains mutated and explored rotamers. More precisely, the backbone motions were not ignored but were treated implicitly, through the protein dielectric constant, ϵ_P. The value used here, $\epsilon_P = 4$, is known to be physically reasonable for proteins. MD simulations further showed that the tested sequences have backbone structures very similar to the wild-type protein and native-like flexibilities.

While our folded state model was physics-based, the design procedure included two other elements that were knowledge-based. For the unfolded state, we assumed a simple, extended peptide model, to which an empirical correction was added that involved type-dependent amino acid chemical potentials[37]. All successful whole-protein redesigns have used similar, knowledge-based unfolded models. Second, we used several filters to choose a handful of sequences for experimental testing, and most of the filters were empirical. Indeed, the folded and unfolded models are imperfect, and while they produced at least three sequences that fold correctly (true positives), they presumably also produced false positives. The empirical filtering did not affect the sequence design, but was used to increase the chances that we would pick true positives for experimental testing. Starting from sequences within 1.5 kcal/mol of the top folding energy, we used the (computed) isoelectric point to reduce the chances of aggregation. We also used negative design, based on the Superfamily fold recognition tool. Indeed, negative design against aggregation or alternate folds was not included in the MC calculations. This left us with 692 designed sequences. Next, we eliminated sequences whose Blosum similarity to natural PDZ sequences was below the average of the 692 remaining sequences. This criterion was not very stringent, because the distribution of the Blosum scores was already very narrow (see Fig. 2, right panel, solid line and Fig. S1). At this point, we were left with 215 sequences. We then eliminated sequences whose structural models included large cavities and ones with a large net charge, which could lead to electrostatic repulsion within the folded structure. Finally, we eliminated sequences with more than 15 "drastic" mutations (corresponding to Blosum scores of − 2 or less). This left us with 16 sequences. We chose 3 that were representative.

The three tested proteins could be overexpressed, had sharp 1D-NMR peaks typical of a folded protein and native-like CD spectra. Two exhibited a shift of their thermal denaturation in the presence of one or two peptides that are known CASK ligands. Evidently, our physics-based folded model and empirical unfolded model can be used to successfully redesign a whole protein, at least with the help of some empirical post-filtering. The expression yields, protein solubilities and stabilities of the designed sequences were lower than for wild-type CASK, so that it was not possible to produce large amounts of pure protein for 2D-NMR or X-ray crystallography. It may be possible to improve this behavior by testing a larger number of designs, by using a more sophisticated filtering of candidate sequences for solubility (beyond estimating the isoelectric point), or by improving the physical model even further. Model improvements might include backbone-dependent rotamers and/or multiple backbone conformations.

The present design method, which combines molecular mechanics, continuum electrostatics, and Boltzmann sampling, is an example of physics-based CPD. It is striking and encouraging that this approach allows whole protein redesign to be done successfully. We expect that the physics-based route will increasingly yield valuable insights and should be a valuable complement to knowledge-based CPD and experimental design.

Received: 18 March 2020; Accepted: 8 June 2020
Published online: 07 July 2020

References

1. Kauffman, S. A. *The Origins of Order, Self-organization and Selection in Evolution* (Oxford University Press, New York, 1993).
2. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.***48**, 545–600 (1997).
3. Wilson, D. S., Keefe, A. D.& Szostak, J. W. The use of mRNA display to select high-affinity protein-binding peptides. *Proc. Natl. Acad. Sci. USA***98**, 3750–3755 (2001).
4. Jackel, C., Kast, P.& Hilvert, D. Protein design by directed evolution. *Annu. Rev. Biochem.***37**, 153–173 (2008).
5. Ptitsyn, O. B. Molten globule and protein folding. *Adv. Protein Chem.***47**, 83–229 (1995).
6. Dahiyat, B. I. & Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **278**, 82–87 (1997).
7. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. A large test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449–460 (2003).
8. Samish, I., MacDermaid, C. M., Perez-Aguilar, J. M. & Saven, J. G. Theoretical and computational protein design. *Annu. Rev. Phys. Chem.* **62**, 129–149 (2011).
9. Pokala, N. & Handel, T. M. Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. *Protein Sci.* **13**, 925–936 (2004).
10. Li, Z., Yang, Y., Zhan, J., Dai, L. & Zhou, Y. Energy functions in de novo protein design: current challenges and future prospects. *Annu. Rev. Biochem.* **42**, 315–335 (2013).
11. Alford, R. F. *et al.* The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

12. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
13. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. High-resolution protein design with backbone freedom. *Science* **1998**, 1462–1467 (1998).
14. Dantas, G. *et al.* High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. *J. Mol. Biol.* **366**, 1209–1221 (2007).
15. Johansson, K. E. *et al.* Computational redesign of thioredoxin is hypersensitive toward minor conformational changes in the backbone template. *J. Mol. Biol.* **428**, 4361–4377 (2016).
16. Xiong, P. *et al.* Protein design with a comprehensive statistical energy function and boosted by experimental selection for fold-ability. *Nat. Commun.* **5**, 5330 (2014).
17. Tian, P., Louis, J. M., Baber, J. L., Aniana, A. & Best, R. B. Co-evolutionary fitness landscapes for sequence design. *Angew. Chem.* **57**, 5674–5678 (2018).
18. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
19. MacKerell, A. D. Jr. Atomistic models and force fields. In *Computational Biochemistry and Biophysics* (eds Becker, O. *et al.*) (Marcel Dekker, New York, 2001).
20. Roux, B. & Simonson, T. Implicit solvent models. *Biophys. Chem.* **78**, 1–20 (1999).
21. Barth, P., Alber, T. & Harbury, P. B. Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. *Proc. Natl. Acad. Sci. USA* **104**, 4898–4903 (2007).
22. Simmerling, C., Strockbine, B. & Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **124**, 11258–11259 (2002).
23. Simonson, T., Carlsson, J. & Case, D. A. Proton binding to proteins: pK_a calculations with explicit and implicit solvent models. *J. Am. Chem. Soc.* **126**, 4167–4180 (2004).
24. Li, J. *et al.* The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins* **79**, 2794–2812 (2011).
25. Panel, N., Sun, Y. J., Fuentes, E. J. & Simonson, T. A simple PB/LIE free energy function accurately predicts the peptide binding specificity of the Tiam1 PDZ domain. *Front. Mol. Biosci.* **4**, Art. 65 (2017).
26. Cochran, F. V. *et al.* Computational de novo design and characterization of a four-helix bundle that selectively binds a nonbiological cofactor. *J. Am. Chem. Soc.* **127**, 1346–1347 (2005).
27. Fry, H. C. *et al.* Computational de novo design and characterization of a protein that selectively binds a highly hyperpolarizable abiological chromophore. *J. Am. Chem. Soc.* **135**, 13914–13926 (2013).
28. Shepherd, T. R. & Fuentes, E. J. Structural and thermodynamic analysis of PDZ-ligand interactions. *Methods Enzymol.* **488**, 81–100 (2011).
29. McLaughlin, R. N. Jr., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **458**, 859–864 (2012).
30. Melero, C., Ollikainen, N., Harwood, I., Karpiak, J. & Kortemme, T. Quantification of the transferability of a designed protein specificity switch reveals extensive epistasis in molecular recognition. *Proc. Natl. Acad. Sci. USA* **111**, 15426–15431 (2014).
31. Cornell, W. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).
32. Lopes, A., Aleksandrov, A., Bathelt, C., Archontis, G. & Simonson, T. Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins* **67**, 853–867 (2007).
33. Mignon, D., Panel, N., Chen, X., Fuentes, E. J. & Simonson, T. Computational design of the Tiam1 PDZ domain and its ligand binding. *J. Chem. Theory Comput.* **13**, 2271–2289 (2017).
34. Simonson, T. *The Proteus Software for Computational Protein Design* (Ecole Polytechnique, Paris, 2019); https://proteus.polytechnique.fr. Accessed 22 May 2020.
35. Wilson, D., Madera, M., Vogel, C., Chothia, C. & Gough, J. The SUPERFAMILY database in 2007: families and functions. *Nucl. Acids Res.* **35**, D308–D313 (2007).
36. Andreeva, A. *et al.* Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* **36**, 419–425 (2008).
37. Simonson, T. *et al.* Computational protein design: the Proteus software and selected applications. *J. Comput. Chem.* **34**, 2472–2484 (2013).
38. Villa, F., Mignon, D., Polydorides, S. & Simonson, T. Comparing pairwise-additive and many-body generalized born models for acid/base calculations and protein design. *J. Comput. Chem.* **38**, 2396–2410 (2017).
39. Ben-Naim, A. *Hydrophobic Interactions* (Plenum Press, New York, 1980).
40. Chothia, C. Hydrophobic bonding and accessible surface area in proteins. *Nature* **248**, 338–339 (1974).
41. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–20 (2011).
42. Mobley, D. L., Bayly, C. I., Cooper, M. D., Shirts, M. R. & Dill, K. A. Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. *J. Chem. Theory Comput.* **5**, 350–358 (2009).
43. Gaillard, T. & Simonson, T. Full protein sequence redesign with an mmgbsa energy function. *J. Chem. Theory Comput.* **13**, 4932–4943 (2017).

Acknowledgements

Some of the calculations were run at the CINES supercomputer center of the French Ministry of Education and Research. We thank Jay Nix and the staff at beamline 4.2.2 (Advanced Light Source, Lawrence Berkeley National Laboratory), Lokesh Gakhar (U of Iowa Protein and Crystallography Facility) and Liping Yu (U of Iowa Carver College of Medicine NMR Facility) for helpful advice. X-ray software used here was installed by SBGrid. We thank the Roy J. Carver Charitable Trust for funding the Carver College of Medicine Medical NMR Facility.

Author contributions

V.O., N.P.: performed design calculations, analyzed and interpreted simulation results. Y.J.S., T.U.: performed biophysical experiments, including X-ray crystallography, analyzed and interpreted results. E.J.F., T.S.: supervised project, analyzed and interpreted results, wrote paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-67972-w.

Correspondence and requests for materials should be addressed to E.J.F. or T.S.

Supplementary Material: A physics-based energy function allows the computational redesign of a PDZ domain

Vaitea Opuu[a,†], Young Joo Sun[b,†], Titus Hou[b], Nicolas Panel[a], Ernesto J. Fuentes[b,*] & Thomas Simonson[a,*]

[a]Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, Palaiseau, France
[b]Dept. of Biochemistry, Carver College of Medicine, University of Iowa, Iowa City, USA

Below, we first provide Material and Methods. Then, we provide supplementary Results. We report sequence similarities between the Proteus designed sequences and the CASK sequence. We provide information on the stability and flexibility of the CASK-based designs in microsecond molecular dynamics (MD) simulations. We report the experimental characterization of PDZ sequences designed using the Tiam1 template structure and the NEA electrostatics model. Finally, we report the crystallographic structure statistics for the apo CASK PDZ domain.

Table of contents

1 Material and Methods

1.1 Computational design methods

Energy function for the folded state

We used the following energy function for the folded state:

$$E = E_{\mathrm{MM}} + E_{\mathrm{GB}} + E_{\mathrm{SA}} \tag{1}$$

E_{MM} is the protein internal energy, taken from the Amber ff99SB molecular mechanics (MM) energy function [1]. E_{GB} is a Generalized Born (GB) implicit solvent contribution [2, 3]:

$$E_{\mathrm{GB}} = \frac{1}{2}\left(\frac{1}{\epsilon_W} - \frac{1}{\epsilon_P}\right)\sum_{ij} q_i q_j \left(r_{ij}^2 + b_i b_j \exp[-r_{ij}^2/4b_i b_j]\right)^{-1/2} \tag{2}$$

Here, ϵ_W and ϵ_P are the solvent and protein dielectric constants (80 and 4, respectively); r_{ij} is the distance between atoms i, j and b_i is the "solvation radius" of atom i [2, 4]. The dependency of the b_i on the protein conformation corresponds to a GB variant we call GB/HCT (for "Hawkins-Cramer-Truhlar") [2, 4]. For some of the design calculations, an additional "Native Environment Approximation", or NEA was used for efficiency [3, 5], where the solvation radius b_i of each particular group (backbone, sidechain or ligand) was computed ahead of time, with the rest of the system having its native sequence and conformation [6, 7]. For the other designs, we computed the solvation radii on the fly during the MC simulation, using a very fast implementation called "Fluctuating Dielectric Boundary," or FDB [7] that uses lookup tables.

The last term in Eq. (1) is a surface area term:

$$E_{\mathrm{SA}} = \sum_i \sigma_i A_i \tag{3}$$

A_i is the exposed solvent accessible surface area of atom i; σ_i is a parameter that reflects each atom's preference to be exposed or hidden from solvent. The solute atoms were divided into four groups with specific σ_i values. The values were -60 (nonpolar), 30 (aromatic), -120 (polar), and -110 (ionic) cal/mol/Å^2. The coefficient for hydrogens was zero. Negative values are physically correct, since the SA term includes favorable protein-solvent dispersion interactions, in addition to hydrophobic effects. Surface areas were computed by the Lee and Richards algorithm [8], implemented in the Proteus software [5], using a 1.5 Å probe radius. Surface burial is not additive, since the same area on a

side chain can be buried by two other residues. To avoid overcounting, a scaling factor was applied to the contact areas involving at least one buried side chain [9]. In previous tests, a value of 0.65 gave the best surface areas, compared to an exact calculation [3, 4].

The unfolded state energy

For a sequence S, the unfolded energy is:

$$E^u = \sum_{i \in S} E^u(t_i, B_i). \tag{4}$$

The sum is over all amino acids; t_i represents the side chain type at position i; B_i represents the buried or exposed character of position i in the folded state. The quantities $E^u(t, B) \equiv E^u_t$ an be thought of as effective chemical potentials of each amino acid type. Their values were chosen empirically, to maximize the likelihood of a set of experimental PDZ sequences. This means that an MC simulation should give overall amino acid frequencies that match those in the experimental sequences [10]. We assigned different values to buried and exposed positions, because we assume residual structure is present in the unfolded state, so that amino acids partly retain their buried/exposed character. Thus, the simulations should reproduce the overall composition of the buried and exposed positions separately. To define the target amino acid frequencies for likelihood maximization, we used a set of PDZ sequences collected earlier [10]. CASK positions were considered buried or exposed based on their solvent-accessible surface area in the CASK 3D structure, with a threshold designed to place roughly half of the positions in either category. Positions in the other PDZ sequences were considered buried or exposed based on a sequence alignment that included CASK: positions aligned with a buried CASK position were buried. Likelihood maximization was initiated with $E^u(t, B)$ values obtained from a non-empirical, tripeptide model [10, 11]. The first iterations then optimized the frequencies of 11 groups of homologous amino acid types [10]. This corresponds to 20 independent, adjustable, unfolded energies (10×2 independent groups). The values after convergence (16 iterations) are reported in Table S1. They differ only moderately from the initial, non-empirical values. 5 more iterations were done to optimize the individual type frequencies. This corresponds to 34 adjustable unfolded energies (17 independent types, since Gly and Pro were not allowed, × 2 regions) [10]. In these iterations, the energies changed very little, by 0.15 kcal/mol on average. Thus, while the number of parameters is large, the departure from the non-empirical values is very small.

Table S1: Unfolded energies (kcal/mol)

	Exposed positions			Buried positions		
	initial[a]	interm.[b]	final[c]	initial[a]	interm.[b]	final[c]
ALA	0.00	0.00	0.00	0.00	0.00	0.00
ARG	-54.76	-56.54	-56.85	-51.37	-51.85	-52.00
ASN	-17.80	-20.01	-20.13	-14.02	-14.34	-14.44
ASP	-18.82	-19.95	-20.11	-14.55	-14.57	-14.76
CYS	-1.64	-1.64	-1.78	-1.06	-1.06	-1.01
GLN	-16.61	-18.82	-19.25	-13.14	-13.46	-13.53
GLU	-18.21	-19.34	-19.80	-14.52	-14.54	-14.52
HIS_δ	7.37	6.94	6.66	10.41	10.57	10.54
HIS_ϵ	8.12	7.69	7.41	10.85	11.01	10.98
HIS^+	10.98	10.55	10.27	12.86	13.02	12.99
ILE	3.06	2.48	2.41	5.50	5.40	5.43
LEU	-2.94	-3.52	-4.03	0.00	-0.10	-0.09
LYS	-11.35	-10.69	-10.88	-8.24	-7.70	-7.65
MET	-3.09	-3.94	-4.26	-2.85	-2.09	-1.90
PHE	-3.18	-3.27	-3.32	0.17	0.77	0.93
SER	-5.24	-5.23	-5.46	-4.45	-4.24	-4.26
THR	-6.68	-6.68	-7.09	-4.84	-4.84	-4.96
TRP	-5.53	-5.62	-5.74	-1.94	-1.34	-1.30
TYR	-10.14	-10.29	-10.36	-5.91	-5.56	-5.50
VAL	-1.66	-2.24	-2.25	-0.05	-0.15	-0.30

[a]Initial values from tripeptide model. [b]Optimized for 11 groups of amino acid types (20 independent parameters). [c]Optimized for each amino acid type, which corresponds to 34 independent, adjustable parameters. With respect to the 20-parameter stage, the mean E^u change was just 0.15 kcal/mol.

Structural model and energy matrix

For CASK, we used a new X-ray structure of the apo PDZ domain, reported here (PDB entry 6NH9). To carry out the MC simulations, an energy matrix was computed using procedures described previously [10]. Briefly, for each pair of amino acid side chains, the interaction energy was computed after 15 steps of energy minimization, with the backbone held fixed and only the interactions of the pair with each other and the backbone included [12]. Side chain rotamers were described by the Tuffery library [13], expanded to include additional hydrogen orientations for OH and SH groups [3]. The energies were stored in an energy table, or "matrix" for use during MC.

Monte Carlo simulations

Sequence design was performed by running long MC simulations where 61 out of 83 positions could mutate freely: all but 7 Gly, 2 Pro and 13 positions that are directly involved in binding the peptide ligand. Non-mutating positions could explore different rotamers. The MC simulations used one- and two-position moves, where either rotamers, amino acid types, or both changed. For two-position moves, the second position was near the first in space. Sampling was enhanced by using Replica Exchange Monte Carlo (REMC), where eight MC simulations ("replicas") were run in parallel, at different temperatures [14]. Periodic swaps were attempted between the conformations of two replicas i, j (adjacent in temperature), subject to a Metropolis acceptance test [14]. Thermal energies ranged from 0.125 to 3 kcal/mol. Simulations were done with the Proteus software [5, 14].

Molecular dynamics simulations

Wild-Type CASK and six sequences designed with Proteus were subjected to MD simulations with explicit solvent and no peptide ligand. The starting structures were taken from the MC trajectory or the crystal structure and slightly minimized with harmonic restraints to maintain the backbone geometry. Each protein was immersed in a solvent box using the CHARMM GUI [15, 16]. The boxes had a truncated octahedral shape. The minimum distance between protein atoms and the box edge was 15 Å. The final models included about 11,000 water molecules. A few sodium or chloride ions were included to ensure overall electroneutrality. The protonation states of histidines were assigned to be neutral, based on visual inspection. MD was performed with periodic boundary conditions, at room temperature and pressure, using Langevin dynamics with a Langevin

Piston Nosé-Hoover barostat [17, 18]. Long-range electrostatic interactions were treated with a Particle Mesh Ewald approach [19]. The Amber ff14SB force field and the TIP3P model [20] were used for the protein and water, respectively. Simulations were run for one microsecond, using the Charmm and NAMD programs [16, 21].

1.2 Protein expression and purification

The codon optimized gene of the human CASK PDZ domain (residues 487–572) was chemically synthesized (GenScript Inc., Piscataway, NJ) and ligated into the pET28a vector (Novagen). The DNA sequence of the pET28a-CASK PDZ vector was verified by automated DNA sequencing (University of Iowa, DNA Facility). Protein expression was conducted in BL21(DE3) (Invitrogen) *E. coli* cells. Typically, *E. coli* cells were grown at 37°C in Luria-Bertani (LB) medium supplemented with kanamycin (15 μg/mL) under vigorous agitation until an absorbance at 600 nm wavelength (A600) reached 0.6-0.8. Cultures were subsequently cooled to 18°C and protein expression was induced by the addition of isopropyl 1-thio-β-d-galactopyranoside (IPTG) to 1 mM final concentration. Induced cells were incubated for an additional 16–18 hrs at 18°C. and harvested by centrifugation. The CASK PDZ domain was purified by cation exchange (SP media, GE-Healthcare) and size-exclusion chromatography (GE-Healthcare). Superdex 75 (S75) size-exclusion chromatography was performed with desired final buffer (20 mM phosphate, pH 6.8, 50 mM NaCl, and 0.5 mM EDTA). The final yield was 50 mg of CASK PDZ protein >98% pure as judged by SDS-PAGE from 1 L of culture. Samples were used immediately or lyophilized and stored at -80°C. The Tiam1 PDZ domain was purified as previously published [22].

The genes encoding the Proteus PDZ designs were codon-optimized for *E. coli* expression and chemically synthesized by GenScript Inc. (Piscataway, NJ). The genes were cloned into a modified pET21a vector (Novagen) that contains a His_6-tag and Tobacco etch virus protease cleavage site at the 5′-end of the multiple cloning site. The nucleotide coding sequence of the pET21a-PDZ vector was verified by automated DNA sequencing (University of Iowa, DNA Facility). Protein expression was conducted in BL21(DE3) (Invitrogen) *E. coli cells*. Typically, cells were grown at 37°C in Luria-Bertani medium supplemented with ampicillin (100 μg/mL) under vigorous agitation until an A600 of 0.6–0.8 was reached. Cultures were subsequently cooled to 18°C and protein expression was induced by the addition of IPTG to 1 mM final concentration. Induced cells were incubated for an additional 16-18 hrs at 18°C and harvested by centrifugation. Proteins

were initially purified by nickel-chelate chromatography (GE-Healthcare). The proteins were further purified by size-exclusion chromatography (Superdex 75, GE Healthcare) using a buffer containing 20 mM phosphate, pH 6.8, 50 mM NaCl, and 0.5 mM EDTA. Samples were used immediately.

1.3 Crystal structure of the wild-type apo CASK PDZ domain

A crystal structure of the apo CASK PDZ domain was determined in this work. High-throughput hanging-drop, vapor-diffusion screens using a Mosquito drop setter (TTP LabTech) were used to determine the crystallization conditions. The CASK PDZ domain was prepared in 20 mM Tris pH 7.5 and 50 mM NaCl. 200 nL of precipitant and PDZ domain (10-30 mg/mL) was used for each screening condition. Initial screening for diffracting crystals was done with an in-house Rigaku RAXIS-IV rotating anode X-ray source. Collection of full X-ray diffraction datasets for structure determination was done at beamline 4.2.2 at the Advanced Light Source (Berkeley, CA). Proper space group handedness was verified by analysis of the electron density.

XDS was used for indexing, integration, and scaling of the diffraction data [23, 24], to 1.85 Å resolution. XSCALE was used to merge multiple datasets. We used PHASER and previously-determined PDZ structures for initial phasing [25]. We used Refmac [26, 27] for the early stages of refinement and PHENIX [28, 29] for the final refinement. Refinement statistics are given in Supplementary Information (Table S1). Manual model building was done based on visualized electron density in Coot [30, 30]. 4.6% of the reflections were randomly selected to be excluded from the refinement and used to calculate R_{free} values. Alignment of structures and generation of figures were done with PyMOL (Schrodinger, LLC, The PyMOL Molecular Graphics System).

1.4 Biophysical characterization of designed proteins

Synthetic peptides

All peptides were chemically synthesized by GenScript Inc. (Piscataway, NJ) and were >95% pure as judged by analytical HPLC and mass spectrometry. Peptides were dansylated at the N-terminus and had a free carboxyl at the C-terminus. The peptides used in this study were derived from the following proteins: Neurexin (residues 1,470–1,477: NKDKEYYV$_{COOH}$), Caspr4 (residues 1,301–1,308: ENQKEYFF$_{COOH}$) and Syndecan1

(residues 303–310: TKQEEFYA$_{COOH}$).

Circular dichroism

Circular dichroism signals were measured using a Jasco J-815 circular dichroism spectropolarimeter. The concentration of each protein ranged from 10 to 20 μM. All proteins were in a buffer composed of 20 mM phosphate, pH 6.8, 50 mM NaCl, and 0.5 mM EDTA. Spectra were taken from the 190 nm to 260 nm wavelength window with a 1 nm data interval at 25°C. Data integration time was 2 seconds and the scanning speed was 100 nm/min.

NMR

Nuclear magnetic resonance (NMR) experiments were carried out at 298 K (calibrated with methanol) on Bruker Avance II 800 MHz (equipped with a CryoProbe), Bruker Avance II 500 MHz, and Varian 600 MHz spectrometers (equipped with room temperature probes). All protein samples were prepared in 20 mM phosphate, pH 6.8, 50 mM NaCl, 0.5 mM EDTA, and 10% (v/v) D_2O with a concentration of 14 μM to 22 μM.

Differential scanning fluorimetry

Standard methodology was used for differential scanning fluorimetry (DSF) [31, 32]. Briefly, DSF was performed using 96-well PCR plates and the Sypro Orange (Thermo Fisher) dye. Each well in the PCR plate had a 20 μL final volume containing 0.25 mg/mL of protein, 300 μM of peptide, and 5x Sypro Orange final concentration (from a 5000x stock) in a buffer containing 20 mM phosphate, pH 6.8, 50 mM NaCl, and 0.5 mM EDTA. The DSF assays were performed using a Bio-Rad CFX96 real-time polymerase chain reaction instrument equipped to read 96-well plates. The protein of interest was thermally denatured from 5°C to 95°C at a ramp rate of 1°C/min. The protein melting/unfolding curves were generated by monitoring changes in Sypro Orange fluorescence (at 610 nm wavelength). Raw fluorescence data were analyzed using DMAN, and the first derivative value from the denaturation data was used to determine the apparent melting temperature [33] ($T_{1/2}$). Each peptide was assayed in triplicate. A 96-well plate containing no peptide was assayed to determine the apparent $T_{1/2}$ of each PDZ domain in the absence of any peptide. A shift of more than 1°C in $T_{1/2}$ indicates binding (based on SEM).

2 Supplementary Results

2.1 Sequence similarities between designed sequences and CASK

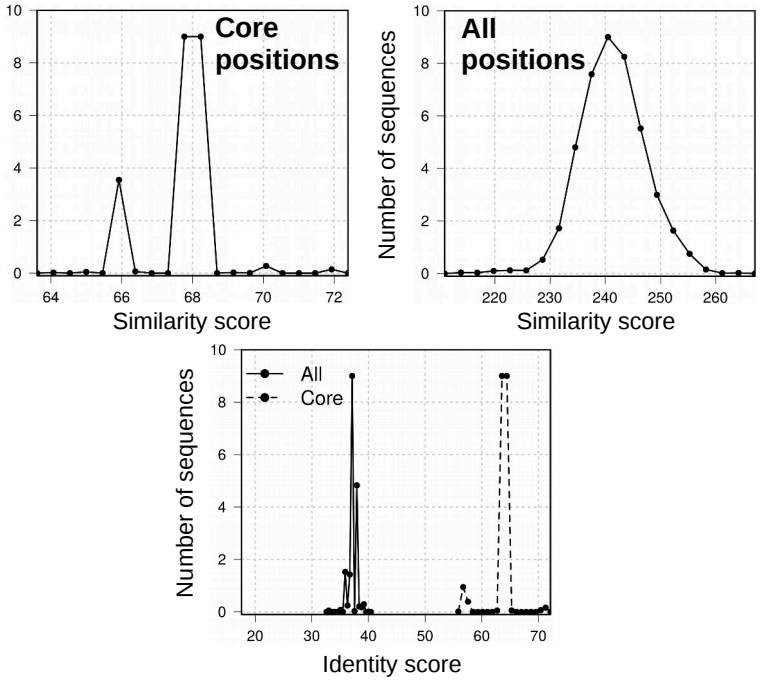


Figure S1: Histograms of Blosum40 similarity scores (above) and sequence identities (below) compared to CASK, for the 2000 lowest-energy designed sequences.

2.2 Stability of the three selected CASK-based designs in molecular dynamics

As a first test of the three selected sequences, FDB1350, FDB1555, and FDB1669, they were subjected to MD simulations using an explicit solvent environment, for 1000 ns. Wild-Type CASK (WT) was also simulated. Convergence of the simulations was good (based on a principal component analysis, not shown). The WT protein was quite stable, with rms deviations from the starting, X-ray structure of 1–1.5 Å (excluding 3–4 residues at each terminus and one very flexible loop, residues 495–502; see Fig. S2). Deviations from its own mean MD structure were similar (Fig. S2). The designed proteins exhibited only slightly larger deviations from the WT X-ray structure (1.2–1.8 Å) and similar, small deviations from their respective mean MD structures, with no visible drift (Fig. S2).

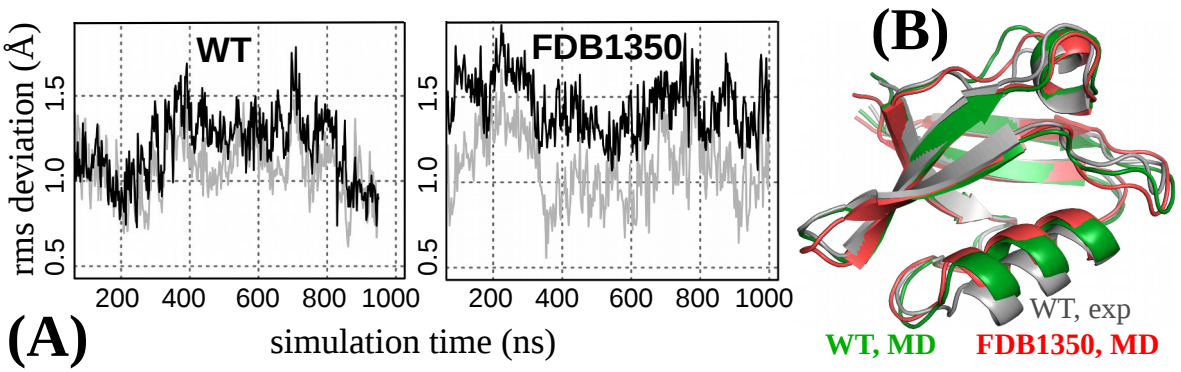


Figure S2: MD simulations of CASK-based designs. **A)** Backbone rms deviations for WT and the FDB1350 designed variant relative to the starting structure (black) and the mean MD structure (grey). **B)** Mean MD structures of WT and designed variant FDB1350.

We also characterized the backbone flexibility of the designed proteins by computing NMR order parameters for the backbone amide groups (Fig. S3). Experimental values were not available for WT CASK, but were available for Tiam1 and a quadruple mutant of Tiam1 [34]. These proteins were also simulated by MD for one microsecond, with and without the peptide ligands Sdc1 and Caspr4, respectively. In Fig. S3, we show the order parameters for both proteins in the apo and holo states, from experiment (circles) and MD (continuous lines) (top two panels). The agreement is very good. Next, we show (Fig. S3, bottom panel) the order parameters for WT CASK and the three selected CASK-based designs, FDB1350, FDB1555, and FDB1669 (apo proteins). Comparing the designed

proteins to WT CASK, the results were similar, with some differences in loop regions. Two designs were slightly less flexible than WT (see positions 492-502 in β_1-β_2, 521-524 in β_3-α_1), while FDB1350 was slightly more flexible (see 492–496 in β_1-β_2 and 559-561 in α_2-β_5). Evidently, the design calculations do not produce overly-rigid or overly-flexible proteins in a systematic way.

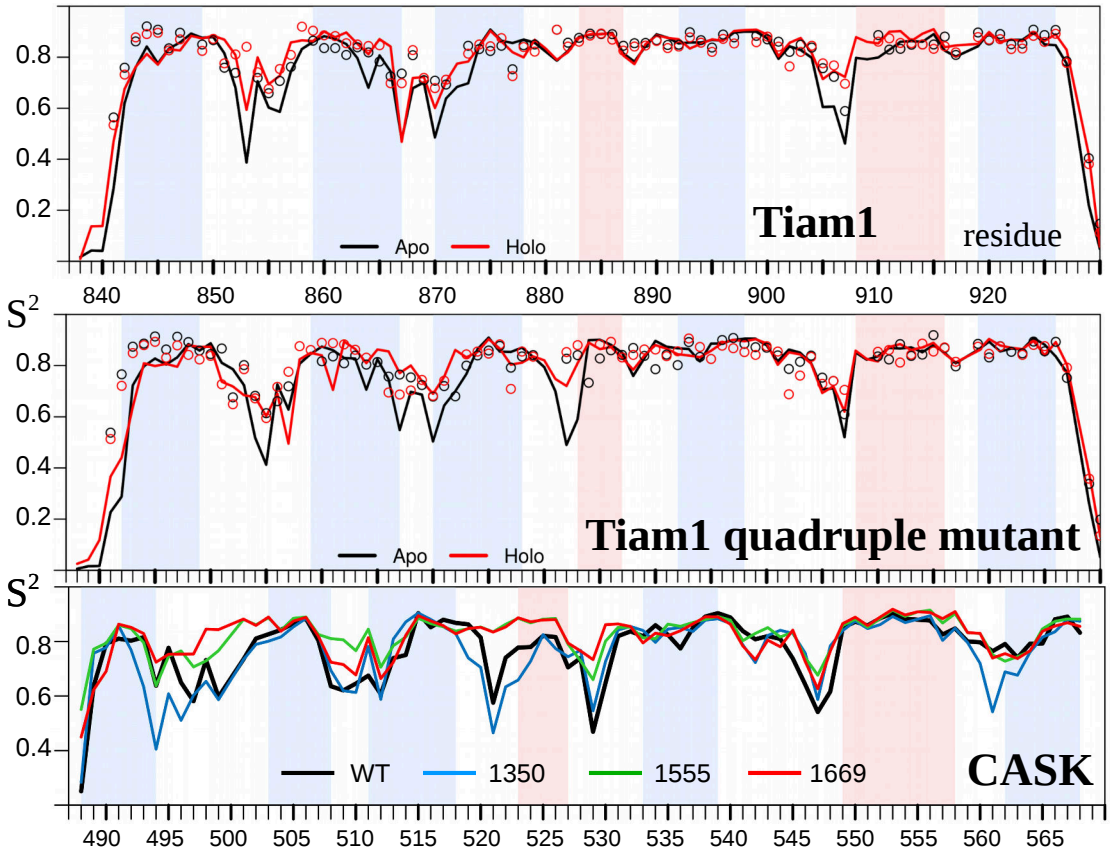


Figure S3: Backbone amide NMR order parameters for natural and designed proteins. **Top panel:** Tiam1 with and without the Sdc1 peptide ligand. Circles are experimental values; lines are from μsec MD simulations. **Middle panel:** analogous data for the Tiam1 quadruple mutant and the Caspr4 peptide. **Bottom panel:** Apo WT CASK and the three designed variants; values from MD.

2.3 Experimental characterization of Proteus designs obtained with the Tiam1 template and the NEA electrostatic model

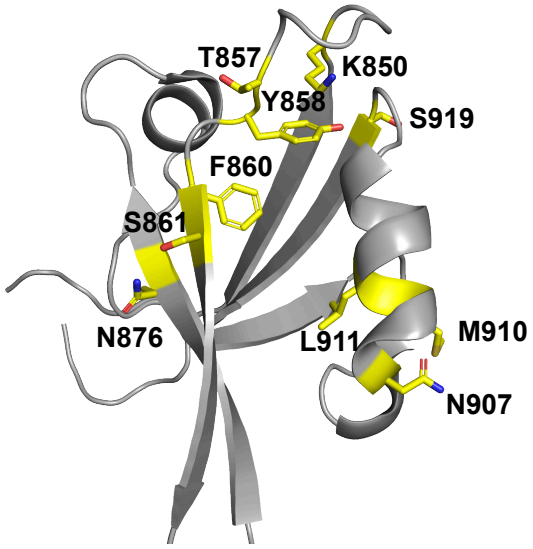


Figure S4: Tiam1 structure. Yellow: 13 positions whose types were fixed in the Proteus designs.

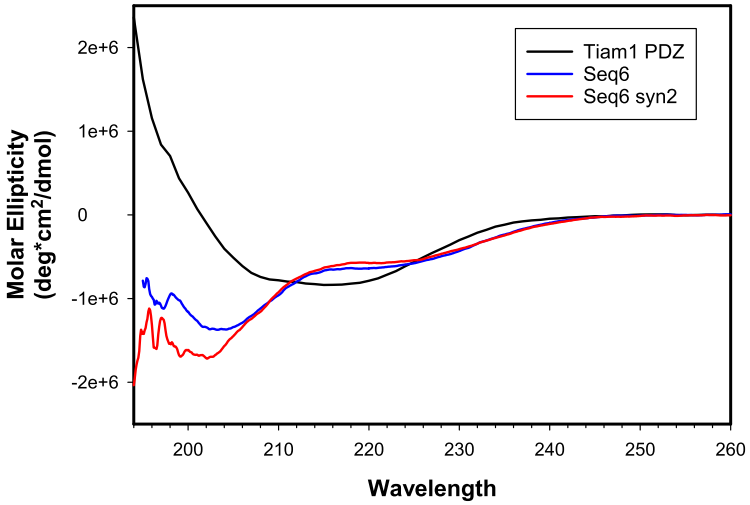


Figure S5: CD spectra of Tiam1 and two designs based on the Tiam1 template and NEA electrostatics.

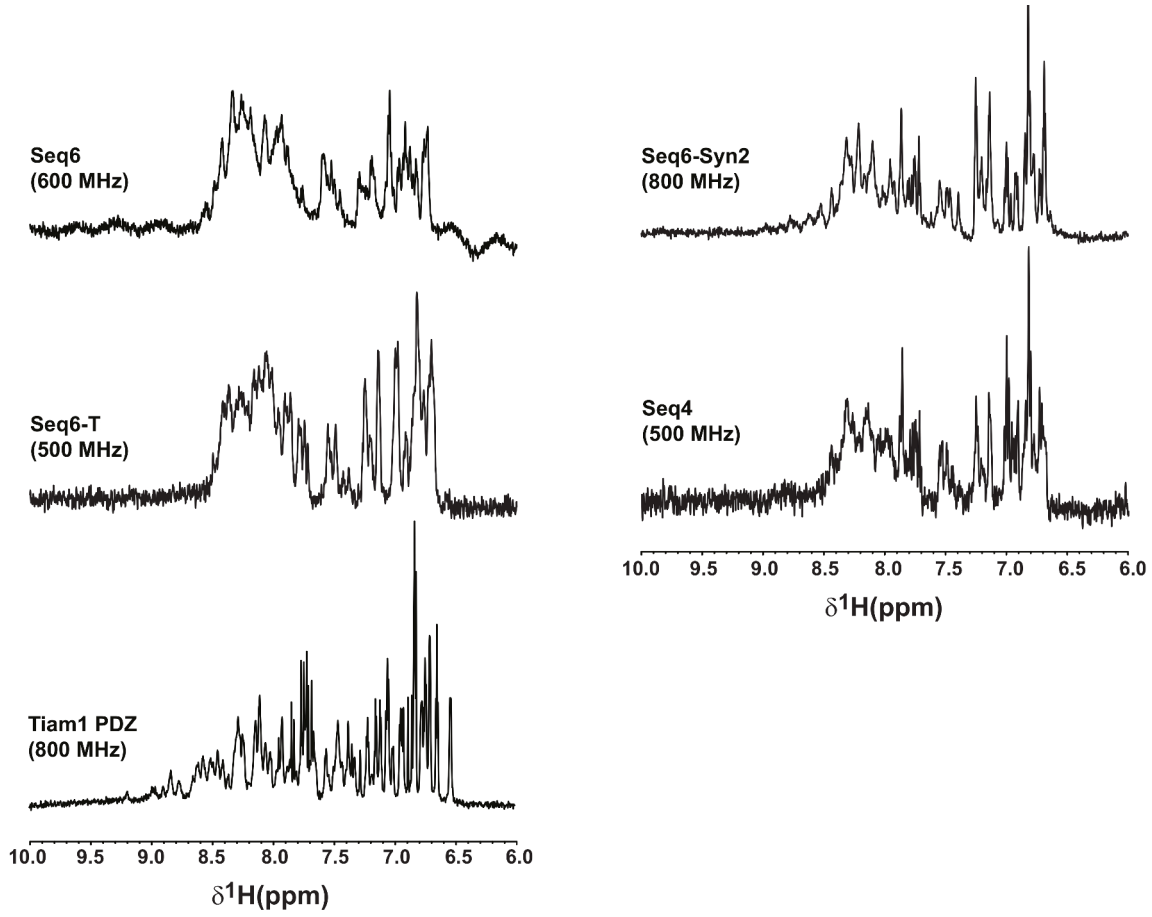


Figure S6: Proton NMR spectra of the Tiam1 PDZ domain and four designs obtained with the Tiam1 template and NEA electrostatics.

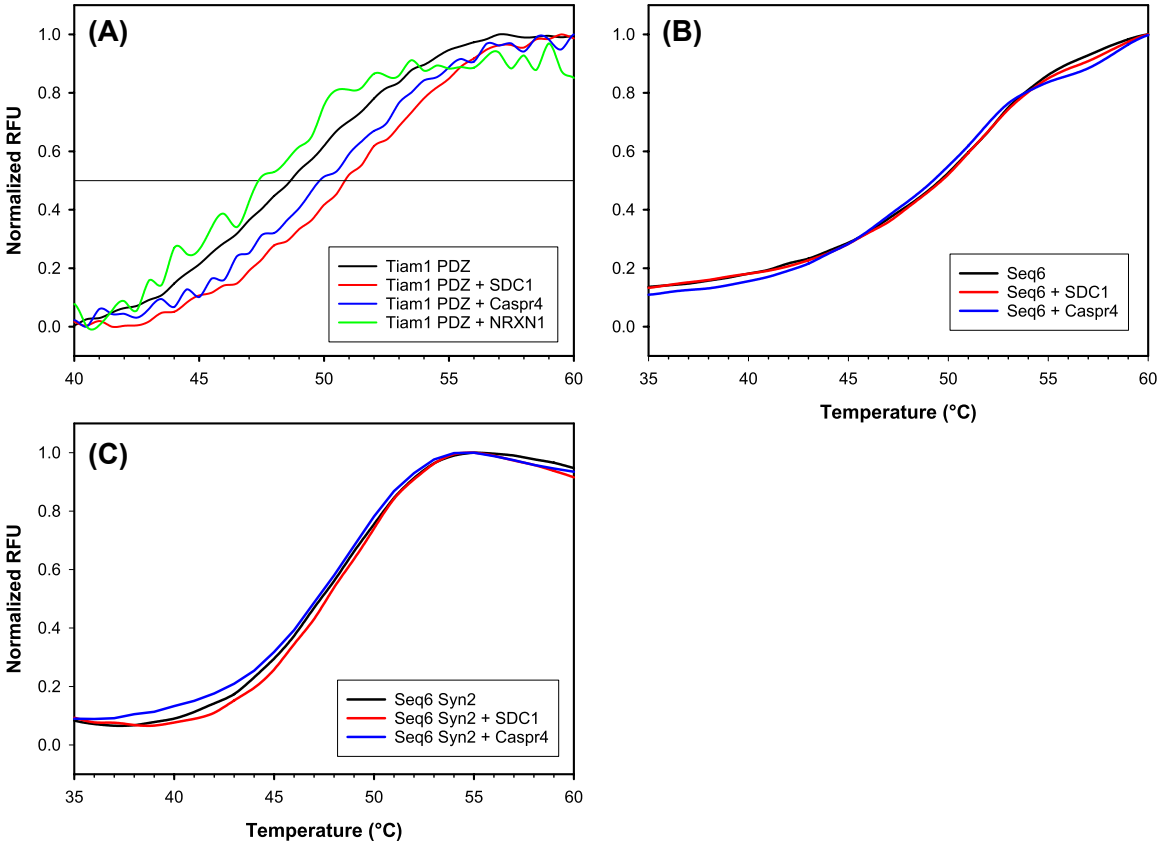


Figure S7: Differential scanning fluorimetry of a natural PDZ domain (Tiam1) and two designs based on the Tiam1 template and the NEA electrostatic model. Signals in the absence and presence of the SDC1, Capr4 and NRXN peptides.

2.4 Human apo CASK PDZ domain X-ray structure statistics

Table S2: Crystallographic statistics for the human apo CASK PDZ domain

Data collection statistics	
Beam line	ALS 4.2.2
Wavelength (Å)	1.0003
Space group	C 1 2 1
Unit cell dimensions (a, b, c) (Å)	61.1, 35.4, 119.5
Unit cell dimensions (α, β, γ)	90°, 90.3°, 90°
Resolution range (Å)	59.8—1.85
Total reflections	37,385 (7,461)
Unique reflections	20,769 (1,910)
Multiplicity	1.8 (1.7)
Completeness (%)	93.7 (93.7)
I/σ (I)	10.4 (2.1)
Wilson B-factor (Å^2)	50.7
Rmeas	0.030 (0.402)
$CC_{1/2}$	99.8 (91.1)
Refinement statistics	
Resolution (Å)	1.85
No. of reflections used in refinement	20,739 (2,705)
No. of reflections used for R_{free}	964 (133)
R_{work}/R_{free}	0.226/0.263
No. of atoms (Protein/Water)	4,188 (4,037/151)
B-factors (Å^2)	53.0
R.M.S.D.[a]	
Bond length (Å)	0.29
Bond angle (degrees)	0.46
Ramachandran plot statistics (%)	
In preferred regions	98.0
In allowed regions	2.0
Outliers	0.0
PDB accession code	6NH9

Numbers in parentheses are for the highest-resolution shell. [a]RMS deviation from ideal values.

References

[1] Cornell, W. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).

[2] Hawkins, G. D., Cramer, C. & Truhlar, D. Pairwise descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **246**, 122–129 (1995).

[3] Gaillard, T. & Simonson, T. Pairwise decomposition of an MMGBSA energy function for computational protein design. *J. Comput. Chem.* **35**, 1371–1387 (2014).

[4] Lopes, A., Aleksandrov, A., Bathelt, C., Archontis, G. & Simonson, T. Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins* **67**, 853–867 (2007).

[5] Simonson, T. *et al.* Computational protein design: the Proteus software and selected applications. *J. Comput. Chem.* **34**, 2472–2484 (2013).

[6] Simonson, T. Protein:ligand recognition: simple models for electrostatic effects. *Curr. Pharma. Design* **19**, 4241–4256 (2013).

[7] Villa, F., Mignon, D., Polydorides, S. & Simonson, T. Comparing pairwise-additive and many-body generalized born models for acid/base calculations and protein design. *J. Comput. Chem.* **38**, 2396–2410 (2017).

[8] Lee, B. & Richards, F. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400 (1971).

[9] Dahiyat, B. I. & Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **278**, 82–87 (1997).

[10] Mignon, D., Panel, N., Chen, X., Fuentes, E. J. & Simonson, T. Computational design of the Tiam1 PDZ domain and its ligand binding. *J. Chem. Theory Comput.* **13**, 2271–2289 (2017).

[11] Pokala, N. & Handel, T. M. Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol.* **347**, 203–227 (2005).

[12] Schmidt am Busch, M., Lopes, A., Mignon, D. & Simonson, T. Computational protein design: software implementation, parameter optimization, and performance of a simple model. *J. Comput. Chem.* **29**, 1092–1102 (2008).

[13] Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.* **8**, 1267–1289 (1991).

[14] Mignon, D. & Simonson, T. Comparing three stochastic search algorithms for computational protein design: Monte Carlo, Replica Exchange Monte Carlo, and a multistart, steepest-descent heuristic. *J. Comput. Chem.* **37**, 1781–1793 (2016).

[15] Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).

[16] Brooks, B. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).

[17] Martyna, G. J., Tobias, D. J. & Klein, M. L. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **101**, 4177–4189 (1994).

[18] Feller, S. E., Zhang, Y., Pastor, R. W. & Brooks, B. R. Constant pressure molecular dynamics simulation: the Langevin piston method. *J. Chem. Phys.* **103**, 4613–4622 (1995).

[19] Darden, T. Treatment of long-range forces and potential. In Becker, O., MacKerrell Jr., A. D., Roux, B. & Watanabe, M. (eds.) *Computational Biochemistry & Biophysics*, chap. 4 (Marcel Dekker, N.Y., 2001).

[20] Jorgensen, W. L., Chandrasekar, J., Madura, J., Impey, R. & Klein, M. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

[21] Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).

[22] Shepherd, T. R. *et al.* The Tiam1 PDZ domain couples to Syndecan1 and promotes cell-matrix adhesion. *J. Mol. Biol.* **398**, 730–746 (2010).

[23] Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Cryst. D* **66**, 133–144 (2010).

[24] Kabsch, W. XDS. *Acta Cryst. D* **66**, 125–132 (2010).

[25] McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658–674 (2007).

[26] Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum likelihood method. *Acta Cryst.* **D53**, 240–255 (1997).

[27] Vagin, A. A. *et al.* REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Cryst. D* **60**, 2184–2195 (2004).

[28] Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Cryst. D* **58**, 1948–1954 (2002).

[29] Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst. D* **66**, 213–221 (2010).

[30] Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Cryst. D* **66**, 486–501 (2010).

[31] Ehrhardt, M. K. G., Warring, S. L. & Gerth, M. L. Screening chemoreceptor-ligand interactions by high-throughput thermal-shift assays. Methods. *Methods Molec. Biol.* **1729**, 281–290 (2018).

[32] Kranz, K. K. & Schalk-Hihi, C. Protein thermal shifts to identify low molecular weight fragments. *Methods Enzym.* **493**, 277–298 (2011).

[33] Wang, C. K., Weeratunga, S. K., Pacheco, C. M. & Hofmann, A. DMAN: a Java tool for analysis of multi-well differential scanning fluorimetry experiments. *Bioinf.* **28**, 439–440 (2012).

[34] Liu, X. *et al.* Distinct roles for conformational dynamics in protein-ligand interactions. *Structure* **24**, 2053–2066 (2016).

Chapter 3

Engineering methionyl-tRNA synthetase for ligand:substrate binding and catalytic power

The following chapter uses the text from the article: ***Adaptive landscape flattening allows the design of both enzyme:substrate binding and catalytic power***, Vaitea Opuu, Giuliano Nigro, Thomas Gaillard, Emmanuelle Schmitt, Yves Mechulam, Thomas Simonson. *Plos Computational Biology*, (2020), 16(1):e1007600.

We report here the redesign of Methionyl-tRNA synthetase (MetRS) binding site for the enzyme/ligand binding using an adaptive Monte Carlo approach. This application allowed us to predict variants that were found active experimentally for the Met ligand. Then, we extended the method to the transition state of the activation reaction. Variants were selected for the first time according to their binding affinity for the transition state.

In complementary work, we studied the effect of native rotamers in the MetRS complex with the transition state. In the work described by the article, we used only the side chain conformations from a rotamer library. We repeated the catalytic efficiency calculations with the native rotamers. We observed a loss in correlation with experiments for the least active variants. However, it seems that it leads to a better selection of true positives.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Adaptive landscape flattening allows the design of both enzyme: Substrate binding and catalytic power

Vaitea Opuu, Giuliano Nigro, Thomas Gaillard, Emmanuelle Schmitt, Yves Mechulam, Thomas Simonson*

Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, Palaiseau, France

* thomas.simonson@polytechnique.fr

OPEN ACCESS

Citation: Opuu V, Nigro G, Gaillard T, Schmitt E, Mechulam Y, Simonson T (2020) Adaptive landscape flattening allows the design of both enzyme: Substrate binding and catalytic power. PLoS Comput Biol 16(1): e1007600. https://doi.org/10.1371/journal.pcbi.1007600

Editor: Alexey Onufriev, Virginia Tech, UNITED STATES

Received: September 20, 2019

Accepted: December 11, 2019

Published: January 9, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Designed enzymes are of fundamental and technological interest. Experimental directed evolution still has significant limitations, and computational approaches are a complementary route. A designed enzyme should satisfy multiple criteria: stability, substrate binding, transition state binding. Such multi-objective design is computationally challenging. Two recent studies used adaptive importance sampling Monte Carlo to redesign proteins for ligand binding. By first flattening the energy landscape of the apo protein, they obtained positive design for the bound state and negative design for the unbound. We have now extended the method to design an enzyme for specific transition state binding, *i.e.*, for its catalytic power. We considered methionyl-tRNA synthetase (MetRS), which attaches methionine (Met) to its cognate tRNA, establishing codon identity. Previously, MetRS and other synthetases have been redesigned by experimental directed evolution to accept noncanonical amino acids as substrates, leading to genetic code expansion. Here, we have redesigned MetRS computationally to bind several ligands: the Met analog azidonorleucine, methionyl-adenylate (MetAMP), and the activated ligands that form the transition state for MetAMP production. Enzyme mutants known to have azidonorleucine activity were recovered by the design calculations, and 17 mutants predicted to bind MetAMP were characterized experimentally and all found to be active. Mutants predicted to have low activation free energies for MetAMP production were found to be active and the predicted reaction rates agreed well with the experimental values. We suggest the present method should become the paradigm for computational enzyme design.

Author summary

Designed enzymes are of major interest. Experimental directed evolution still has significant limitations, and computational approaches are another route. Enzymes must be stable, bind substrates, and be powerful catalysts. It is challenging to design for all these properties. A method to design substrate binding was proposed recently. It used an adaptive Monte Carlo method to explore mutations of a few amino acids near the substrate. A

bias energy was gradually "learned" such that, in the absence of the ligand, the simulation visited most of the possible protein mutations with comparable probabilities. Remarkably, a simulation of the protein:ligand complex, including the bias, will then preferentially sample tight-binding sequences. We generalized the method to design binding *specificity*. We tested it for the methionyl-tRNA synthetase enzyme, which has been engineered in order to expand the genetic code. We redesigned the enzyme to obtain variants with low activation free energies for the catalytic step. The variants proposed by the simulations were shown experimentally to be active, and the predicted activation free energies were in reasonable agreement with the experimental values. We expect the new method will become the paradigm for computational enzyme design.

Introduction

One of the most important challenges in computational protein design (CPD) is to modify a protein so that it will bind a given ligand [1–4]. This is essential for problems like enzyme design, biosensor design, and building tailored protein assemblies. To design ligand binding means optimizing a free energy difference between bound and unbound states. This two-state optimization is not directly tractable by the most common CPD methods, such as simulated annealing, plain Monte Carlo (MC), or simple branch-and-bound and dead end elimination methods [4, 5]. Rather, most studies have used either heuristic methods that optimize the *bound state* energy [1–4, 6], or enumeration methods that are rigorous but expensive and explore a limited free energy range [7–10].

Recently, a new approach was proposed, using Monte Carlo simulation and importance sampling. The energy landscape in sequence space is flattened adaptively over the course of a simulation, thanks to a bias potential [11]. Flattening can be done for the bound state, the unbound state, or both [12]. Remarkably, this leads to a situation where sequence variants are sampled according to a Boltzmann distribution controlled by the *binding free energy*, exactly the quantity we want to select for. Several variations have been employed, including one that used molecular dynamics instead of MC [13]. The method allows sequences to be designed for binding affinity, but also binding *specificity*. This is especially important for enzyme design, since catalytic power is directly related to the enzyme's ability to preferentially stabilize the transition state [14]. We apply the method here to an enzyme of biological and technological importance, methionyl-tRNA synthetase (MetRS). We demonstrate that the method can be used to design an enzyme for its catalytic power.

Each aminoacyl-tRNA synthetase (aaRS) attaches a specific amino acid to a tRNA that carries the corresponding anticodon, establishing the genetic code [15]. Two reactions are catalyzed. In the first, the amino acid reacts with ATP to give aaAMP and pyrophosphate. In the second, tRNA reacts with aaAMP. For MetRS, the first reaction does not require tRNA. Several aaRSs have been engineered experimentally to bind noncanonical amino acids (ncAAs) [16–20]. Obtaining an aaRS that binds an ncAA and uses it as a substrate is a key step to allow the ncAA to become part of an expanded genetic code [17, 20, 21]. The ncAA can then be genetically encoded and incorporated into proteins by the cellular machinery. Several MetRS variants that accepted the ncAA azidonorleucine (AnL) as a substrate were obtained earlier by experimental directed evolution [22]. The AnL azide group can be used for protein labeling and imaging.

The design procedure has two stages. First, a bias potential is optimized adaptively over the course of a MC simulation of the apo protein. The adaptation method is closely analogous to

the Wang-Landau and metadynamics approaches [23, 24]. The bias is chosen so that all the allowed residue types achieve comparable probabilities at all mutating positions. This implies that the free energy landscape in sequence space has been flattened, and the bias of each sequence is approximately the opposite of its apo free energy. In the second stage, the holo state is simulated. The bias is included in the energy function, "subtracting out" the apo free energy. Thus, the method achieves positive design for the bound state and negative design for the unbound. The sequences sampled in the second stage are distributed according to their binding free energies, with tight binders exponentially enriched.

In an analogous procedure, a bias potential can be optimized for the protein bound to one ligand, say L. Then a complex with another ligand is simulated, say, L', including the bias. The sequences sampled preferentially in the second simulation are those with a strong binding free energy difference between the two ligands, *i.e.*, the most *specific* binders. Importantly, L' can be an activated, transition state ligand, while L is the non-activated substrate. In this case, the first simulation flattens the ground state landscape, while the second preferentially samples sequences that stabilize the transition state, relative to the ground state. Thus, the method can be used to select directly for low activation free energies. It is then straightforward to rank the sampled sequences based on their catalytic efficiency, the ratio between the rate constant for the catalytic step, k_{cat} and the Michaelis constant K_M.

Here, we report CPD calculations that aim to increase the binding of several ligands by MetRS. We first considered AnL. Three residues in the active site were allowed to mutate. The CPD method was tested for its ability to recover the known experimental variants [22]. The top six experimental variants were visited by the MC simulations and were highly ranked among the predicted sequences. We next considered the natural ligand methionyl adenylate (MetAMP). Another set of three residues near the ligand side chain were allowed to mutate. The wildtype sequence was highly ranked by the computational design. 17 other sequences among the top 40 predictions were tested experimentally and all found to be active. The computed binding free energy differences between variants were mostly in good agreement with the experimental values, obtained from kinetic measurements of the enzyme reaction. Next, we predicted MetRS variants that were specifically designed to bind the transition state for the enzymatic reaction Met+ATP → MetAMP+PP_i. The wildtype enzyme was highly ranked among 5832 possible variants, and for 20 variants that were characterized experimentally, the transition state binding free energies from the simulations were in good agreement with the values deduced from the experimental reaction rates. These calculations represent the first time an enzyme is specifically designed to optimize its transition state binding free energy relative to ground state binding, *i.e.*, its catalytic power. We expect the method will become the paradigm for computational design of enzymes.

Materials and methods

Theoretical approach: Designing for ligand binding

Stage 1: Adaptive apo simulation. We consider a polypeptide, with or without a bound ligand. Below, we will use a fixed backbone geometry, but the method is valid with a flexible backbone. Side chains can explore a few discrete conformations, or rotamers, and a few selected positions are allowed to mutate. In a first stage, we perform a MC exploration of the protein with no ligand, using the usual Metropolis-Hastings scheme [25–27]. We gradually increment a bias potential until all the side chain types at the mutating positions have roughly equal populations, thus flattening the free energy landscape. We number the mutating positions arbitrarily

1, . . ., p. The bias E^B at time t has the form:

$$E^B(s_1(t), s_2(t), \ldots, s_p(t); t) = \sum_i E_i^B(s_i(t); t) + \sum_{i<j} E_{ij}^B(s_i(t), s_j(t); t) \quad (1)$$

Here, $s_i(t)$ represents the side chain type at position i. The first sum is over single amino acid positions; the second is over pairs. The individual terms are updated at regular intervals of length T. At each update, whichever sequence variant ($s_1(t)$, $s_2(t)$, . . ., $s_p(t)$) is populated is penalized by adding an increment $e_i^B(s_i(t); t)$ or $e_{ij}^B(s_i(t), s_j(t); t)$ to each corresponding term in the bias. The increments have the form:

$$e_i^B(s_i(t); t) = e_0 \exp\left[-E_i^B(s_i(t); t)/E_0\right] \quad (2)$$

$$e_{ij}^B(s_i(t), s_j(t); t) = e_0 \exp\left[-E_{ij}^B(s_i(t), s_j(t); t)/E_0\right] \quad (3)$$

where e_0 and E_0 are constant energies. Thus, the increments decrease exponentially as the bias increases. This scheme is adapted from well-tempered metadynamics [24, 28, 29]. The individual bias terms depend on the system history, and can be written:

$$E_i^B(s; t) = \sum_{n; nT<t} e_i^B(s; nT)\delta_{s,s_i(nT)} \quad (4)$$

$$E_{ij}^B(s, s'; t) = \sum_{n; nT<t} e_{ij}^B(s, s'; nT)\delta_{s,s_i(nT)}\delta_{s',s_j(nT)} \quad (5)$$

where $\delta_{a,b}$ is the Kronecker delta. Over time, the bias for the most probable states grows until it pushes the system into other regions of sequence space. Two-position biases were implemented in the Proteus software [30, 31] during this work.

Stage 2: Biased holo simulation. In the second stage, the protein:ligand complex is simulated using the bias potential from stage 1. The sampled population of a sequence S is normalized to give a probability, denoted $\tilde{p}_H(S)$, where the subscript means "holo" and the tilde indicates that the bias is present. The apo state probability $\tilde{p}_A(S)$ was obtained in stage 1. Both probabilities can be converted into free energies $\tilde{G}$:

$$\begin{aligned} \tilde{p}_X(S) &= \frac{1}{Z_X} \exp\left(-\tilde{G}_X(S)/kT\right) \\ \tilde{G}_X(S) &= -kT \ln \tilde{p}_X(S) - kT \ln Z_X \end{aligned} \quad (6)$$

where $X = A$ or H and Z_X is a normalization factor that depends on X but not S. We also have a relation between the free energies with and without the bias:

$$\tilde{G}_X(S) = G_X(S) + E^B(S) \quad (7)$$

whose (straightforward) derivation is given in the Supporting Appendix (S1 File). Note that if the apo state flattening were ideal, $\tilde{p}_A(S)$ would be a constant, so that (from Eqs 6 and 7) $E^B(S) = -G_A(S)$, up to an additive constant. Thus, the ideal bias is the opposite of the apo free energy.

The binding free energy relative to a reference sequence R can be deduced from the populations. We have:

$$\begin{aligned} \Delta\Delta\tilde{G}(S) &= \left(\tilde{G}_H(S) - \tilde{G}_A(S)\right) - \left(\tilde{G}_H(R) - \tilde{G}_A(R)\right) \\ &= -kT \ln \frac{\tilde{p}_H(S)}{\tilde{p}_H(R)} + kT \ln \frac{\tilde{p}_A(S)}{\tilde{p}_A(R)} \\ &= \Delta\Delta G(S) \end{aligned} \tag{8}$$

Since the bias is the same in the bound and unbound states, it cancels out from $\Delta\Delta\tilde{G}(S)$, which is equal to the relative binding free energy *in the absence of bias*, $\Delta\Delta G(S)$. While the bias does not appear explicitly in (8), it is essential for accurate sampling. Perfect flattening, however, is not usually achieved, nor is it needed.

In the holo state, the probability of a sequence S (with bias) is:

$$\tilde{p}_H(S) \propto \exp\left(-\frac{\tilde{G}_H(S)}{kT}\right) = \exp\left(-\frac{G_H(S) + E^B(S)}{kT}\right) \approx \exp\left(-\frac{G_H(S) - G_A(S)}{kT}\right) \tag{9}$$

Thus, holo sampling follows a Boltzmann distribution governed by $G_H(s) + E^B(S)$, which is approximately the binding free energy $G_H(S) - G_A(S)$. This is exactly the quantity we want to design for. If the apo state is well-flattened, the biased holo simulation will be exponentially enriched in tight binders.

Energy function and matrix

The energy was computed using either an MMGBLK or an MMGBSA function ("molecular mechanics + Generalized Born + Lazaridis-Karplus" or "Surface Area"):

$$E = E_{\mathrm{MM}} + E_{\mathrm{GB}} + E_{\mathrm{LK|SA}} \tag{10}$$

The MM term used the Amber ff99SB force field [30, 32]. The SA term was described earlier [33–35]. The LK term and its parameterization were described earlier [35]. The GB term corresponds to a variant very similar to the one used in Amber, detailed in previous articles [33, 36, 37]. To make the calculation efficient, we compared two strategies. The first used a Native Environment Approximation (NEA), where the GB solvation radii for a given side chain were computed with the rest of the system in its native conformation [36, 38]. The second used a "Fluctuating Dielectric Boundary" (FDB) method, where the GB interaction between two residues I, J was expressed as a polynomial function of their solvation radii [39]. These were kept up to date over the course of the MC simulation, so the GB interaction could be deduced with little additional calculation [37, 39]. The solvent dielectric constant was 80; the protein one was 4.0 with the GBSA variants and 6.8 with GBLK [35]. Each solvent model is referred to by its GB variant and nonpolar term; for example, the FDBLK model combines FDB with LK.

To allow very fast MC simulations, we precomputed an energy matrix for each system [34, 40]. For each pair of residues I, J and all their allowed types and rotamers, we performed a short energy minimization (15 conjugate gradient steps) [30]. The backbone was fixed (in its crystal geometry) and the energy only included interactions between the two side chains and with the backbone. At the end of the minimization, we computed the interaction energy between the two side chains. Side chain–backbone interaction energies were computed similarly (and formed the matrix diagonal) [30].

Structural models

MetRS:AnL and MetRS:MetAMP complexes. For MetRS:AnL, we started from the crystal structure of a complex between a triple mutant of *E. coli* MetRS and AnL (PDB code 3H9B) [41]. The protein mutations were L13S, Y260L, H301L. We refer to this mutant as SLL. The protein backbone was held fixed. Side chains more than 20 Å from the ligand were held fixed. The other side chains were allowed to explore rotamers, taken from the Tuffery library, augmented to allow multiple orientations for certain hydrogen atoms [42, 43]. Side chains 13 and 301 were allowed to mutate into the following 14 types: ACDEHIKLMNQSTV; position 260 was allowed to mutate into the same types, except that Tyr replaced Asp. Thus, there were $14^3 = 2744$ possible sequences in all. Histidine protonation states at non-mutating positions were assigned by visual inspection of the 3D structure. System preparation was done using the protX module of the Proteus design software [31].

For MetRS:MetAMP, we started from a crystal complex (PDB code 1PG0) between *E. coli* MetRS and a methionyl adenylate (MetAMP) analogue [44]. The protein backbone was held fixed. Side chains more than 20 Å from the ligand were held fixed. The other side chains were allowed to explore rotamers [42, 43]. Side chains 13, 256 and 297 were allowed to mutate into all types except Gly or Pro, for a total of 5832 possible sequences in all. Histidine protonation states at non-mutating positions were assigned by visual inspection of the 3D structure.

Unfolded state. The unfolded state energy was estimated with a tri-peptide model [45]. For each mutating position, side chain type, and rotamer, we computed the interaction between the side chain and the tri-peptide it forms with the two adjacent backbone and C_β groups. Then, for each allowed type, we computed the energy of the best rotamer and averaged over mutating positions. The mean energy for each type was taken to be its contribution to the unfolded state energy. The contributions of the mutating positions were summed to give the total unfolded energy.

Ligand force field and rotamers

Force field. For the AnL azido group, we used atomic charges and van der Waals parameters obtained earlier for azidophenylalanine [46]. Parameters for the implicit solvent energy terms were assigned by analogy to existing groups. For methionyl adenylate (MetAMP), we mostly used existing Met and AMP parameters. For atoms close to the Met:AMP junction, we used atomic charges computed earlier for ThrAMP (G. Monard, personal communication) from *ab initio* quantum chemistry, in a manner consistent with the rest of the Amber force field [32]. Van der Waals parameters for atoms near the junction were assigned the same types as in Met or AMP. Parameters for bond lengths, angles and dihedrals involving junction atoms were taken from the experimental geometry of MetAMP. Stiffness parameters were assigned by analogy to existing parameters. The complete set of parameters for AnL and MetAMP is in Supplementary Material (S2 and S3 Files, respectively).

Rotamers. AnL was positioned in the protein complex so that its backbone had the position occupied in the MetRS:AnL crystal structure [41]. The ligand's side chain was allowed to explore rotamers. These were defined by the usual side chain rotamers of Met [42, 43]. We started by positioning Met in the pocket by superimposing it on AnL in the mutant MetRS:AnL crystal complex (PDB code 3H9B). We then positioned the 17 Met side chain rotamers from the Tuffery library. We extracted the AnL side chain from the experimental complex and superimposed it on each of the 17 Met rotamers, producing 17 AnL conformers. Finally, for each one, we performed a short energy minimization with the AnL backbone held fixed. The 17 minimized conformers defined the AnL rotamers. Notice that with this procedure, the azido group always had the same orientation relative to the aliphatic part of the AnL side

chain. For MetAMP, we allowed the Met rotamers from the Tuffery library, with the rest of the ligand held fixed. The ϕ and ψ dihedral angles around the MetAMP C_α were not allowed to rotate and the whole AMP moiety stayed fixed.

Modeling the MetRS transition state complex

MetRS catalyzes two reactions. In the first, Met reacts with ATP to give MetAMP and pyrophosphate. In the second, tRNA reacts with MetAMP. Here, we considered the first reaction, which occurs in the absence of tRNA. A model for the ground state ligands Met + ATP was first obtained, starting from the crystal complex between MetAMP and PP_i (PDB code 3KFL). The covalent structure was reset to that of Met + ATP and the geometry was adjusted by a short energy minimization. The complex included a magnesium ion. Next, a model for the activated ligand $[Met:ATP]^{\ddagger}$ was obtained, starting from the Met + ATP complex. First, a phosphate and carboxylate fragment were positioned in a geometry close to the expected pentacoordinate transition state arrangement [44, 47–49] and an *ab initio* energy minimization was done, including planarity constraints for the phosphorus and three oxygens. This led to a length of 2.4 Å for the P–O bonds perpendicular to the plane. Next, the molecular mechanics model was constructed. A covalent bond was introduced between the reacting Met carboxylate oxygen and the α phosphorus atom. The lengths for this bond and the symmetric one on the other side of the phosphorus were set to 2.4 Å. Planarity restraints were imposed on the phosphorus and the three α phosphate oxygens. A short energy minimization was done (with molecular mechanics). This led to an α phosphate geometry with three oxygens in plane and two perpendicular (Fig 1), as expected for in-line attack of the Met carboxylate on the

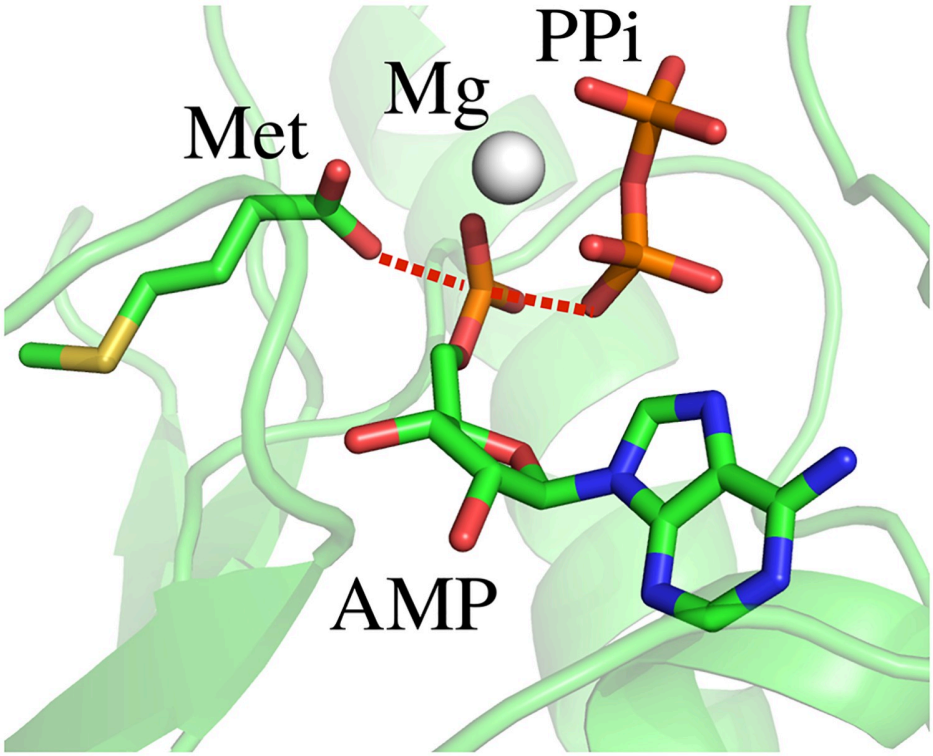


Fig 1. MetRS transition state for MetAMP formation. Closeup of the ligands.

https://doi.org/10.1371/journal.pcbi.1007600.g001

phosphate [44, 47–49]. *Ab initio* atomic charges were then computed for the entire activated ligand in this geometry, from a Merz-Kollman population analysis of the HF/6-31G* wavefunction [32], using Gaussian 9.0. The magnesium ion, which bridges the α, β and γ phosphates, was included in the calculation. The resulting charges were applied to atoms close to the α phosphate group, while other atoms kept their usual Met or ATP charges. Small manual adjustments were made to establish the correct total charge of -4. The final Mg charge was +1.5. Charges are in Supplementary Material (S3 File).

The geometry of the protein around the ligands was relaxed slightly by performing a short, restrained molecular dynamics simulation, with the ligands held fixed. The entire system was placed in a large box of explicit TIP3P water [50]. Harmonic restraints were applied to nonhydrogen atoms, with force constants that decreased gradually from 5 to 0.5 kcal/mol/Å^2 over 575 ps of dynamics, performed with the NAMD program [51]. The final protein geometry was used for the design calculations.

Monte Carlo simulations

To optimize the bias potential, we performed MC simulations of the apo state with bias updates every $T = 1000$ steps, with $e_0 = 0.2$ kcal/mol and $E_0 = 50$ kcal/mol [12]. During the first 10^8 MC steps, we optimized a bias potential including only single-position terms. There were $p = 3$ mutating positions, which all contributed to the bias. In the second stage, we ran MC or (in one case: MetAMP complex with the FDBSA solvent model) Replica Exchange MC (REMC) simulations of 5.10^8 MC steps [27], using 8 replicas with thermal energies (kcal/mol) of 0.17, 0.26, 0.39, 0.59, 0.88, 1.33, 2.0 and 3.0. Temperature swaps were attempted every 500 steps. All the replicas experienced the same bias potential. Both stages used 1- and 2-position moves.

Experimental mutagenesis and kinetic assays

Purification of wildtype and mutant MetRS. Throughout this study, we used a His-tagged M547 monomeric version of *E. coli* MetRS, fully active, both in vitro and in vivo [41]. The gene encoding M547 MetRS from pBSM547+ [52, 53] was subcloned into pET15blpa [54] to overproduce the His-tagged enzyme in *E. coli* ([55]). Site-directed mutations were generated using the QuickChange method [56], and the whole mutated genes verified by DNA sequencing. The enzyme and its variants were produced in BLR(DE3) *E. coli* cells. Transformed cells were grown overnight at 37˚C in 0.25 L of TBAI autoinducible medium containing 50 μg/ml ampicillin. They were harvested by centrifugation and resuspended in 20 ml of buffer A (10 mM Hepes-HCl pH 7.0, 3 mM 2-mercaptoethanol, 500 mM NaCl). They were disrupted by sonication (5 min, 0˚C), and debris was removed by centrifugation (15,300 G, 15 min). The supernatant was applied on a Talon affinity column (10 ml; Clontech) equilibrated in buffer A. The column was washed with buffer A plus 10 mM imidazole and eluted with 125 mM imidazole in buffer A. Fractions containing tagged MetRS were pooled and diluted ten-fold in 10 mM Hepes-HCl pH 7.0, 10 mM 2-mercaptoethanol (buffer B). These solutions were applied on an ion exchange column Q Hiload (16 mL, GE-Healthcare), equilibrated in buffer B containing 50 mM NaCl. The column was washed with buffer B and eluted with a linear gradient from 5 to 500 mM NaCl in buffer B (2 ml/min, 10 mM/min). Fractions containing tagged MetRS were pooled, dialyzed against a 10 mM Hepes-HCl buffer (pH 7.0) containing 55% glycerol, and stored at -20˚C. The homogeneity of the purified MetRS was estimated by SDS-PAGE to be higher than 95%.

Measurement of ATP-PPi exchange activity. Prior to activity measurements, MetRS was diluted in a 20 mM Tris-HCl buffer (pH 7.6) containing 0.2 mg/ml bovine serum albumin

(Aldrich) if the concentration after dilution was less than 1 μM. Initial rates of ATP-PPi exchange activity were measured at 25˚C as described [57]. In brief, the 100 μl reaction mixture contained Tris-HCl (20 mM, pH 7.6), MgCl2 (7 mM), ATP (2 mM), [^{32}P]PPi (1800-3700 Bq, 2 mM) and various concentrations (0-16 mM) of the Met amino acid. The exchange reaction was started by adding catalytic amounts of MetRS (20 μl). After quenching the reaction, ^{32}P-labeled ATP was adsorbed on charcoal, filtered, and measured by scintillation counting. kcat and K_M values were derived from iterative nonlinear fits of the theoretical equation to the experimental values using either MC-fit [58] or Origin (Origin Lab).

Results

Designing MetRS to bind azidonorleucine

As a first test, we searched for MetRS variants with strong azidonorleucine (AnL) binding. Positions 13, 260 and 301 were allowed to mutate, for comparison to the earlier experimental data [22]. 14 types were allowed at each position (see Methods), for a total of 2744 possible sequences. We compared three variants of the solvent model, which gave similar results. The first stage was to optimize a bias potential that flattened the free energy landscape in sequence space for apo MetRS. We used a bias potential including single-position terms only. After the adaptation period, we ran a further simulation of 10^8 MC steps to determine the biased populations. With the FDBLK solvent model, 2099 sequences were visited at least 1000 times, thanks to the adaptive bias. The second stage was to simulate the MetRS:AnL complex in the presence of the bias. 1957 sequences were visited at least 1000 times in both the first and second stages. For these, we used the sampled populations to deduce the AnL binding free energy (Eq 8), relative to the X-ray sequence. The overall computation time for system setup, energy matrix precalculation and both MC stages was about one day (per solvent model). Sequences sampled with and without the bias and ligand are shown in Fig 2 as logos.

Experimental directed evolution had revealed 21 active variants [22]. 13 of them were sampled by the computations and are listed in Table 1. 8 others either were not sampled or were predicted to have low stability. Each variant is referred to by the sequence of the three mutating positions; for example, the X-ray variant is SLL. The top six experimental sequences are the ones that were observed in multiple clones. The others were seen in just one clone [22]. The top six were all sampled by the computations and had good predicted stabilities and affinities (Table 1). SML was ranked the highest, 17th. The five others had lower ranks, between 45 and 104, but they were all within 1.4 kcal/mol of the top predicted variant (which was HMS). Other predicted variants may also be active, even though they were not revealed by the directed evolution experiments. For the SLL variant, the predicted rotamers for binding site residues were in good agreement with the X-ray structure (Supplementary Material; S1 File). The results in Table 1 were obtained with the FDBLK solvent treatment. The FDBSA solvent model gave similar results, while NEASA was slightly poorer (not shown), probably due to its simpler GB treatment [37].

We also searched for MetRS variants that maximized the AnL binding *specificity*, relative to Met. A bias potential was adaptively optimized for the MetRS:Met complex, then used in a simulation of the MetRS:AnL complex. The mutating positions and allowed types were the same as above. Specificity ranks are included in Table 1. Three of the top six experimental variants had high specificity ranks. The top experimental variant NLL was 36th, the next-best experimental variant SLL was 2nd, AQL was 18th, and CLL was 3rd. Thus, among the top 40 specificity ranks, there were 4 sequences that are known to be active. Evidently, selecting for specificity can help reveal active variants.

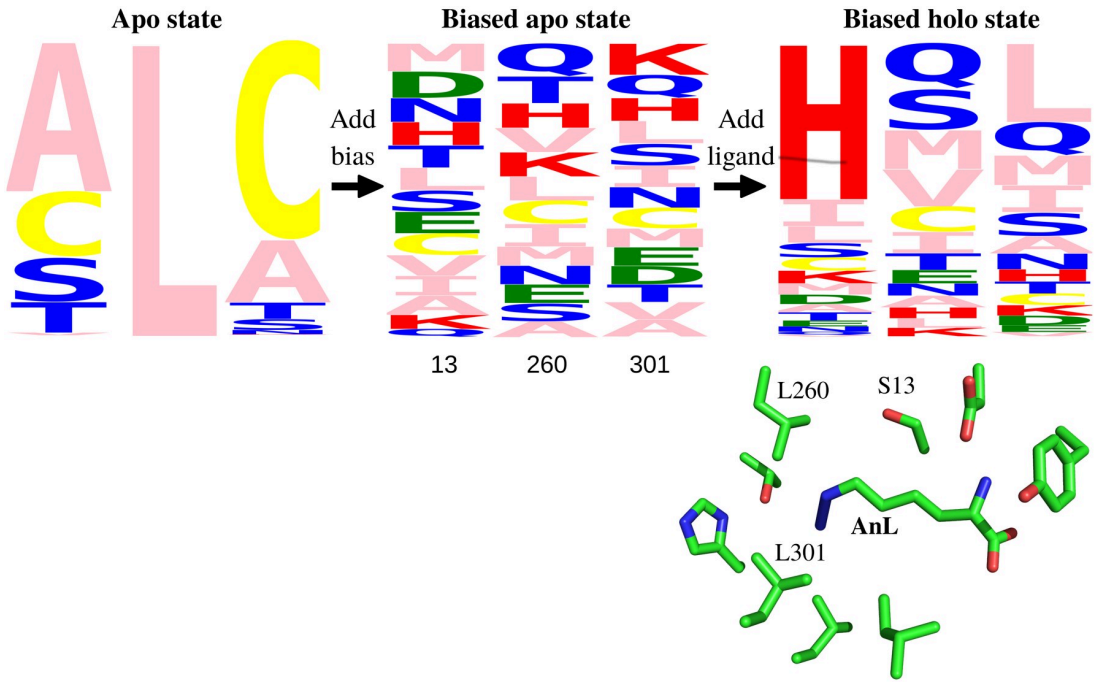


Fig 2. MetRS sequence logos. Sequences sampled without and with the AnL ligand (FDBLK solvent model) are shown in the form of logos, including the three mutating positions, 13, 260, 301. The logos represent the apo state (left), the biased apo state (middle), and the biased holo state (right). The height of each letter measures the frequency of its type. The 3D view below is a closeup of azidonorleucine (AnL) in the binding pocket, with selected side chains.

https://doi.org/10.1371/journal.pcbi.1007600.g002

Table 1. MetRS redesigned for AnL binding affinity or specificity.

a seq.	*b* pop.	*c* fold	*d* bind	*e* rank	*f* spec.	*g* rank	*a* seq.	*b* pop.	*c* fold	*d* bind	*f* rank	*e* spec.	*g* rank
NLL	62	6.7	0.3	104	5.7	36	CVL	1	6.9	-0.4	23	10.9	164
SLL	12	0.0	0.0	55	0.0	2	ACL	1	5.0	0.2	86	11.0	175
SML	4	4.6	-0.5	17	8.3	74	SCM	1	-0.9	0.4	123	18.8	589
AVL	3	6.7	-0.1	45	11.0	165	SLV	1	-2.3	1.4	688	7.4	57
AQL	2	4.2	0.0	57	3.3	18	SNL[h]	1	7.6	0.0	–	10.2	–
CLL	2	-0.6	0.1	73	1.0	3	SSL[h]	1	7.2	-0.1	–	10.3	–
							STL[h]	1	7.2	0.6	–	10.2	–

[a]Sequence at the designed positions 13, 260, 301, ranked by

[b]population among the experimental clones.

[c]Folding and

[d]binding free energies (kcal/mol) relative to the X-ray sequence SLL.

[e]Rank based on affinity or

[g]specificity.

[f]Specificity, defined by the binding free energy difference between AnL and Met (relative to SLL).

[h]Not ranked, since folding free energy is above the 7 kcal/mol threshold.

Calculations used the FDBLK solvent.

https://doi.org/10.1371/journal.pcbi.1007600.t001

Redesigning MetRS to bind MetAMP

As a second test, we searched for MetRS variants with a high affinity for the natural ligand methionyl adenylate (MetAMP). These should include the wildtype (WT) sequence and close homologs. Three positions close to the Met side chain (Fig 3), 13, 256, and 297 were allowed to mutate into all types except Gly or Pro, for a total of 5832 possible sequences. The first stage was to optimize a bias potential that flattened the free energy landscape in sequence space for apo MetRS. We performed calculations with both the FDBSA and the FDBLK variants of the solvent model, which gave similar results. We report the FDBSA results, since they were obtained first and were the basis for choosing which sequences to test experimentally. Selected FDBLK results are also reported. With the FDBSA solvent model, using Replica Exchange MC, 4178 variants were visited at least 1000 times.

The second stage was to simulate the MetRS:MetAMP complex in the presence of the bias potential. For sequences visited at least 1000 times in both stages (528 sequences), we used the sampled populations to deduce the MetAMP binding free energy (Eq 8), relative to the wild-type (WT) sequence LAI. The folding energy of each variant was also estimated (see Methods) and sequences less stable than WT by 5 kcal/mol or more were discarded. The top 20 remaining sequences, with the largest binding free energies, are shown in Table 2. The top sequence, CDV, had an Asp at position 256, positioned to form a salt bridge with the MetAMP ammonium group. Its binding free energy, relative to WT, was -1.4 kcal/mol. The next 19 variants had types similar to WT. Their computed binding free energies were close to WT, with relative values between -0.2 and 0.6 kcal/mol. The WT sequence was sixth overall. Among the top 40 variants, 17 mutants were produced experimentally. They were representative of the computational variants, while providing ease of construction (see Methods). CDV was left out, as the A256D mutation, selected for binding, might reduce the catalytic activity. All 17 tested variants had detectable activity, a 100% success rate for the design procedure. One other sequence, SAI, was tested experimentally and found to be active, but did not show up in the MC simulation. Thus the method produced one false negative along with 17 true positives.

Going further, we made a quantitative comparison between the computed and experimental binding free energies. The experimental dissociation constants were estimated from the Michaelis constants K_M. In the experimental conditions (excess ATP) and under the usual Michaelis-Menten assumptions [14, 59], K_M represents the dissociation constant for Met binding in the presence of bound ATP. Here, we computed relative binding free energies for binding MetAMP, not Met. Nevertheless, we expect that these MetAMP binding free energy changes can be compared to the experimental Met binding free energy changes; *i.e.*, we make the additional assumption that the relative effects of the mutations will be conserved going from MetAMP to Met+ATP.

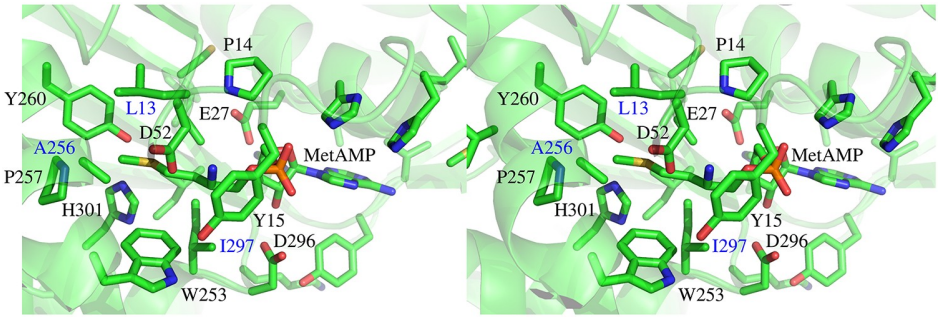


Fig 3. MetRS:MetAMP complex. Binding site closeup (stereo). Mutating side chains are 13, 256, 297.

https://doi.org/10.1371/journal.pcbi.1007600.g003

Table 2. MetRS redesigned for MetAMP binding by mutating positions 13, 256, 297.

			__binding__					___binding___	
rank	variant	[a]folding	[b]comp.	[b]exp.	rank	variant	[a]folding	[b]comp.	[b]exp.
1	CDV	4.5	-1.36		11	LAC	-0.3	0.25	1.8
2	MAV	1.3	-0.23	1.8	12	MAT	4.6	0.28	2.4
3	MAI	2.5	-0.20		13	LSV	0.4	0.29	
4	LAV	-1.3	-0.16	1.8	14	LAA	-0.6	0.31	3.8
5	MAC	2.3	-0.09	2.3	15	CAV	-8.8	0.34	2.8
6	**LAI**	0.0	0.00	0.0	16	CAI	-7.4	0.37	1.2
7	MAA	2.0	0.02		17	MSC	4.1	0.45	
8	MSV	3.1	0.11	3.4	18	MCV	1.0	0.46	
9	MSI	4.4	0.15	2.2	19	MCI	2.3	0.48	
10	LSI	1.6	0.20		20	MSA	3.8	0.56	
					21	LAT	1.8	0.59	2.2
26	CAC	-7.9	0.69	3.0	28	SAI	-3.5	0.72	1.2
51	SAC	-4.0	1.11	3.0	68	LAS	1.3	1.34	3.4
70	SSI	-1.9	1.35	2.2	81	SSC	-2.2	1.45	3.6
	MST	6.2	0.98	3.5		MSS	5.8	1.64	3.4

Calculations with the FDBSA solvent model.

[a]Folding and

[b]MetAMP binding free energies (kcal/mol) from computations and experiment, relative to the WT sequence LAI.

https://doi.org/10.1371/journal.pcbi.1007600.t002

Certain mutations at position 297 involved significant changes in the side chain volume, where the largest type, Ile or the smallest type, Ala was introduced or removed. For these, the computed binding free energies departed significantly from the experimental ones. However, if these two types were excluded, there were 25 point mutations between experimental variants, and for these, agreement was very good. The computed binding free energy differences had an rms error of just 0.52 kcal/mol and a mean unsigned error (mue) of 0.43 kcal/mol. The correlation between the experimental and computed sets was 0.52. Fig 4 shows the binding free energy changes. Note that the good agreement supports the assumption that the experimental K_M values are good proxies for the relative MetAMP binding free energies.

With the FDBLK solvent model, results were similar. The WT variant was ranked slightly lower, 20th. The top sequence was SAN, with a binding free energy of -1.3 kcal/mol relative to the WT. 7 of the 17 experimental sequences were ranked among the top 20 predictions. The computed and experimental binding free energy changes associated with point mutations are shown in Supplementary Material (Figure B in S1 File). Excluding (as above) mutations involving the types Ile or Ala at position 297, the mue and rms error were 0.76 and 0.98 kcal/mol, respectively, only slightly larger than with FDBSA.

Redesigning MetRS for catalytic power

For enzyme design, it is of great interest to select for a low activation free energy [14]. Therefore, we considered a model of the transition state complex (Fig 1). The ATP α phosphorus was bound to five oxygens: three coplanar and two perpendicular, corresponding to in-line attack of the Met carboxylate. In the first stage, we simulated a competing, ground state complex between MetRS, Met, and ATP. The same three binding pocket residues as above, 13, 256, and 297 were allowed to mutate into all types except Gly, Pro. We used the FDBLK solvent model. We optimized a bias potential during the MC simulation, flattening the free energy

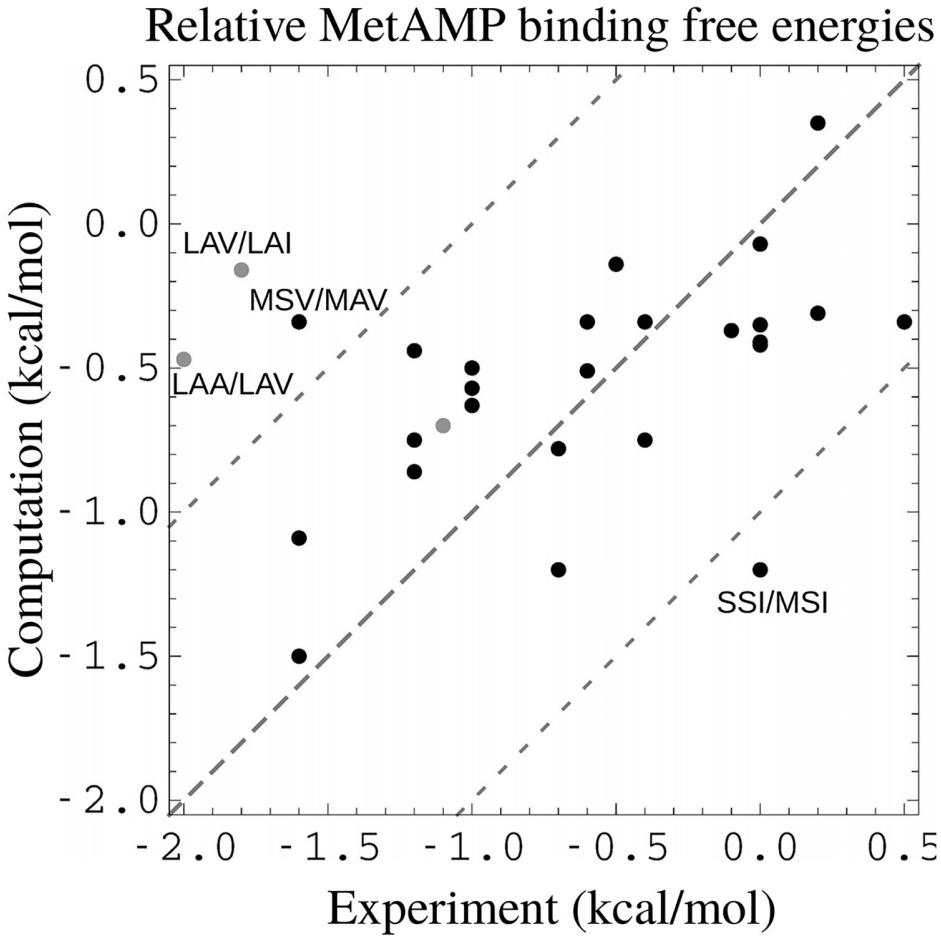


Fig 4. MetRS:MetAMP binding free energies, relative to the wildtype protein (WT). Shown are data for 28 point mutations. 3 gray points correspond to two mutations at position 297 (labeled) that change the side chain volume, plus one involving a variant (MST) that was predicted to be weakly stable (above our 5 kcal/mol threshold, see text) but was produced and measured experimentally nevertheless. Two other mutations with sizable errors are labeled.

https://doi.org/10.1371/journal.pcbi.1007600.g004

surface in sequence space. In the second stage, we simulated the transition state complex, with the bias included. All the variants that had been tested experimentally (Table 2) were sampled (WT and 19 variants, including five that Proteus had predicted (with FDBLK) to be above our 5 kcal/mol instability threshold). For each one, from the sampled populations, we deduced the free energy difference (Eq 8) between its ground state and transition state complexes, *i.e.*, its activation free energy. From transition state theory [14], this difference can be identified with the log of the catalytic reaction rate, k_{cat}. We also computed the Met dissociation free energies for the ground state complexes, which can be identified with the Met Michaelis constants, K_M. We first simulated the ground state complex with ATP but no Met, flattening its free energy surface with an adaptive bias. We then simulated the MetRS+Met+ATP complex, including the bias. From the sampled populations, we deduced the Met binding free energy of each variant, relative to WT (Eq 8). The overall protocol is schematized in Fig 5.

Fig 6 compares the k_{cat}/K_M ratios from experiment and simulations. We refer to them as catalytic efficiencies. We recall that they represent the 2nd order rate constant for the reaction of Met with the MetRS:ATP complex. Fig 6 shows the quantities $kT \log (k_{cat}/K_M) / (k_{cat}/K_M)_{WT}$,

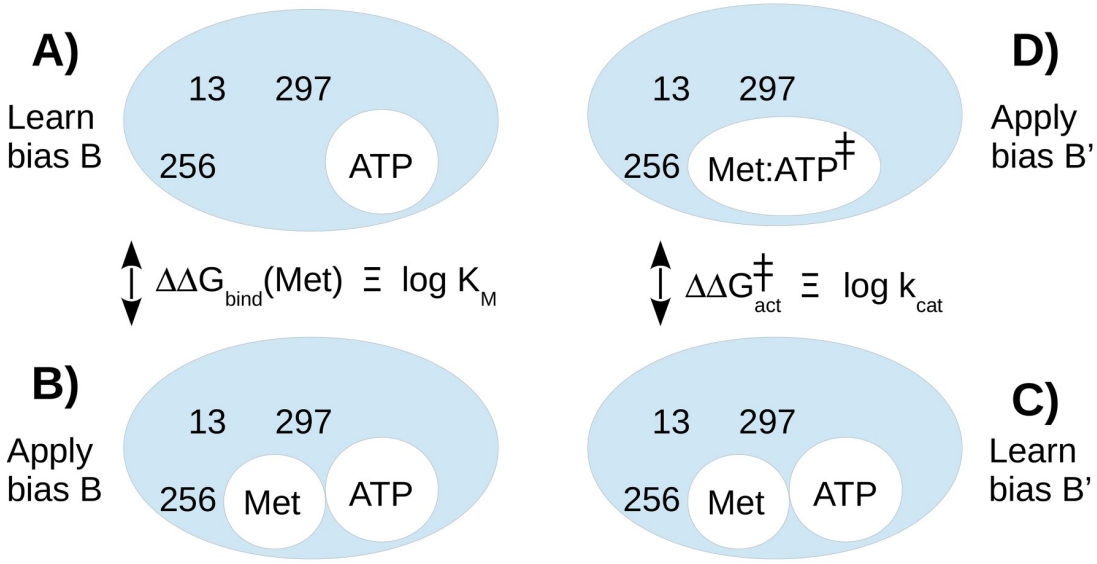


Fig 5. Computational scheme used to obtain the catalytic efficiencies k_{cat}/K_M. A) A bias B is optimized to flatten the sequence landscape of the enzyme without the Met ligand. Mutating positions are 13, 256, 297. **B)** The same bias B is used to simulate the complex including Met. Sequences are populated according to their Met binding affinities. **C)** A bias B' is optimized to flatten the sequence landscape of the complex including Met. **D)** B' is used to simulate the transition state complex. Sequences are populated according to their activation free energies. The lefthand simulations yield the predicted K_M values. The righthand simulations yield the predicted k_{cat} values.

https://doi.org/10.1371/journal.pcbi.1007600.g005

which express the catalytic efficiencies on a log scale, in thermal energy units, relative to the WT value. The figure includes WT and 19 other experimental variants. 5 of these had low predicted stabilities and are shown as gray points. WT defines the origin. For the other 14 points, agreement between calculations and experiment is quite good, with a correlation of 0.73 and mean errors of 1.36 kcal/mol (rms) and 1.18 kcal/mol (mue). Experimentally, WT has the largest efficiency. Computationally, two variants are predicted to be slightly better, by 0.9 and 0.2 kcal/mol, respectively, which is less than the mean error. Overall, by designing directly for a low activation free energy, we retrieve all the experimental variants and reproduce the catalytic efficiencies semi-quantitatively.

Discussion

Adaptive importance sampling solves the design problem for ligand binding and specificity. It applies positive design to one state (say, bound) and negative design to the other (unbound). It provides quantitative values for relative binding free energies or activation free energies. Variants sampled for one criterion, such as activation free energy (k_{cat}), can be reranked *a posteriori* based on another criterion, such as k_{cat}/K_M. *A posteriori* reranking or filtering does not leave out any important solutions; rather, the initial selection brings in too many solutions (*e.g.*, unstable variants), which are then filtered out at very little cost. In the first stage of the procedure, the sampling is very aggressive, if not exhaustive. In the second stage, it does not need to be exhaustive, since the best designs are exponentially enriched, and the unsampled variants are the ones with poor affinities or specificities. If one wants to reveal weak binders or perform reranking on another property, one can also flatten the energy landscape in the second stage. One can also use a more aggressive bias in one or both stages, including two-position biases. Replica Exchange MC can also be used to increase sampling. Using plain MC, one-position

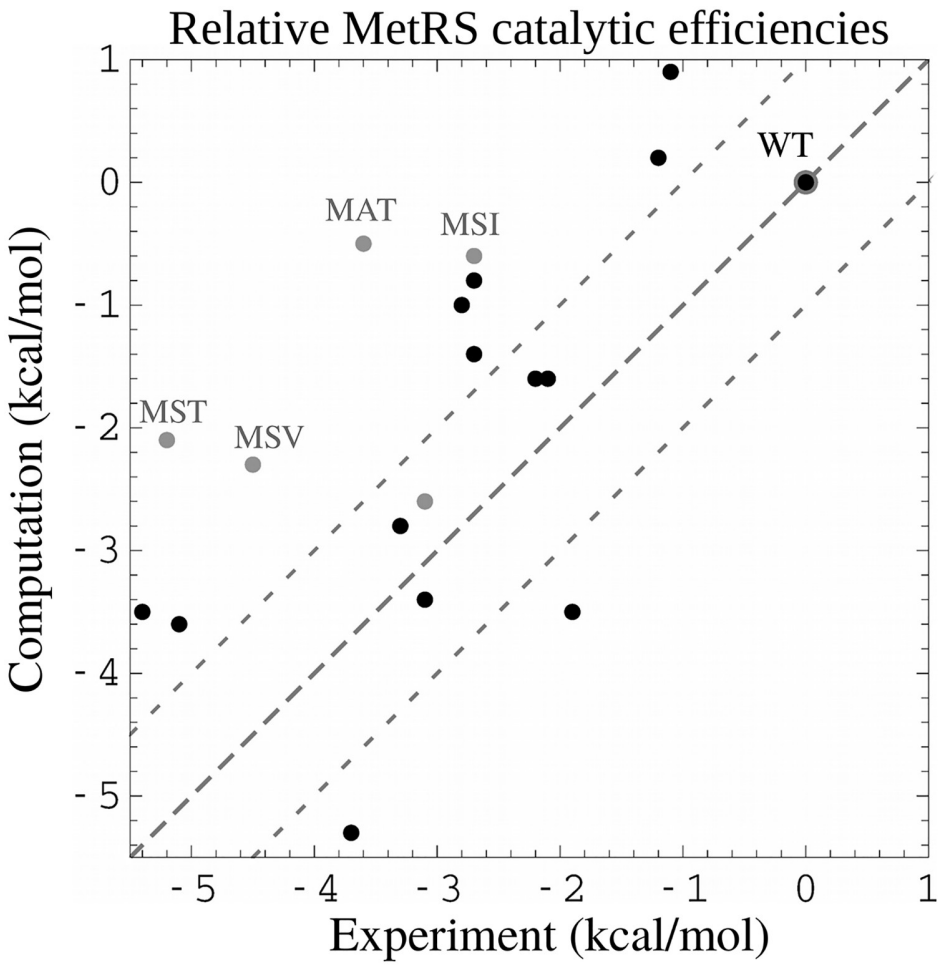


Fig 6. MetRS catalytic efficiencies kT log (k_{cat}/K_M) / $(k_{cat}/K_M)_{WT}$ relative to the wildtype (kcal/mol). Four gray points correspond to variants that were predicted to be weakly stable but were produced and measured experimentally nevertheless. Results obtained with the FDBLK solvent model.

https://doi.org/10.1371/journal.pcbi.1007600.g006

biases and no flattening of the holo state, our simulations produced 200 MetRS variants, enriched in tight binders, spanning a 7–8 kcal/mol range of binding free energies.

A difficulty when designing ligand binding is to choose one or more poses for the ligand. Here, we redesigned MetRS in cases where the ligand pose was known from an X-ray structure for one sequence: the SLL sequence in the AnL case and the wildtype sequence in the MetAMP case. For these ligands, we used the experimental ligand pose and protein backbone conformation. Three residues close to the ligand were then allowed to mutate. Not surprisingly, the calculations produced designed sequences that were homologous to the X-ray sequence. The experimental binding free energies in the MetAMP case were well-reproduced (the AnL values are not known). It is likely that other poses exist that would be compatible with other mutations, and would possibly lead to even stronger binding. The exploration of such alternate poses was left aside in this work. For the transition state complex, the position of the ligands could also be inferred with some confidence, since the enzyme achieves catalysis with little reorganization or motion of the substrates [15], and the modeled transition state geometry of the α phosphate was intermediate between that seen in two Met RS X-ray structures: the

MetRS product (adenylate) complex and the reactant (ATP) complex (PDB 4QRE). By designing the protein to stabilize this ligand pose, we may have biased the results towards native-like solutions. Here, too, the experimental relative activation free energies were well-reproduced, supporting the structural model.

Another important model component is the implicit solvent model. Here, we used a carefully-parameterized Generalized Born variant [33], a physically-plausible value of the protein dielectric constant and an "FDB" computational scheme that maintains the many-body nature of the GB model. The simpler, NEA scheme gave somewhat poorer results, similar to another recent study [35]. For nonpolar contributions to solvation, we compared a Surface Area (SA) treatment and a Lazaridis-Karplus (LK) treatment, which gave similar results. In the reported calculations, no water molecules were modelled explicitly. We also tested a model where three waters in the MetRS active site were explicitly represented: those that directly coordinate the Mg^{2+} ion in the substrate and transition state complexes for MetAMP formation. With both the FDBSA and FDBLK treatments, their explicit representation led to k_{cat} values well within the mean error of the calculations (relative to experiment). Most $kT \log (k_{cat}/K_M) / (k_{cat}/K_M)_{WT}$, values were with 0.2–0.3 kcal/mol of those reported above. Overall, the results were reasonably robust with respect to model details, with FDB giving improved performance.

Agreement with experiment was very good for three MetRS redesign test problems: redesign to bind the AnL ncAA, redesign to bind the natural intermediate MetAMP, and redesign for catalytic power for the reaction that produces MetAMP. Except for the earlier AnL data [22], the experiments were done in this work. Transition state modeling was done simply, by combining two X-ray structures and running a standard quantum chemistry protocol for atomic charges, consistent with the usual Amber force field [32]. All the procedures were carried out with the Proteus software, which is freely available to academics (https://proteus.polytechnique.fr). An entire calculation (setup, matrix calculation, MC simulations, postprocessing) lasted around one day on a 16-core desktop computer. We expect the present adaptive MC method will become the paradigm for computational enzyme design in the future.

Supporting information

S1 File. Supplementary appendix. This file includes a short theoretical derivation, some explanation of force field parameters, atomic charges for the MetRS transition state ligands, a figure showing the MetRS:AnL complex structure, and MetRS:MetAMP binding free energy results obtained with the FDBLK solvent model.
(PDF)

S2 File. Azidonorleucine force field information. This file contains the "topology" or 2D structure of AnL, including the atomic charges, followed by energy parameters for covalent bonds, angles, dihedrals, impropers, van der Waals terms and Generalized Born.
(FF)

S3 File. MetAMP force field information. This file contains the "topology" or 2D structure of MetAMP, including the atomic charges, followed by energy parameters for covalent bonds, angles, dihedrals, impropers, van der Waals terms and Generalized Born.
(FF)

Acknowledgments

We thank Christine Lazennec-Schurdevin for technical assistance, Alexandrine Daniel for preliminary MetRS:AnL calculations and Francesco Villa and David Mignon for many helpful discussions.

Author Contributions

Conceptualization: Emmanuelle Schmitt, Yves Mechulam, Thomas Simonson.

Data curation: Emmanuelle Schmitt, Yves Mechulam, Thomas Simonson.

Formal analysis: Vaitea Opuu, Thomas Simonson.

Investigation: Vaitea Opuu, Giuliano Nigro, Thomas Gaillard, Emmanuelle Schmitt, Yves Mechulam, Thomas Simonson.

Methodology: Vaitea Opuu, Emmanuelle Schmitt, Yves Mechulam, Thomas Simonson.

Project administration: Emmanuelle Schmitt, Yves Mechulam, Thomas Simonson.

Software: Vaitea Opuu, Thomas Simonson.

Supervision: Emmanuelle Schmitt, Yves Mechulam, Thomas Simonson.

Writing – original draft: Thomas Simonson.

Writing – review & editing: Emmanuelle Schmitt, Yves Mechulam, Thomas Simonson.

References

1. Malisi C, Schumann M, Toussaint NC, Kageyama J, Kohlbacher O, Höcker B. Binding Pocket Optimization by Computational Protein Design. PLoS One. 2012; 7:e52505. https://doi.org/10.1371/journal.pone.0052505 PMID: 23300688
2. Feldmeier K, Hoecker B. Computational protein design of ligand binding and catalysis. Curr Opin Chem Biol. 2013; 17:929–933. https://doi.org/10.1016/j.cbpa.2013.10.002 PMID: 24466576
3. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, et al. Computational design of ligand-binding proteins with high affinity and selectivity. Nature. 2013; 501:212–218. https://doi.org/10.1038/nature12443 PMID: 24005320
4. Stoddard B, editor. Methods in Molecular Biology: Design and Creation of Ligand Binding Proteins. Springer Verlag, New York; 2016.
5. Samish I, MacDermaid CM, Perez-Aguilar JM, Saven JG. Theoretical and computational protein design. Ann Rev Phys Chem. 2011; 62:129–149. https://doi.org/10.1146/annurev-physchem-032210-103509
6. Simonson T, Ye-Lehmann S, Palmai Z, Amara N, Bigan E, Wydau S, et al. Redesigning the stereospecificity of tyrosyl-tRNA synthetase. Proteins. 2016; 84:240–253. https://doi.org/10.1002/prot.24972 PMID: 26676967
7. Shen Q, Tian H, Tang D, Yao W, Gao X. Ligand-K* sequence elimination: a novel algorithm for ensemble-based redesign of receptor-ligand binding. Trans Comp Biol Bioinf. 2014; 11:573–578. https://doi.org/10.1109/TCBB.2014.2302795
8. Viricel C, Simoncini D, Allouche D, de Givry S, Barbe S, Schiex T. Approximate counting with deterministic guarantees for affinity computation. In: LeThi HA, Dinh TP, Nguyen NT, editors. Adv. Intell. Syst. Comput. vol. 360. Springer, New York; 2015. p. 165–176.
9. Hallen MA, Donald BR. COMETS (Constrained Optimization of Multistate Energies by Tree Search): A provable and efficient protein design algorithm to optimize binding affinity and specificity with respect to sequence. J Comp Biol. 2016; 23:311–321. https://doi.org/10.1089/cmb.2015.0188
10. Karimi M, Shen Y. iCFN: an efficient exact algorithm for multistate protein design. Bioinf. 2018; 34:i811–820. https://doi.org/10.1093/bioinformatics/bty564
11. Bhattacherjee A, Wallin S. Exploring protein-peptide binding specificity through computational peptide screening. PLoS Comp Biol. 2013; 7:e1003277. https://doi.org/10.1371/journal.pcbi.1003277
12. Villa F, Panel N, Chen X, Simonson T. Adaptive landscape flattening in amino acid sequence space for the computational design of protein:peptide binding. J Chem Phys. 2018; 149:072302. https://doi.org/10.1063/1.5022249 PMID: 30134674
13. Hayes RL, Armacost KA, Vilseck JZ, Brooks CL III. Adaptive landscape flattening accelerates sampling of alchemical space in multisite lambda dynamics. J Phys Chem B. 2017; 121:3626–3635. https://doi.org/10.1021/acs.jpcb.6b09656 PMID: 28112940
14. Jencks WP. Catalysis in chemistry and enzymology. Dover, New York; 1986.

15. Ibba M, Francklyn C, Cusack S, editors. Aminoacyl-tRNA Synthetases. Landes Bioscience, Georgetown; 2005.
16. Xie J, Schultz PG. A chemical toolkit for proteins: an expanded genetic code. Nat Rev Molec Cell Biol. 2006; 7:775–782. https://doi.org/10.1038/nrm2005
17. Young TS, Schultz PG. Beyond the canonical twenty amino acids: expanding the genetic lexicon. J Biol Chem. 2010; 285:11039–11044. https://doi.org/10.1074/jbc.R109.091306 PMID: 20147747
18. Liu CC, Schultz PG. Adding new chemistries to the genetic code. Ann Rev Biochem. 2010; 79:413–444. https://doi.org/10.1146/annurev.biochem.052308.105824 PMID: 20307192
19. Neumann-Staubitz P, Neumann H. The use of unnatural amino acids to study and engineer protein function. Curr Opin Struct Biol. 2016; 38:119–128. https://doi.org/10.1016/j.sbi.2016.06.006 PMID: 27318816
20. Chin JW. Expanding and reprogramming the genetic code. Nature. 2017; 550:53–60. https://doi.org/10.1038/nature24031 PMID: 28980641
21. Wang L, Brock A, Herberich B, Schultz PG. Expanding the genetic code of *Escherichia coli*. Science. 2001; 292:498–500. https://doi.org/10.1126/science.1060077 PMID: 11313494
22. Tanrikulu IC, Schmitt E, Mechulam Y, Goddard W III, Tirrell DA. Discovery of Escherichia coli methionyl-tRNA synthetase mutants for efficient labeling of proteins with azidonorleucine in vivo. Proc Natl Acad Sci USA. 2009; 106:15285–15290. https://doi.org/10.1073/pnas.0905735106 PMID: 19706454
23. Wang FG, Landau DP. Efficient, multiple-range random walk algorithm to calculate the density of states. Phys Rev Lett. 2001; 86:2050–2053. https://doi.org/10.1103/PhysRevLett.86.2050 PMID: 11289852
24. Laio A, Gervasio F. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Rep Prog Phys. 2008; 71:art. 126601. https://doi.org/10.1088/0034-4885/71/12/126601
25. Frenkel D, Smit B. Understanding molecular simulation, Chapter 3. Academic Press, New York; 1996.
26. Grimmett GR, Stirzaker DR. Probability and random processes. Oxford University Press, Oxford, United Kingdom; 2001.
27. Mignon D, Simonson T. Comparing three stochastic search algorithms for computational protein design: Monte Carlo, Replica Exchange Monte Carlo, and a multistart, steepest-descent heuristic. J Comput Chem. 2016; 37:1781–1793. https://doi.org/10.1002/jcc.24393 PMID: 27197555
28. Barducci A, Bussi G, Parrinello M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. Phys Rev Lett. 2008; 100:art. 020603. https://doi.org/10.1103/PhysRevLett.100.020603 PMID: 18232845
29. Dama JF, Parrinello M, Voth GA. Well-tempered metadynamics converges asymptotically. Phys Rev Lett. 2014; 112:art. 240602. https://doi.org/10.1103/PhysRevLett.112.240602 PMID: 24996077
30. Simonson T, Gaillard T, Mignon D, Schmidt am Busch M, Lopes A, Amara N, et al. Computational protein design: the Proteus software and selected applications. J Comput Chem. 2013; 34:2472–2484. https://doi.org/10.1002/jcc.23418 PMID: 24037756
31. Simonson T. The Proteus software for computational protein design. Ecole Polytechnique, Paris: https://proteus.polytechnique.fr; 2019.
32. Cornell W, Cieplak P, Bayly C, Gould I, Merz K, Ferguson D, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc. 1995; 117:5179–5197. https://doi.org/10.1021/ja00124a002
33. Lopes A, Aleksandrov A, Bathelt C, Archontis G, Simonson T. Computational sidechain placement and protein mutagenesis with implicit solvent models. Proteins. 2007; 67:853–867. https://doi.org/10.1002/prot.21379 PMID: 17348031
34. Gaillard T, Simonson T. Pairwise decomposition of an MMGBSA energy function for computational protein design. J Comput Chem. 2014; 35:1371–1387. https://doi.org/10.1002/jcc.23637 PMID: 24854675
35. Michael E, Polydorides S, Simonson T, Archontis G. Simple models for nonpolar solvation: parametrization and testing. J Comput Chem. 2017; 38:2509–2519. https://doi.org/10.1002/jcc.24910 PMID: 28786118
36. Polydorides S, Simonson T. Monte Carlo simulations of proteins at constant pH with generalized Born solvent, flexible sidechains, and an effective dielectric boundary. J Comput Chem. 2013; 34:2742–2756. https://doi.org/10.1002/jcc.23450 PMID: 24122878
37. Villa F, Mignon D, Polydorides S, Simonson T. Comparing pairwise-additive and many-body Generalized Born models for acid/base calculations and protein design. J Comput Chem. 2017; 38:2396–2410. https://doi.org/10.1002/jcc.24898 PMID: 28749575
38. Polydorides S, Amara N, Aubard C, Plateau P, Simonson T, Archontis G. Computational protein design with a generalized Born solvent model: application to Asparaginyl-tRNA synthetase. Proteins. 2011; 79:3448–3468. https://doi.org/10.1002/prot.23042 PMID: 21563215

39. Archontis G, Simonson T. Proton binding to proteins: a free energy component analysis using a dielectric continuum model. Biophys J. 2005; 88:3888–3904. https://doi.org/10.1529/biophysj.104.055996 PMID: 15821163

40. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science. 1997; 278:82–87. https://doi.org/10.1126/science.278.5335.82 PMID: 9311930

41. Schmitt E, Tanrikulu IC, Yoo TH, Panvert M, Tirrell DA, Mechulam Y. Switching from an Induced-Fit to a Lock-and-Key Mechanism in an Aminoacyl-tRNA Synthetase with Modified Specificity. J Mol Biol. 2009; 394:843–851. https://doi.org/10.1016/j.jmb.2009.10.016 PMID: 19837083

42. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn. 1991; 8:1267–1289. https://doi.org/10.1080/07391102.1991.10507882 PMID: 1892586

43. Gaillard T, Panel N, Simonson T. Protein sidechain conformation predictions with an MMGBSA energy function. Proteins. 2016; 84:803–819. https://doi.org/10.1002/prot.25030 PMID: 26948696

44. Crépin T, Schmitt E, Mechulam Y, Sampson PB, Vaughan MD, Honek JF, et al. Use of analogues of methionine and methionyl adenylate to sample conformational changes during catalysis in *Escherichia coli* methionyl-tRNA synthetase. J Mol Biol. 2003; 332:59–72. https://doi.org/10.1016/s0022-2836(03)00917-3 PMID: 12946347

45. Pokala N, Handel TM. Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J Mol Biol. 2005; 347:203–227. https://doi.org/10.1016/j.jmb.2004.12.019 PMID: 15733929

46. Druart K, Palmai Z, Omarjee E, Simonson T. Protein:ligand binding free energies: a stringent test for computational protein design. J Comput Chem. 2016; 37:404–415. https://doi.org/10.1002/jcc.24230 PMID: 26503829

47. Arnez JG, Augustine JG, Moras D, Francklyn CS. The first step of aminoacylation at the atomic level in histidyl-tRNA synthetase. Proc Natl Acad Sci USA. 1997; 94:7144–7149. https://doi.org/10.1073/pnas.94.14.7144 PMID: 9207058

48. Zurek J, Bowman A, Sokalski W, Mulholland A. MM and QM/MM modeling of threonyl-tRNA synthetase: Model testing and simulations. Struct Chem. 2004; 15:405–414. https://doi.org/10.1023/B:STUC.0000037896.80027.2c

49. Banik S, Nandi N. Aminoacylation Reaction in the Histidyl-tRNA Synthetase: Fidelity Mechanism of the Activation Step. J Phys Chem B. 2010; 114:12301–2311. https://doi.org/10.1021/jp910730s

50. Jorgensen WL, Chandrasekar J, Madura J, Impey R, Klein M. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983; 79:926–935. https://doi.org/10.1063/1.445869

51. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem. 2005; 26:1781–1802. https://doi.org/10.1002/jcc.20289 PMID: 16222654

52. Mellot P, Mechulam Y, Le Corre D, Blanquet S, Fayat G. Identification of an amino acid region supporting specific methionyl-tRNA synthetase:tRNA recognition. J Mol Biol. 1989; 208:429–443. https://doi.org/10.1016/0022-2836(89)90507-x PMID: 2477552

53. Schmitt E, Meinnel T, Panvert M, Mechulam Y, Blanquet S. Two acidic residues of *Escherichia coli* methionyl-tRNA synthetase act as negative discriminants towards the binding of noncognate tRNA anticodons. J Mol Biol. 1993; 233:615–628. https://doi.org/10.1006/jmbi.1993.1540 PMID: 8411169

54. Guillon L, Schmitt E, Blanquet S, Mechulam Y. Initiator tRNA binding by e/aIF5B, the eukaryotic/archaeal homologue of bacterial Initiation Factor IF2. Biochemistry. 2005; 44:15594–15601. https://doi.org/10.1021/bi051514j PMID: 16300409

55. Nigro G, Bourcier S, Lazennec-Schurdevin C, Schmitt E, Marlière P, Mechulam Y. Use of β3-methionine as an amino acid substrate of Escherichia coli methionyl-tRNA synthetase Journal of Structural Biology, in press, https://doi.org/10.1016/j.jsb.2019.107435

56. Braman J, Papworth C, A G. Site-directed mutagenesis using double-stranded plasmid DNA templates. Methods Molec Biol. 1996; 57:31–44.

57. Schmitt E, Meinnel T, Blanquet S, Mechulam Y. Methionyl-tRNA synthetase needs an intact and mobile KMSKS motif in catalysis of methionyl adenylate formation. J Mol Biol. 1994; 242:566–577. https://doi.org/10.1006/jmbi.1994.1601 PMID: 7932711

58. Dardel F. Comp App Biosci. 1994; 10:273–275.

59. Thompson D, Plateau P, Simonson T. Free energy simulations reveal long-range electrostatic interactions and substrate-assisted specificity in an aminoacyl-tRNA synthetase. ChemBioChem. 2006; 7:337–344. https://doi.org/10.1002/cbic.200500364 PMID: 16408313

Supplementary Appendix: Adaptive landscape flattening allows the design of both enzyme:substrate binding and catalytic power

Vaitea Opuu, Giuliano Nigro, Thomas Gaillard, Emmanuelle Schmitt, Yves Mechulam & Thomas Simonson*

Laboratoire de Biochimie, Ecole Polytechnique, Palaiseau, France

Relation between free energies with and without the bias potential

The free energy of a sequence S is denoted $G_X(S)$, where X indicates the apo or holo system. $G_X(S)$ is defined by a Boltzmann average over all possible conformations r:

$$e^{-\beta G_X(S)} = \int e^{-\beta E_X(S,r)}\, dr \tag{1}$$

where β is the inverse of the thermal energy kT. The bias potential $E^B(S)$ depends only on the sequence, not r. Therefore, for the free energy $\tilde{G}_X(S)$ in the presence of the bias, we have

$$e^{-\beta \tilde{G}_X(S)} = \int e^{-\beta(E_X(S,r)+E^B(S))}\, dr = e^{-\beta E^B(S)} \int e^{-\beta E_X(S,r)}\, dr = e^{-\beta E^B(S)} e^{-\beta G_X(S)} \tag{2}$$

which gives Eq. (7) in the main text.

Force field parameters for AnL and MetAMP

Force field information is given in the files AnL.ff and MetAMP.ff. Each file contains the "topology" or 2D structure of each molecule, including the atomic charges. This is followed by the energy parameters for covalent bonds, angles, dihedrals and impropers, van der Waals and Generalized Born terms. The data are in the format of the Proteus software, with comments for clarity. With minor reformatting, they can also be read by XPLOR, CHARMM and NAMD.

Atomic charges for the Met + ATP → MetAMP + PP_i transition state

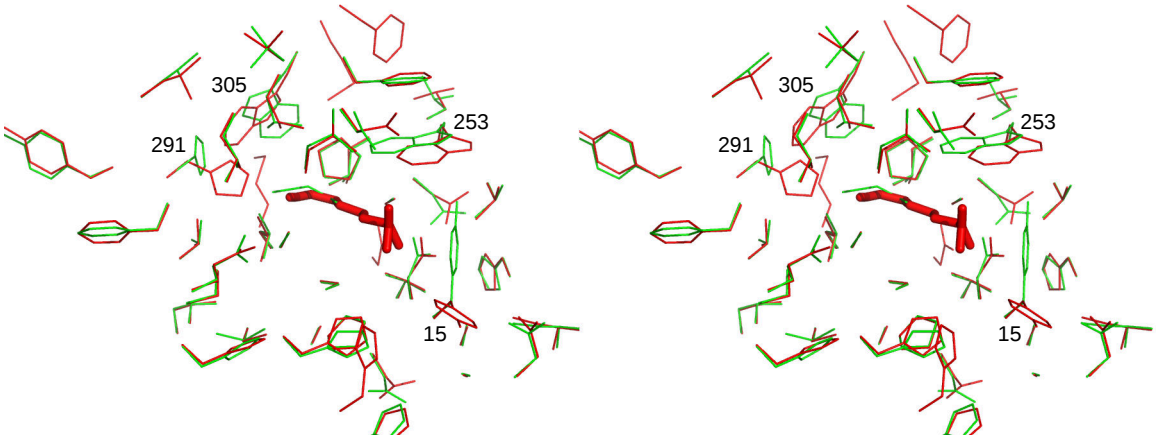
The transition state charges are given below, in the form of a Proteus topology file:

```
ATOM MG   TYPE=MG  CHARGE= 1.5000    ATOM O2A  TYPE=O2  CHARGE=-0.7016
                                     ATOM O3A  TYPE=OA  CHARGE=-0.8680
ATOM N    TYPE=N3  CHARGE=-0.3025    ATOM O5'  TYPE=OS  CHARGE=-0.4478
ATOM HN1  TYPE=H   CHARGE= 0.2770    ATOM C5'  TYPE=CT  CHARGE= 0.0558
ATOM HN2  TYPE=H   CHARGE= 0.2770    ATOM H5'1 TYPE=H1  CHARGE= 0.0679
```

```
ATOM HN3  TYPE=H   CHARGE= 0.2770    ATOM H5'2 TYPE=H1  CHARGE= 0.0679
ATOM CA   TYPE=CT  CHARGE= 0.0204    ATOM C4'  TYPE=CT  CHARGE= 0.1065
ATOM HA   TYPE=HP  CHARGE= 0.0741    ATOM H4'  TYPE=H1  CHARGE= 0.1174
ATOM CB   TYPE=CT  CHARGE= 0.0297    ATOM O4'  TYPE=OS  CHARGE=-0.3548
ATOM HB2  TYPE=HC  CHARGE= 0.0195    ATOM C1'  TYPE=CT  CHARGE= 0.0394
ATOM HB3  TYPE=HC  CHARGE= 0.0195    ATOM H1'  TYPE=H2  CHARGE= 0.2007
ATOM CG   TYPE=CT  CHARGE=-0.0027    ATOM N9   TYPE=N*  CHARGE=-0.0251
ATOM HG2  TYPE=H1  CHARGE= 0.0394    ATOM C8   TYPE=CK  CHARGE= 0.2006
ATOM HG3  TYPE=H1  CHARGE= 0.0394    ATOM H8   TYPE=H5  CHARGE= 0.1553
ATOM SD   TYPE=S   CHARGE=-0.2782    ATOM N7   TYPE=NB  CHARGE=-0.6073
ATOM CE   TYPE=CT  CHARGE=-0.0580    ATOM C5   TYPE=CB  CHARGE= 0.0515
ATOM HE1  TYPE=H1  CHARGE= 0.0638    ATOM C6   TYPE=CA  CHARGE= 0.7009
ATOM HE2  TYPE=H1  CHARGE= 0.0638    ATOM N6   TYPE=N2  CHARGE=-0.9019
ATOM HE3  TYPE=H1  CHARGE= 0.0638    ATOM HN61 TYPE=H   CHARGE= 0.4115
ATOM C    TYPE=C   CHARGE= 0.9610    ATOM HN62 TYPE=H   CHARGE= 0.4115
ATOM O    TYPE=O   CHARGE=-0.7856    ATOM N1   TYPE=NC  CHARGE=-0.7615
ATOM OXP  TYPE=OA  CHARGE=-0.7517    ATOM C2   TYPE=CQ  CHARGE= 0.5875
                                     ATOM H2   TYPE=H5  CHARGE= 0.0473
ATOM PG   TYPE=P   CHARGE= 1.4463    ATOM N3   TYPE=NC  CHARGE=-0.6997
ATOM O1G  TYPE=O3  CHARGE=-1.0141    ATOM C4   TYPE=CB  CHARGE= 0.3053
ATOM O2G  TYPE=O3  CHARGE=-0.9438    ATOM C3'  TYPE=CT  CHARGE= 0.2022
ATOM O3G  TYPE=O3  CHARGE=-0.9153    ATOM H3'  TYPE=H1  CHARGE= 0.0615
ATOM PB   TYPE=P   CHARGE= 1.5390    ATOM C2'  TYPE=CT  CHARGE= 0.0670
ATOM O1B  TYPE=O2  CHARGE=-0.9582    ATOM H2'  TYPE=H1  CHARGE= 0.0972
ATOM O2B  TYPE=O2  CHARGE=-0.8900    ATOM O2'  TYPE=OH  CHARGE=-0.6139
ATOM O3B  TYPE=OS  CHARGE=-0.6252    ATOM HO2' TYPE=HO  CHARGE= 0.4186
ATOM PA   TYPE=P5  CHARGE= 1.2530    ATOM O3'  TYPE=OH  CHARGE=-0.6541
ATOM O1A  TYPE=O2  CHARGE=-0.6138    ATOM HO3' TYPE=HO  CHARGE= 0.4376
```

Structure of the SLL:AnL complex from Proteus and experiment

Figure A: Complex between AnL and the SLL MetRS mutant. Red: lowest-energy Proteus structure; green: X-ray. Side chains close to the ligand; the 4 largest deviations are labeled.



MetRS:MetAMP binding free energies with the FDBLK solvent model

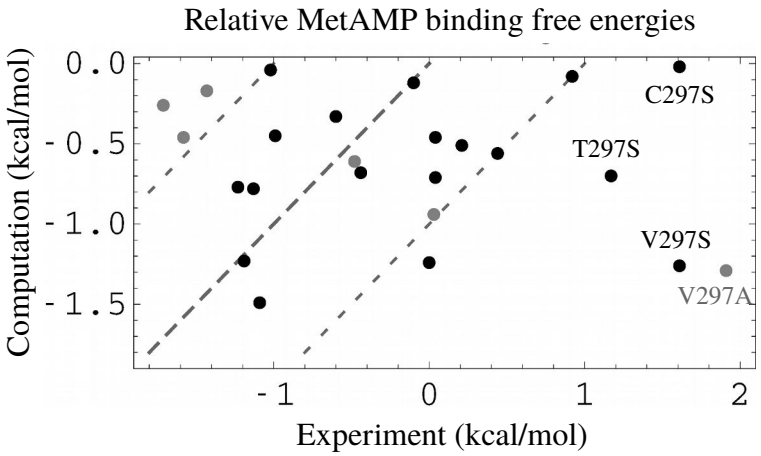


Figure B: MetRS:MetAMP binding free energies relative to WT, with the FDBLK solvent. Data are for 27 point mutations, as in the main text (Fig. 4). The largest errors are labeled.

3.1 The effect of native rotamers

We tested the effect of native rotamers on the MetRS complex with the transition state [α-Met:ATP]‡. We repeated the earlier estimations of catalytic efficiencies for α-Met, using MetRS with its backbone relaxed in the context of the wild type sequence. Positions 13, 256, and 297 were allowed to vary. Computed values are then compared to the values obtained previously without native rotamers.

3.1.1 Results

We considered two states: MetRS:ATP and MetRS:[α-Met:ATP]‡. We flattened the sequence space of both states with simulations of 10^8 steps. Then, we performed two biased simulations of the same length and computed the catalytic efficiencies. These simulations were produced with the FDBLK solvent model.

Figure 3.1 compares the catalytic efficiencies obtained with and without the native rotamers. The estimates are less accurate with the native rotamers when compared to experiments since the sampling is now biased toward the native-like variants. The L13M mutation present in {MAC, MAV, MAT, MSV, MST, MSA} showed a high loss of accuracy. SAI and CAI are no longer predicted as better variants than the wild type LAI. Indeed, figure 3.2 shows that the wild type sequence is the most active variant in the new simulations.

3.1.2 Conclusions

The fixed backbone approximation introduces a bias in the side chain packing. Combined with a small number of rotamers for some side chain types, it is sometimes impossible to have a satisfying packing, even for the wild type sequence. Thus, Phe has only 3 conformations in the library used and is one of the most difficult side chains to place. Therefore, it is necessary to introduce the native rotamers so the wild type conformation can be correctly reproduced. This allowed us to eliminate false positives. However, the estimates are now less accurate when compared to experiments. Variants with the L13M mutation are underestimated. This variant involves a small change of the backbone geometry (Y. Mechulam, E. Schmitt, and G. Nigro, personal communication). Therefore, this mutation is not fully consistent with the fixed backbone CPD. However, we expect that the loss of accuracy in favor of better discrimination

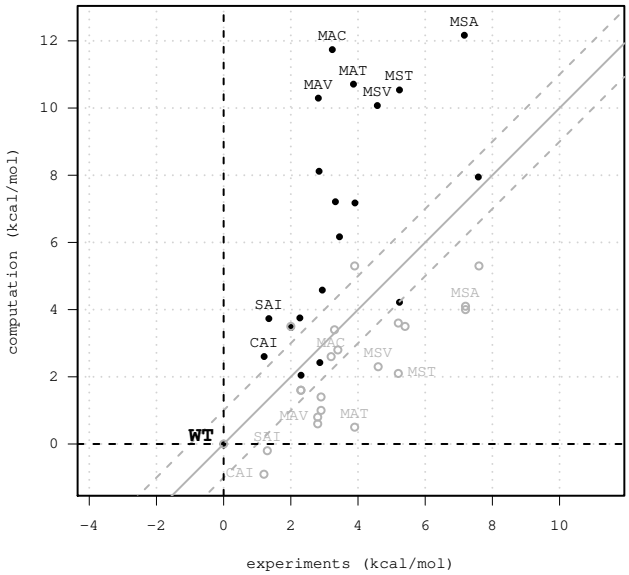


Figure 3.1: **Comparison between experiments/predictions of catalytic efficiency for MetRS:[α-Met:ATP]‡** with and without native rotamers. Catalytic efficiencies with (respectively withtout) native rotamers are represented by black (gray) dots.

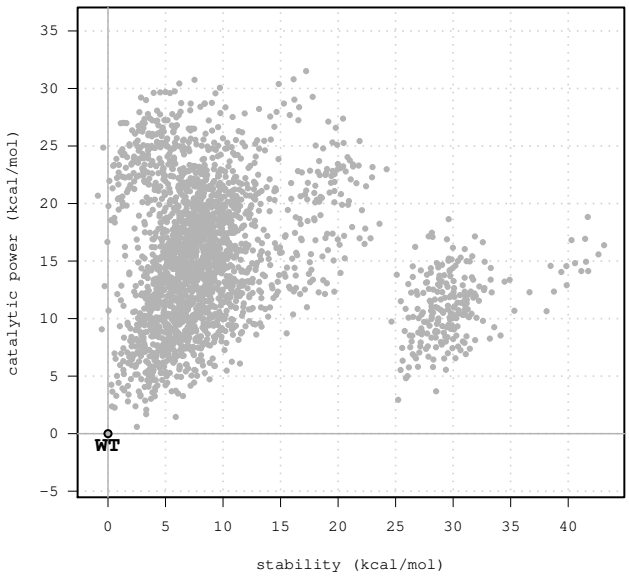


Figure 3.2: **Distribution of all the variants at positions 13, 256, and 297, sampled according to catalytic efficiencies and stabilities.** Values are computed relative to the wild type sequence LAI.

of true positives will be beneficial overall. The use of a richer rotamer library may improve the accuracy ([Shapovalov and Dunbrack, 2011]).

Chapter 4

Engineering methionyl-tRNA synthetase for β amino acid activity: background and methods

The incorporation of β amino acids into proteins is an important biotechnological challenge. Each canonical amino acid (*i.e.* α) has two β homologs (figure 4.1) which have an additional carbon atom between the carboxylate and amine groups. Incorporating such molecules into proteins could enhance available backbone geometries. One standard approach is to engineer an appropriate aminoacyl-tRNA synthetase (aaRS).

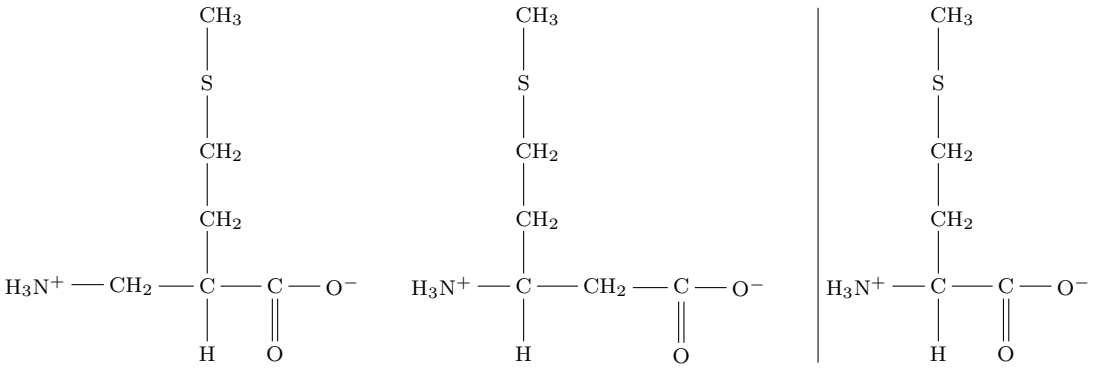


Figure 4.1: **β-Methionines (left) and α-Methionine (right) examples**

Here, we present the engineering of Methionyl-tRNA synthetase (MetRS), which allows the aminoacylation of tRNA(Met), for the incorporation of two β amino acids. Experimental directed evolution has been used to redesign aaRSs to accept unnatural amino acids (uaa)

([Tanrikulu et al., 2009]). However, these approaches have limitations. Computational design is another possibility. In a previous project ([Opuu et al., 2020a]), we designed active MetRS variants for Met and recovered known variants for azidonorleucine (ANL), a Met homolog. We used a Monte Carlo algorithm that performs an adaptive sampling with a few positions allowed to mutate. Then, we considered the transition state of the aminoacylation reaction, to select variants by their catalytic power. Now, we apply this method to the redesign of MetRS for β-Met and β-Val activity. Wild type MetRS can process β-Met but with a weak activity. Hence, we searched for variants with better activity than the wild type variant.

First, we recall the relationship between biochemical constants (k_{cat}, K_M et $\frac{k_{cat}}{K_M}$) and free energy. Next, we detail the MC approach for free energy estimations. Then, we briefly present the activation reaction associated with the transition state we modeled and some structural properties of MetRS and β amino acids.

Next, we present a first calculation strategy we used for the search of active variants for β-Met and β-Val. First, we considered the MetRS complexes with β-MetAMP and β-ValAMP, the products of the reaction. We sampled MetRS variants according to ligand binding, with three positions allowed to mutate {13, 256, 297}. To investigate further these three positions, we then considered transition state binding for both β amino acids.

Finally, we introduce a new method to pick positions to mutate according to binding free energy. We will apply this method to the MetRS complex with the [β-Met:ATP]‡ or [β-Val:ATP]‡ transition states. It allows to select quartets of positions to study in detail. With these positions varying, we then produce mutations for β-Met or β-Val activity.

4.1 Enzyme kinetics and standard free energy

To sample enzyme variants, we use the binding to enzyme substrates, reaction products, and transition states. Now, we recall the Mechaelis-Menten kinetic model that we use to describe binding affinity and catalytic power.

4.1.1 Protein ligand binding

Portein ligand binding and its specificity is due to the side chain composition and backbone geometry of the binding site. Let E be an enzyme that binds to a ligand S:

$$E + S \rightleftharpoons ES. \tag{4.1}$$

The binding affinity of E for S is measured by the ratio of equilibrium concentrations K_a:

$$K_a = \frac{[ES]}{[E][S]} = \frac{1}{K_d} \tag{4.2}$$

[E], [S], and [ES] are the equilibrium concentrations of enzyme E, substrat S, and ES complex. K_d is the dissociation constant. We derive a binding free energy from K_a, the association constant:

$$\Delta G_b = -kT \times ln(K_a^*) \tag{4.3}$$

k (kcal.K^{-1}.mol^{-1}) is the Boltzmann constant, T (K) is the temperature, and ΔG_b (kcal/mol) is the standard binding free energy. K_a^* is the association constant:

$$K_a^* = \frac{\frac{[ES]}{C^0}}{\frac{[E][S]}{C^0 C^0}} \tag{4.4}$$

Where $C^0 = 1$ M. Free energy is computed at standard concentrations although C^0 may not be explicitly written ([General, 2010]).

To improve the binding of a given ligand L, the quantity we are interested in is the relative free energy. Let E be an enzyme and E' a variant of E, the relative binding free energy is defined as follows:

$$\Delta\Delta G_b(E \rightarrow E') = -kT \times ln(\frac{K'_a}{K_a}) \tag{4.5}$$

K'_a is the binding constant of E' for L. Here, the dependence on C^0 is canceled.

4.1.2 Michaelis-Menten model

The activation reaction can be modeled with the Michaelis-Menten approach. First, we recall the model, then we detail the principles of catalytic power.

4.1.2.1 The model

We consider the non-covalent binding between enzyme E and substrat S, then the transformation into a product P:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_{cat}}{\rightharpoonup} E + P \tag{4.6}$$

k_1 ($M^{-1}s^{-1}$) a second order reaction rate constant for complex formation. k_{-1} (s^{-1}) is a dissociation rate constant. k_{cat} (s^{-1}) measures the rate of product formation. For equilibrium concentrations, the association constant is given by: $K_a = k_1/k_{-1}$. The enzyme concentration is conserved, $[E]_0 = [E] + [ES]$.

Elementary reactions for the evolution of each species lead to the following system of differential equations:

$$\begin{aligned} d[ES]/dt &= k_1[E][S] - (k_{-1} + k_{cat})[ES] \\ d[E]/dt &= (k_{-1} + k_{cat})[ES] - k_1[E][S] \\ d[S]/dt &= k_{-1}[ES] - k_1[E][S] \\ d[P]/dt &= k_{cat}[ES] \end{aligned} \tag{4.7}$$

To express the product formation, we assume that the system is in a quasi-stationary state ([Briggs and Haldane, 1925]) such that the bound enzyme concentration is time-independent. Now we have:

$$\frac{[E][S]}{[ES]} = \frac{k_{-1} + k_{cat}}{k_1} = K_M \tag{4.8}$$

We recognize the Michaelis constant, K_M (M)

Since the enzyme concentration is conserved, one can substitute $[E]$ by $[E]_0 - [ES]$ in 4.8. In addition, we assume that we are in an early stage of the process, so $[S] \equiv [S]_0$. We obtain the bound enzyme concentration as a function of initial concentrations:

$$[ES] = \frac{[E]_0[S]_0}{[S]_0 + K_M} \tag{4.9}$$

Now, we can substitute $[ES]$ in the last equation of system 4.7 to obtain:

$$d[P]/dt = k_{cat}[ES] = k_{cat}\frac{[E]_0[S]_0}{[S]_0 + K_M} = \frac{V_{max}[S]_0}{[S]_0 + K_M} \tag{4.10}$$

This gives:

$$d[P]/dt = \frac{k_{cat}}{K_M}[E][S]_0 \tag{4.11}$$

As established by Mechaelis-Menten (figure 4.2), enzyme kinetics is described by the initial concentrations $[S]_0$ and $[E]_0$, and two constants k_{cat} and K_M. K_M can approximate K_d if $k_{cat} \ll k_{-1}$.

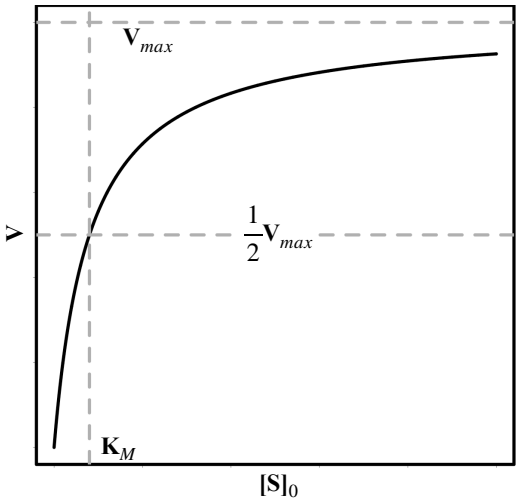


Figure 4.2: **Michaelis-Menten kinetic model for the initial rate.**

4.1.2.2 Binding to the transition state according to Michaelis Menten

In the Michaelis-Menten model, catalytic power is measured by the product formation rate. A high power corresponds to strong transition state binding. The catalytic efficiency is defined as the second order reaction rate $\frac{k_{cat}}{K_M}$ ($M^{-1}s^{-1}$) that measure the effectiveness of enzymes. It measures the transition state binding. Indeed, Figure 4.3 shows the sequence of states for a given system along the reaction coordinate for the non-catalyzed path (in red) and the catalyzed one. First, a complexe ES is formed with the binding free energy ΔG_b, then the substrat is activated with the activation free energy ΔG_a, and the product is released. The limiting step is the activation for the formation of the transition state denoted $S^{\ddagger}$. Natural enzymes stabilize the transition state such that the activation free energy is lower ($\Delta G_a < \Delta G_a^*$).

The catalytic power can be decomposed into two energetic contributions. First, the binding free energy:

$$\frac{1}{K_M} \approx K_a = exp(-\frac{\Delta G_b}{kT}) \tag{4.12}$$

Then, thanks to the transition state theory ([Jencks, 1987, Garcia-Viloca, 2004, Marti et al., 2004]), we have the activation free energy:

$$k_{cat} \propto exp(-\frac{\Delta G_a}{kT}) \tag{4.13}$$

The catalytic power can be expressed as follow:

$$\begin{aligned} \frac{k_{cat}}{K_M} &\propto exp(-\frac{\Delta G^{\ddagger}}{kT}) \\ \Delta G^{\ddagger} &= \Delta G_a + \Delta G_b \end{aligned} \tag{4.14}$$

Therefore, the relative catalytic power between E and one of its variants E' is:

$$\begin{aligned} (\frac{k_{cat}}{K_M})'/(\frac{k_{cat}}{K_M}) &= exp(-\frac{\Delta\Delta G^{\ddagger}(E \to E')}{kT}) \\ -kT \times ln((\frac{k_{cat}}{K_M})'/(\frac{k_{cat}}{K_M})) &= \Delta\Delta G^{\ddagger}(E \to E') \end{aligned} \tag{4.15}$$

$(\frac{k_{cat}}{K_M})'$ is the relative catalytic power. $\Delta\Delta G^{\ddagger}(E \to E')$ is the change of binding free energy for the transition state $S^{\ddagger}$.

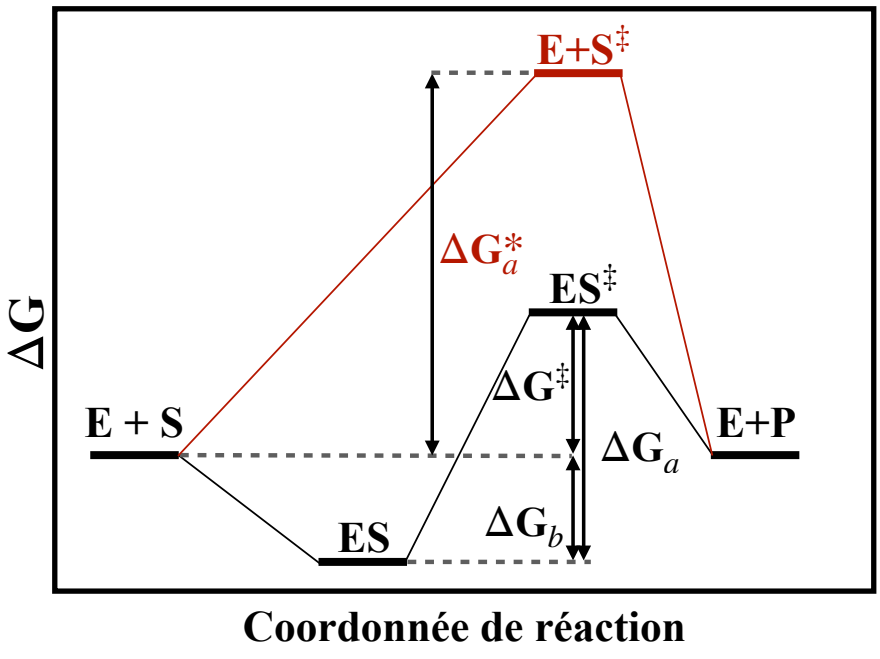


Figure 4.3: **Free energy profil along the reaction coordiante** The non-catalyzed path is showed in red with an acitivation energy denoted ΔG_a^*.

The binding process is accompanied by a lost of entropy. The overall binding free energy is $\Delta G_b = \Delta H - T\Delta S$ where ΔH is the enthalpy and ΔS the entropy. This is due to the loss of geometrical degrees of freedom, compensated by molecular interactions in the binding site. For $ES \rightarrow ES^{\ddagger}$, the enzyme binding site is pre-organized, so the stabilization of the transition state is mainly due to the enthalpic contribution. Many studies showed that the electrostatic contribution is predominant ([Warshel, 1978, Jindal et al., 2017]).

4.2 Biological context

4.2.1 Methionine aminoacylation reaction

aaRSs binds a specific canonical amino acid to its cognate tRNA. This reaction is called aminoacylation. MetRS catalyses the aminoacylation of Met in two steps (4.4). First, α-Met-adenylate (MetAMP) is produced and pyrophosphate (PPi) is released. Then, Met is transferred to its cognate tRNA from the MetAMP molecule.

We consider the part of the reaction, in the absence of tRNA. MetRS binds a molecule of adenosine triphosphate (ATP) and one amino acid. Complex formation is accompanied by change of conformation of the activation, or KMSKS loop (figure 4.6). The conformation associated to the complex is called active. The active conformation has its KMKSK motif positioned to stabilize the tri-phosphate fragment ([Schmitt et al., 1994]).

Once the complex MetRS:Met:ATP is formed, a nucleophilic attack occurs on the ATP α phosphate (in the presence of a Mg^{2+} ion). It leads to transition state [Met:ATP]‡ formation where P-O bonds form a plane and two are perpendicular to the plane (figure 4.5) ([Leatherbarrow et al., 1985]). The KMSKS loop stabilizes the adenine fragment in the reverse conformation ([Denessiouk and Johnson, 2003]). Next, PPi is released and MetAMP is produced.

$$\begin{array}{ccccc}
\text{MetRS+ATP+Met} & \underset{K_{ATP}}{\rightleftharpoons} & \text{MetRS:ATP+Met} & & \text{MetRS:MetAMP+PPi} \\
K_{Met} \Big\| & & K'_{Met} \Big\| & & \Big\| K_{pp} \\
\text{MetRS:Met+ATP} & \underset{K'_{ATP}}{\rightleftharpoons} & \text{MetRS:Met:ATP} & \underset{k_b}{\overset{k_f}{\rightleftharpoons}} & \text{MetRS:MetAMP:PPi}
\end{array}$$

Figure 4.4: **Met aminoacylation catalyzed by MetRS.** [Nigro et al., 2020]

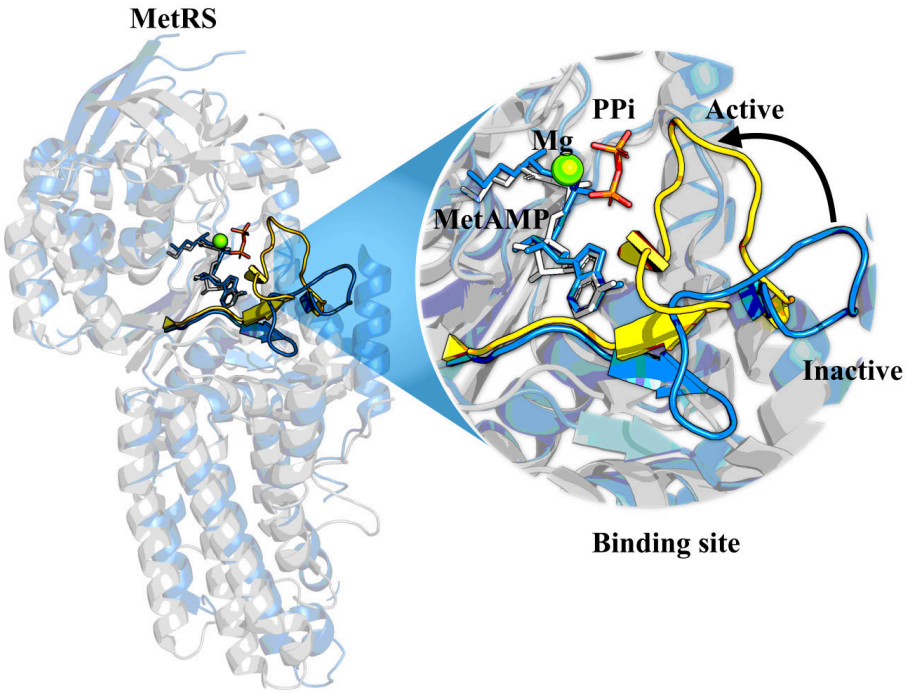


Figure 4.5: **`KMSKS` loop conformations and the binding site**. MetRS *E. coli.* (code 1PG0) is shown in blue. KMSKS loop is shown in yellow with the active conformation (code 3KFL from *L. major*).

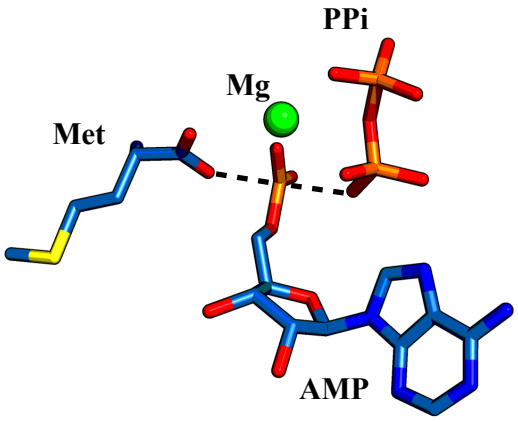


Figure 4.6: **Met transition state model [Met:ATP]‡**.

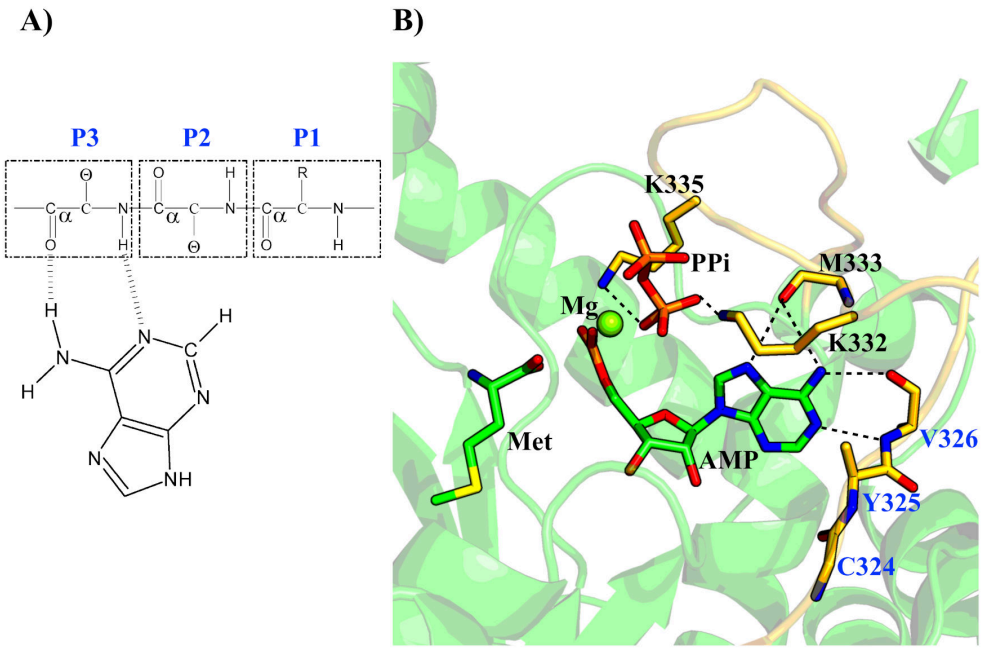


Figure 4.7: **Close view of the MetRS binding site with adnine fragment and pyrophosphate.** A) Figure adapted from [Denessiouk and Johnson, 2003] shows the reverse adenine binding conformation. B) It shows the stabilization of pyrophosphate and adenine fragments.

4.2.2 β amino acids

β amino acids have an additional methylene in between the amino and carboxylate groups (figure 4.8). They can allow new geometries for protein backbones and reduce the recognition by proteases ([Daura et al., 2001]). They can also change the alternation of side chain orientations.

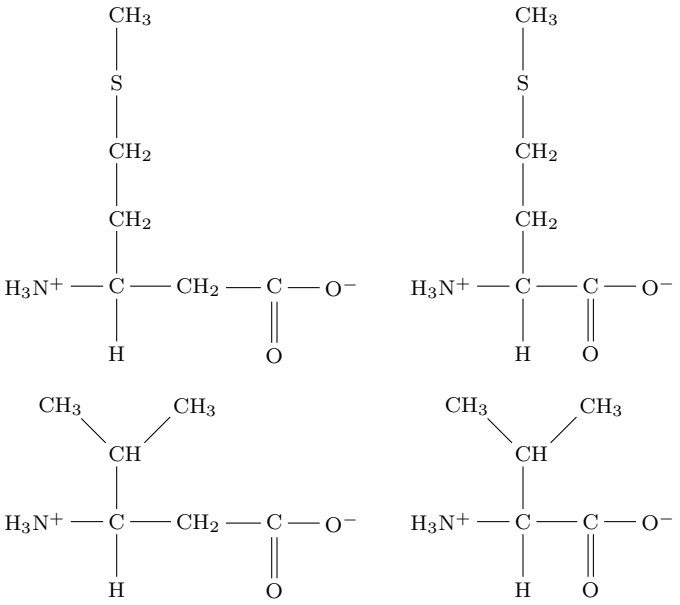


Figure 4.8: **β amino acids, the examples of β-Methionine and β-Valine.** In the right are shown canonical amino acids. In the left are shown β amino acids.

4.3 Theoretical methods

Here, we use Monte Carlo approaches implemented in the Proteus software ([Simonson et al., 2013, Simonson, 2019]). First, we recall the Monte Carlo sampling and how free energy is derived from simulations. Next, we present a new method to select mutating positions.

4.3.1 Design of proteins with a Monte Carlo approach

Proteus is based on three components: a discrete confromational space, an energy function based on molecular mechanics, and a Monte Carlo search algorithm.

4.3.1.1 Bound and unbound modelisation of a polypeptide

CPD needs a rapid evaluation of energies. To achieve this, we use an approximation in which the backbone is held fixed and the side chain flexibility is modeled with a discrete set of conformations called rotamers. Here, we use the Tuffery library ([Tuffery et al., 1997]), which has 17 rotamers for Met for example (figure 4.9). Once the ligand is placed in the binding site, we assign a set of rotamers to model its flexiblity as well (based on an existing library).

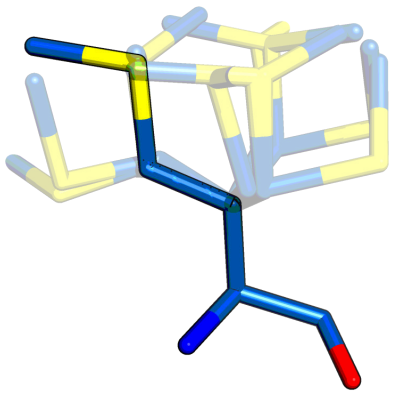


Figure 4.9: **Methionine rotamers**

4.3.1.2 Energy function

The energy function takes the form $E_{tot} = E_{vac} + \Delta G_{solv}$, where E_{vac} is the potential energy in gas phase and ΔG_{solv} is the free energy of solvation. For the potential energy, we use the ff99sb force field ([Cornell et al., 1996]). The electrostatic contribution is modeled with a Coulomb potential. Hydrogen bonds are included in the electrostatic contribution with their specific partial charge parameters.

Generalized Born (GB) model the polar solvent contributions ([Schaefer and Karplus, 1996]), and Lazaridis-Karplus model is used for the non-polar solvent contributions ([Lazaridis and Karplus, 1999, Michael et al., 2017]). Here, we used two recent GB variants ([Villa et al., 2017]): the Native Environment Approximation (NEA), where GB radii are computed in the native conformation, and the Fluctuating Dielectric Boundary (FDB) in which GB radii take into account the changes of each atom's environment along the MC simulation.

4.3.1.3 Monte Carlo exploration

Monte Carlos allows us to sample a probability distributions. Here, it allows us to sample variants S of a protein according to Boltzmann distribution, $P(S) = \frac{1}{Z} exp(-\frac{E(S)}{kT})$. $E(S)$ is the variant energy, Z is the partition function, k is Boltzmann constant, and T is the temperature.

First, we consider a polypeptide with a sequence $S = S_1, S_2, ...S_p$. We assign to each position S_i a rotamer r_i. We denote $C(S) = (r_1, r_2, ..., r_p)$ the conformation of S. We consider two sampling moves, the change of a rotamer ($\mathrm{r_i} \rightarrow \mathrm{r'_i}$) or the change of a type ($\mathrm{S_i} \rightarrow \mathrm{S'_i}$). Next, we create an ergodic and reversible Markov chain using the Metropolis-Hasting algorithm ([Hastings, 1970]). A Markov chain is a sequence of moves in the protein sequence/conformation space $S \rightarrow S' \rightarrow ... \rightarrow S''$. The sampling of variants along this sequence converges to a unique distribution, the Boltzmann distribution. The probability of a move depends on the change of energy $\Delta E(S \rightarrow S')$. Let $P(S \rightarrow S')$ be the probability to chose the move $S \rightarrow S'$. The probability of such a move is $\pi(S \rightarrow S') = P(S \rightarrow S')acc(S \rightarrow S')$, where $acc(S \rightarrow S') = min(1, exp(-\frac{\Delta E(S \rightarrow S')}{kT})\frac{P(S \rightarrow S')}{P(S' \rightarrow S)})$ is the acceptance probability. Under a detailed balance hypothesis, the populations of sequences $N(S)$ at equilibrium follow the Boltzmann distribution.

Since populations follow the Boltzmann distribution, each sequence has the probablitiy $P(S) = \frac{1}{Z} \sum_C exp(-\frac{E(C(S))}{kT})$ where $C(S)$ are the conformations. Therefore, $-kTln(P(S')/P(S)) = \Delta G(S \rightarrow S')$ gives the free energy change for a sequence mutation. We have:

$$\begin{aligned} \frac{P(S')}{P(S)} &= exp\big(-\frac{\Delta G(S \rightarrow S')}{kT}\big) \\ \Delta G(S \rightarrow S') &= G(S') - G(S) \end{aligned} \tag{4.16}$$

$G(S)$ is the free energy of S (respectively S').

Now, we consider the complex with a ligand L. Populations in the bound state obtained from a MC simulation still follow equation 4.16. Bound state free energies are denoted with an L subscript. We can derive the relative free energy of binding:

$$\Delta\Delta G(S \rightarrow S') = \Delta G_L(S \rightarrow S') - \Delta G(S \rightarrow S') \tag{4.17}$$

With two simulations, one can derive relative binding free energies of a ligand L if mutations have been sampled in both independent simulations. This raises one important issue. Since

sequences are populated exponentially according to free energy differences, only a fraction of sequences are sampled in each simulation. In practice, only a handful of sequences overlap in the two simulations. Therefore, we need a more sophisticated method.

4.3.1.4 Adaptive landscape flattening with Monte Carlo simulation

To tackle the sampling issue, we use a method called Adaptive Landscape Flattening (ALF) ([Villa et al., 2018, Bhattacherjee and Wallin, 2013]). For a polypeptide of p positions, first we consider it in the absence of ligand. Then, calculations will be repeated in the presence of a ligand L. Corresponding free energies will be denoted with an index u ou b, for unbound or bound. ALF is based on two separate MC simulations. First, we develop a bias potential such that all sequences are sampled with comparable probabilities. Next, we apply the bias to a simulation. It allows us to compute free energy changes for almost the entire sequence space considered.

For the first step, the bias potential $E_u^B(S;t)$ of a sequence S at time t takes the following form ([Villa et al., 2018, Villa and Simonson, 2018]):

$$E_u^B(S;t) = \sum_i E_i^B(S_i(t);t) + \sum_{i<j} E_{ij}^B(S_i(t), S_j(t);t) \tag{4.18}$$

$E_i^B(S_i;t)$ is the bias term associated with side chain type S_i at position i. $E_{ij}^B(S_i(t), S_j(t);t)$ is another bias term for the position pair i and j. After a segment of T MC steps, the bias is incremented as follows ([Villa and Simonson, 2018, Villa et al., 2018]):

$$\begin{aligned} e_i^B(S_i(t);t) &= e_0 \times \exp(-E_i^B(S_i(t);t)/E_0) \\ e_{ij}^B(S_i(t), S_j(t);t) &= e_0 \times \exp(-E_i^B(S_i(t), S_j(t);t)/E_0) \end{aligned} \tag{4.19}$$

where $e_i^B(S_i(t);t)$ and $e_{ij}^B(S_i(t), S_j(t);t)$ are increments added to the corresponding bias terms. e_0 is the initial bias increment. This update rule is borrowed from well tempered metadynamic ([Barducci et al., 2008]). E_0 controls the speed at which the increment decreases (high values mean a low decrease rate). At the end of the adaptive simulation, we have obtained the bias $E_u^B(S) \equiv E_u^B(S;t)$. In pratice, the bias doesn't need to flatten perfectly the sequence space distribution.

In a second step, we produce a biased simulation in which we apply the bias learned above. Sequence probabilities are now governed by:

$$\begin{aligned} \frac{\tilde{P}(S')}{\tilde{P}(S)} &= \exp(-\Delta\tilde{G}_u(S \to S')/kT) \\ \Delta\tilde{G}_u(S \to S') &= [G_u(S') + E_u^B(S')] - [G(S) + E_u^B(S)] \end{aligned} \tag{4.20}$$

$\tilde{G}_u(S \to S')$ is the free energy change in the biased simulation. If the bias flattened the distribution perfectly, one would have $\tilde{G}_u(S \to S') = 0$. We convert population ratios into free energies and remove the bias contribution ($E_u^B(S)$) to obtain the mutation free energy changes:

$$\begin{aligned} \Delta\Delta G_u(S \to S') &= -kT \times \ln\left(\frac{\tilde{P}(S')}{\tilde{P}(S)}\right) - \Delta E_u^B(S \to S') \\ \Delta E^B(S \to S') &= E^B(S') - E^B(S) \end{aligned} \tag{4.21}$$

Now, we apply the same procedure to the complex with the ligand. First, we build a bias potential (E_b^B) from an adaptive simulation. Next, we produce a simulation including the bias. We obtain the mutation free energy changes in the bound state $\Delta G_b(S \to S')$. Finally, we subtract the unbound energy to obtain the relative binding free energy:

$$\Delta\Delta G(S \to S') = \Delta G_b(S \to S') - \Delta G_u(S \to S') \tag{4.22}$$

If L was a transition state $L^\ddagger$, $\Delta\Delta G(S \to S')$ would be the catalytic efficiency. Also, we can apply this procedure to a second ligand, say L', in order to derive selectivity estimations.

The method can be slightly modified to allow the sampling of variants directly on their binding free energy. Instead of applying the bias derived from the bound state ($E_b^B(S)$), we apply the unbound state bias ($E_u^B(S)$) to sample sequences in the bound state. If the bias achieves near-perfect flattening, sequences are populated according to their binding free energy:

$$\begin{aligned} \Delta\tilde{G}^\ddagger(S \to S') &= [G_b(S) + E_u^B(S)] - [G_b(S) + E_u^B(S)] \\ \Delta\tilde{G}^\ddagger(S \to S') &\approx [G_b(S) - G_u(S)] - [G_b(S) - G_u(S)] \\ \Delta\tilde{G}^\ddagger(S \to S') &\approx \Delta G^\ddagger(S \to S') - \Delta G^\ddagger(S \to S') \\ \Delta\tilde{G}^\ddagger(S \to S') &\approx \Delta\Delta G^\ddagger(S \to S') \end{aligned} \tag{4.23}$$

Finally, we apply equation 4.22, where the bias cancels out since it is the same in both states. In practice, ALF can be applied on a small number of mutable positions, 4-5 positions, since it needs to sample heavily each sequence in order to make a robust statistical estimation.

4.3.1.5 Screening method for position selection

A pertinent selection of mutable positions is crucial for ALF, since only 4-5 positions are allowed to mutate. To fully explore a binding site of 20 positions, we would need to consider 4845 possible sets of 4 positions. With four simulations per quartet, we would have to perform 19380 MC simulations. Only a fraction of these combinations would produce good variants. Here, we describe a screening method that allows the selection of positions of interest. First, we select 20-30 positions within a distance threshold to the ligand (panel A, figure 4.10). We form pairs of positions whenever the C_α-C_α distance is below 12 Å (panel B, figure 4.10). For each such pair of positions, we perform an ALF to evaluate the catalytic efficiency of each pair of residue types (panel C, figure 4.10). Figure 4.10 shows eight ALFs, one for each pair.

For each pair of positions, we select the best pair of residue types with a catalytic efficiency threshold. We use the average catalytic efficiency to score the pair. The residues observed in those pairs are stored and will be used to restrain the mutation space of future quartets. Finally, the score of each quartet is the average score of the pairs of which it consists. All 4845 quartets can be evaluated with this score. A few will be selected for further investigation.

This score reflects the quality of the pairs that compose a quartet. We assume that a score based on pairs is enough to describe the quality of a given quartet. This might to simplistic if there are too many correlations within the quartet.

4.4 Structural models

4.4.1 KMSKS loop conformations

The *Escherichia coli* (*E. coli*) MetRS model started with three crystal structures (tableau 4.1). First, we used *E. coli* MetRS structure (PDB 1PG0 [Crepin et al., 2003]) with the KMSKS loop in the inactive state (in blue, figure 4.6). Then, the active conformation was modeled with *Leishmania major* (*L. major*) MetRS structure (PDB 3KFL [Larson et al., 2011]) with

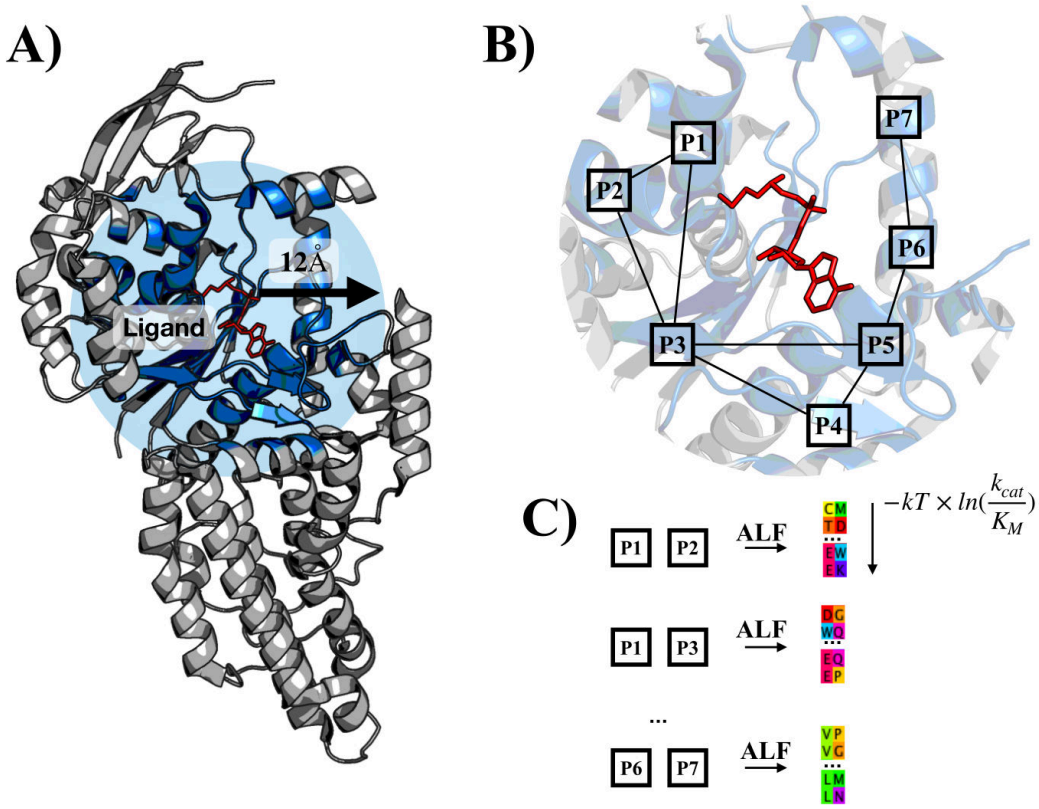


Figure 4.10: **Representation of the selection of positions and pairs.** A) Only positions close to the binding are considered. B) Closed pairs only are considered. For the sake of simplicity, only seven positions are shown by numbered squares. C) Catalytic efficiency is computed for each pair of residues with an ALF. Residue pairs are sorted according to catalytic efficiency.

the alignment of common ligand fragments in both structures. *L. major* was mutated into the *E. coli* sequence with Scwrl4 ([Krivov et al., 2009]). Finally, we adjusted its geometry using 40 steps of conjugate gradient minimization to obtain a model of *E. coli* MetRS with the KMSKS loop in the active conformation. Adenylate and pyrophosphate fragments were used to align ATP in the binding site. The Mg^{2+} ion was already in 3KFL structure and was transferred to the new model. We used visual inspection to assign histidine protonation states. Other ionisable groups were held in their standard protonation state. We call this complex MetRS:ATP.

Tableau 4.1: **Experimental structures used to build the active model for *E. coli* MetRS**

	1PG0	3KFL	6SPN
Resolution (Å)	1.9	2.00	1.45
ligand	Methionine phosphinate	Methionyladenylate + PPi	β-Methionine
KMSKS conformation	inactive	active	inactive
organism	*E. coli*	*L. major*	*E. coli*
MG^{2+}	Non	Oui	Non

4.4.2 Ligand: force field and catalytic pose

For the β-Met side chain pose, we used a recent complex of *E. coli* MetRS with β-Met (PDB 6SPN [Nigro et al., 2020]). The experimental data showed two conformations for the carboxylate fragment (figure 4.11). The B conformation is closer to a catalytic geometry although it is the less populated (30% occupancy in the crystal). We aligned MetRS:ATP with that structure to form RS:β-Met (figure 4.14). If the ligand in the binding site is the canonical Met, we call it RS:α-Met.

To create the complex with the transition state [β-Met:ATP]‡, we started from the RS:β-Met complex and [Met:ATP]‡ active geometry (figure 4.14). We apply harmonic constraints to maintain the B conformation of the carboxylate. We add a bond of 2.4 Å between the P_α and the carboxylate. We apply planar constrains to the phosphate α fragment. Then, we use a few steps of minimization to obtain the β-Met transition state [β-Met:ATP]‡. We call this complex RS:[β-Met:ATP]‡. If the transition state in the binding site is the canonical Met, we call it MetRS:[α-Met:ATP]‡. For β-Val models, we aligned the side chain fragment to the β-Met side chain.

β-Met flexibility is limited to its side chain. The other fragments are held fixed. To model

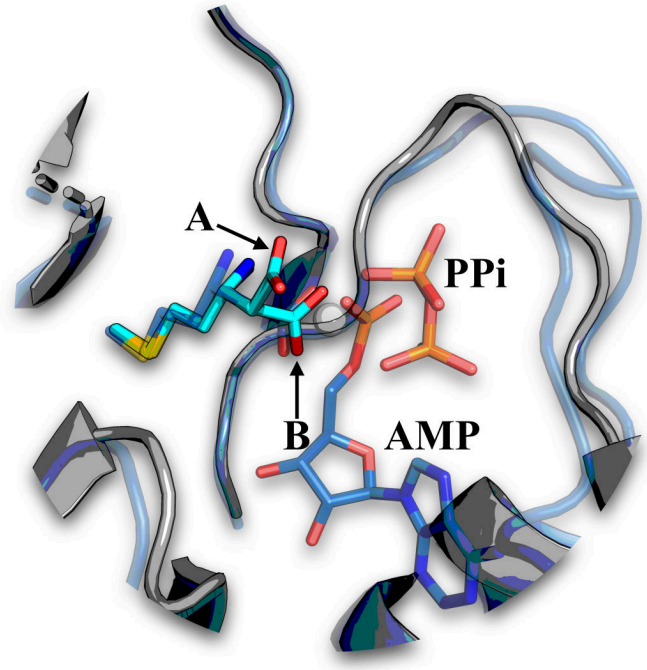


Figure 4.11: **β-Met observed conformations in the crystal structure (6SPN)** A conformation is the most populated (70% occupancy). B is the second conformation observed (30% occupancy). In blue and transparent, we show the β-Met transition state. In grey in the crystal structure 6SPN backbone.

Met flexibility, we used Tuffery Met rotamers ([Tuffery et al., 1997]). β-Val rotamers (initially three) are enriched with intermediate χ angles, to give 17 β-Val rotamers (figure 4.12). Table 4.2 shows the list of structural models.

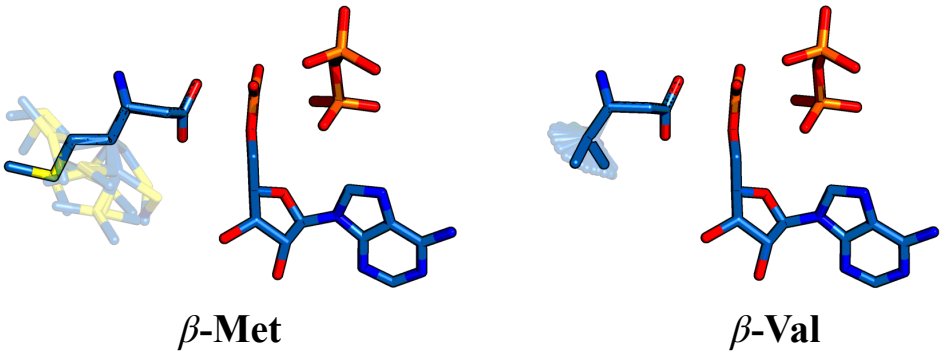


Figure 4.12: **[β-Met:ATP]‡** et [β-Val:ATP]‡ rotamers.

Tableau 4.2: **List of structural models used for catalytic efficiency, stability, and affinity estimations**

	MetRS:ATP	RS:β-Met	RS:[β-Met:ATP]‡	MetRS:[α-Met:ATP]‡	RS:[β-Val:ATP]
organism	*E. coli*	*E. coli*	*E. coli*	*E. coli*	*E. coli*
KMSKS conformation	active	active	active	active	active
ligand	ATP	β-Met+ATP	[β-Met:ATP]‡	[Met:ATP]‡	[β-Val:ATP]‡
Mg^{2+}	oui	oui	oui	oui	oui

4.4.3 Backbone relaxation

We used short MD based relaxation, since we have observed that the sampling and free energy calculations are sensitive to this step. It seems that energy minimization procedures may overspecialize the backbone to the wild type sequence in one state, say APO. Then, the relative catalytic efficiency estimates may be shifted. This can impact the final free energy calculations. Conversely, it can overspecialize the structure for the bound state and shifts the wild type value in to the opposite direction.

First, systems are truncated at 25 Å from the ligand P_α and solvated in a large box of TIP3P water ([Jorgensen et al., 1983]). Then, we perform 100 conjugate gradient minimization steps. Harmonic restraints were applied to nonhydrogen atoms with force constants that decreased gradually from 5 to 0.5 kcal/mol/Å^2 except for groups near the truncation sphere. 575 ps of MD were performed with NAMD ([Phillips et al., 2005]). For the RS:[β-Met:ATP]‡ complex, we first performed the relaxation with RS:[β-Met:ATP]‡ and we aligned the transition state geometry on it. The protein geometry remains the same in both complexes.

4.4.4 Unfolded state

The unfolded state allows us to estimate roughly the decrease of stability of predicted variants. We model the unfolded state of a sequence as an extended peptide where the energy is the sum of position dependent terms ([Pokala and Handel, 2005]). For each mutable position, we computed the energy between atoms in the side chain, local backbone, and the two adjacent C_α positions. The energy term for a given side chain type is the average of the best rotamers at each position. We call this energy the reference energy (tableau 4.3). The sum of reference energies of a given sequence is its energy in the unfolded state.

Stability is the free energy difference between the unfolded and folded state. The folded

Tableau 4.3: **Reference energies (ref. ener.) for the 18 side chain types.** Mutation to Pro and Gly are not allowed.

type	ref. ener.	type	ref. ener.
ALA	6.80	ILE	10.15
ARG	-18.79	LEU	5.99
ASN	0.31	LYS	4.61
ASP	-4.23	MET	5.74
CYS	5.91	PHE	9.50
GLN	2.58	SER	3.58
GLU	-1.24	THR	3.07
HIS_δ	23.27	TRP	11.13
HIS_ϵ	22.50	TYR	6.21
HIP_+	27.14	VAL	5.46

state energy is computed for the complex MetRS:ATP. It implies that ATP and Mg^{2+} are already in the binding site, and the KMSKS loop is in the active conformation. Stability is therefore a more complex quantity since it takes into account the binding of ATP and Mg^{2+} and the loop conformation change.

4.4.5 Catalytic efficiency estimation

For the catalytic efficiency, we considered the process MetRS:ATP $\rightarrow$ RS:[β-Met:ATP]‡ since the experiments are performed under ATP saturation ([Nigro, 2019, Nigro et al., 2020]). MetRS:ATP is the complex of MetRS with ATP and Mg^{2+}, with the KMSKS loop in the active conformation. The catalytic efficiency does not take into account ATP binding or the loop conformation change.

ATP binding mode, in general, involves the Watson-Crick side of the base ([Denessiouk and Johnson, 20 Denessiouk and Johnson, 2003], [Denessiouk et al., 2001]) as shown in figure 4.7. We don't allow mutations in positions involved in the recognition. We assume that ATP binding is therefore constant for all the variants we will produce. The stability criterion will help filter out ones that may not reproduce such geometries and binding.

Figure 4.13 shows a view of MetRS binding sites in the states considered (table 4.2). These states allow us to compute catalytic efficiency for Met, β amino acids, and also the selectivity between activated ligands. The selectivity quantifies the preference for a ligand compare to another. It is measured by the binding free energy difference between two activated ligand considered ($\Delta\Delta G^\ddagger$). As shown in figure 4.13, the selectivity in favor of β-Met with respect to

α-Met involved the change MetRS:[α-Met:ATP]‡ $\rightarrow$ RS:[β-Met:ATP]‡.

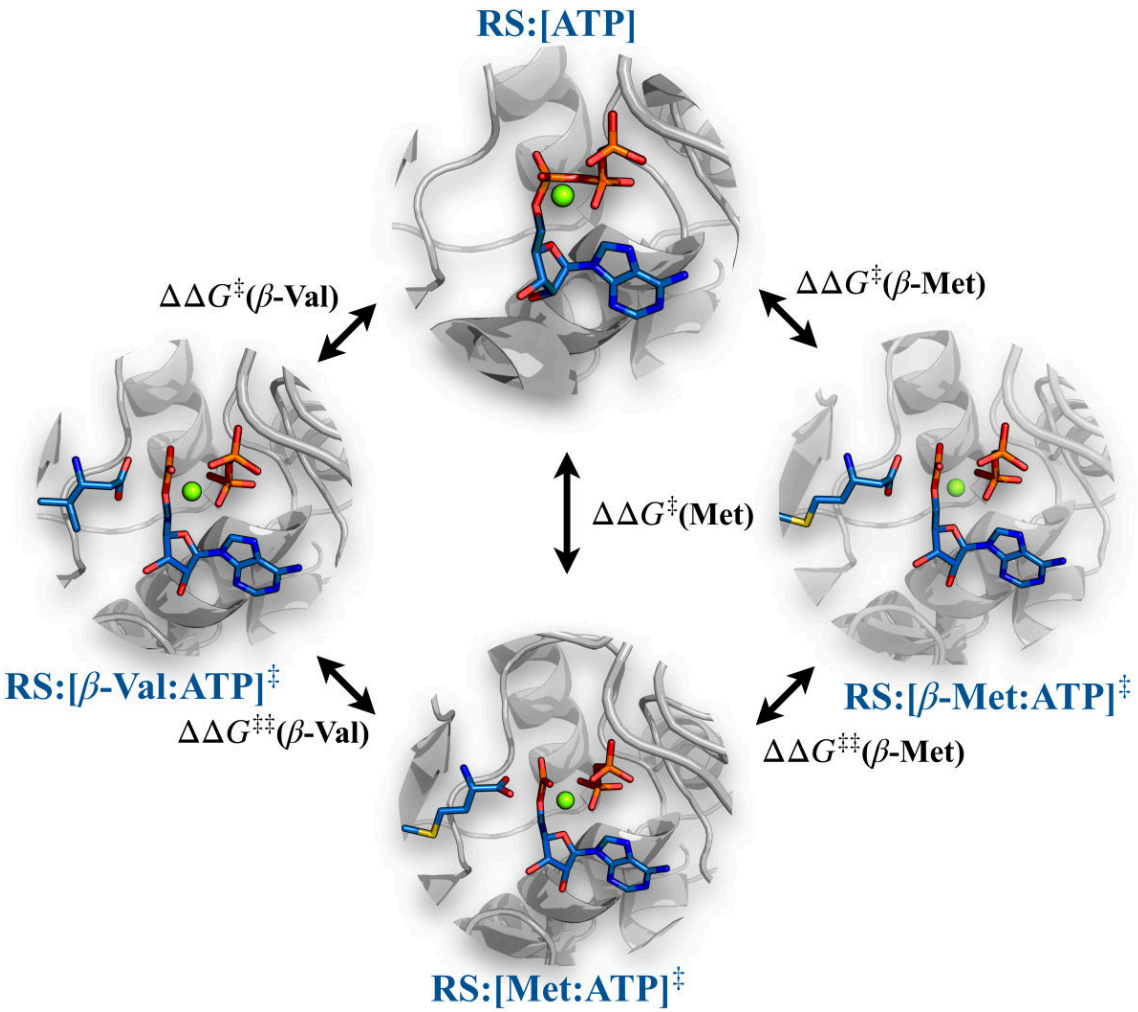


Figure 4.13: **Four structured states to evaluate the catalytic efficiency and selectivity.** We show a closed view of the four states {MetRS:ATP, MetRS:[α-Met:ATP]‡, RS:[β-Met:ATP]‡, RS:[β-Val:ATP]‡} binding sites. MetRS is in grey. Relative free energy changes are represented by arrows.

4.5 Numerical methods

Now, we present the parameters we used for the energy function and MC simulations. Finally, we present the screening approach we used for the selection of positions to design in both β-Met and β-Val complexes.

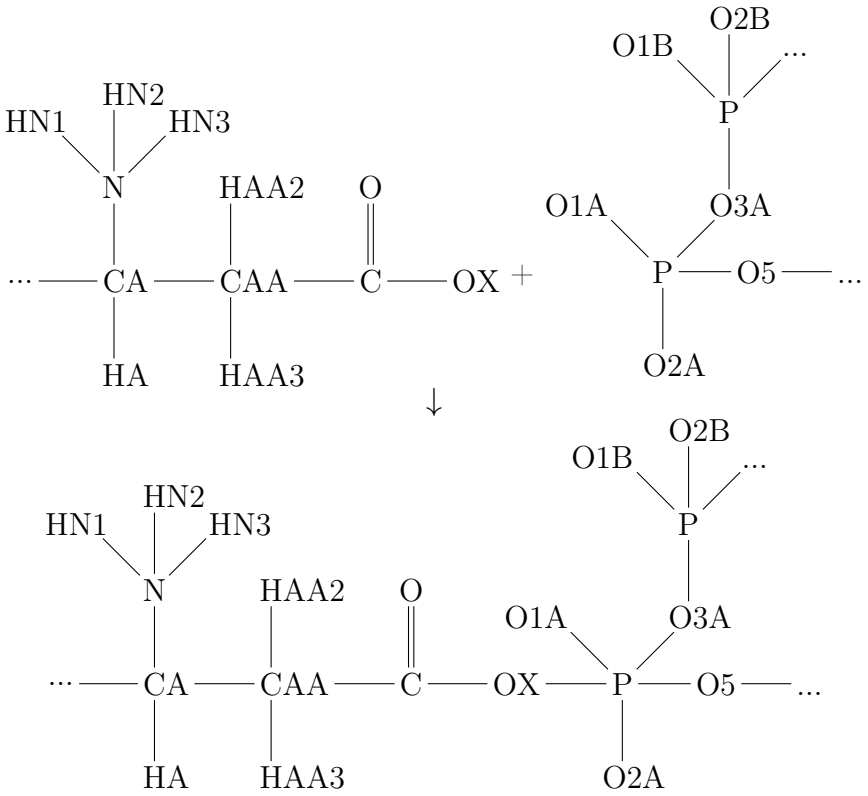


Figure 4.14: **Junction atoms for Met+ATP & [β-Met:ATP]‡**. β amine and carboxylate fragments (top left). α and β phosphate is showed in the top right corner. The transition state is showed in the bottom.

4.5.1 Energy function

The partial charges of ribose, adenine and side chain fragments were derived from existing parameters in ff99SB from analog fragments. For the junction atoms, we performed an HF/6-31G* *ab initio* calculation. Then, partial charges were chosen to reproduce the electrostatic potential according to Merz and Kollman ([Cornell et al., 1995]). These calculations were performed with Gaussian 9. This procedure is consistent with the Amber force field. β-Val charge parameters for the junction, the ATP, and the adenine fragments are the same. Bonded and van der Waals parameters were assigned by analogy to the α-Met model ([Opuu et al., 2020a]). Finally, bonded terms were assigned by analogy to existing parameters.

Tableau 4.4: **Junction charge parameters for β amino acids in RS:β-Met and RS:[β-Met:ATP]‡** ligands

Atomic name	Atomic type	RS:β-AA	RS:[β-AA:ATP]‡
MG	MG	1.5000	1.5000
N	N3	-0.3025	-0.3025
HN1	H	0.2770	0.2770
HN2	H	0.2770	0.2770
HN3	H	0.2770	0.2770
CA	CT	-0.2298	-0.2298
HA	HP	0.1208	0.1208
CAA	CT	-0.0620	-0.0620
HAA2	HC	-0.0554	-0.0554
HAA3	HC	-0.0554	-0.0554
C	C	0.8326	0.9610
O	O	-0.7856	-0.7856
OX	OX	-0.7257	-0.7517
PB	P	1.3586	1.4664
O1B	O2	-0.8280	-0.9582
O2B	O2	-0.8933	-0.8900
O3B	OS	-0.5746	-0.6252
PA	PA	1.2412	1.1805
O1A	O2	-0.6153	-0.6138
O2A	O2	-0.7853	-0.7016
O3A	OS	-0.7561	-0.8680
O5'	OS	-0.5025	-0.4478
C5'	CT	0.0558	0.0558

4.5.2 Parameters of MC simulations

4.5.2.1 Pair designs

Pair scores were computed from the aggressive ALF procedure applied to each pair of positions considered. Mutations to GLY and PRO are not allowed. 324 pairs of residues need to be visited per position pair. First, we flatten the sequence space for MetRS:ATP. Then, we flatten the activated states, MetRS:[α-Met:ATP]‡ and MetRS:[α-Met:ATP]‡.

We used 5 millions steps MC simulations with a bias update frequency T = 1000 steps, e_0 = 0.2 kcal/mol, E_0 = 40 kcal/mol, and the NEALK solvent model ([Villa et al., 2018]). Table 4.5 shows the other MC parameters. Only pair bias terms are used here. Then, biased simulations are performed. When the simulation is done, we remove from the mutation space types that need more than 10 kcal/mol of bias to be sampled at room temperature (kT=0.6).

If the bias does not allow the sampling of the wild type sequence and at least 95% of the mutation space, we add another iteration of ALF on top of the current bias. When the sampling

is sufficient, we compute the catalytic efficiency with respect to the wild type sequence.

4.5.2.2 Quartet designs

The quadruplets with high pair scores are investigated with a second MC simulation. We used an aggressive ALF procedure with the more accurate solvent FDBLK. We performed 10^8 MC steps to flatten the quadruplet sequence space at room temperatures for all the states: MetRS:ATP, RS:[β-Met:ATP]‡, and RS:[β-Val:ATP]‡. The bias is refined until the flattening is sufficient.

Tableau 4.5: **The list of MC simulation parameters.** Rot is the probability of changing one rotamer. Mut is the probability of one mutation. Rot-Rot is the probability of changing two rotamers. Rot-Mut is the probability of changing a rotamer then a type.

Parameters	Pair	Quartet
Energy function	NEALK	FDBLK
Number of MC steps	5.10^6	10^8
kT	0.6	0.6
Number of replicas	1	1
Bias parameters		
e_0	0.2	0.2
E_0	40	40
t	1000	1000
Move probabilities(a)		
Rot-Mut	0.1	0.1
Rot-Rot	0.9	0.9
Mut	0.1	0.1
Rot	0.9	0.9

4.5.3 Selection of mutable positions with binding site screening

First, we chose 19 positions from the first and second layers in the binding site within 20 Å of the α phosphorus (except Pro and Gly). Positions in the KMSKS loop are not allowed to mutate since they stabilize the PP_i fragment ([Schmitt et al., 1994]). Active positions are shown in table 4.6 and figure 4.15. Then, position pairs are formed when the C_α-C_α distance is less than 12 Å. 263 flexible positions are modelled with Tuffery rotamer libary ([Tuffery et al., 1997]) enriched with the native rotamers.

We considered 87 of the 171 possible pairs for the 19 positions selected. For each pair, we first flattened the position pair space with a MC of 5.10^6 steps. Then, a biased simulation

Tableau 4.6: **19 active positions selected in the binding site with a distance threshold of 12 Å.**

positions	types	positions	types	positions	types
11	CYS	50	ALA	253	TRP
12	ALA	51	ASP	256	ALA
13	LEU	52	ASP	293	ILE
15	TYR	97	THR	296	ASP
16	ALA	251	TYR	297	ILE
17	ASN	252	VAL	300	PHE
24	HIS				

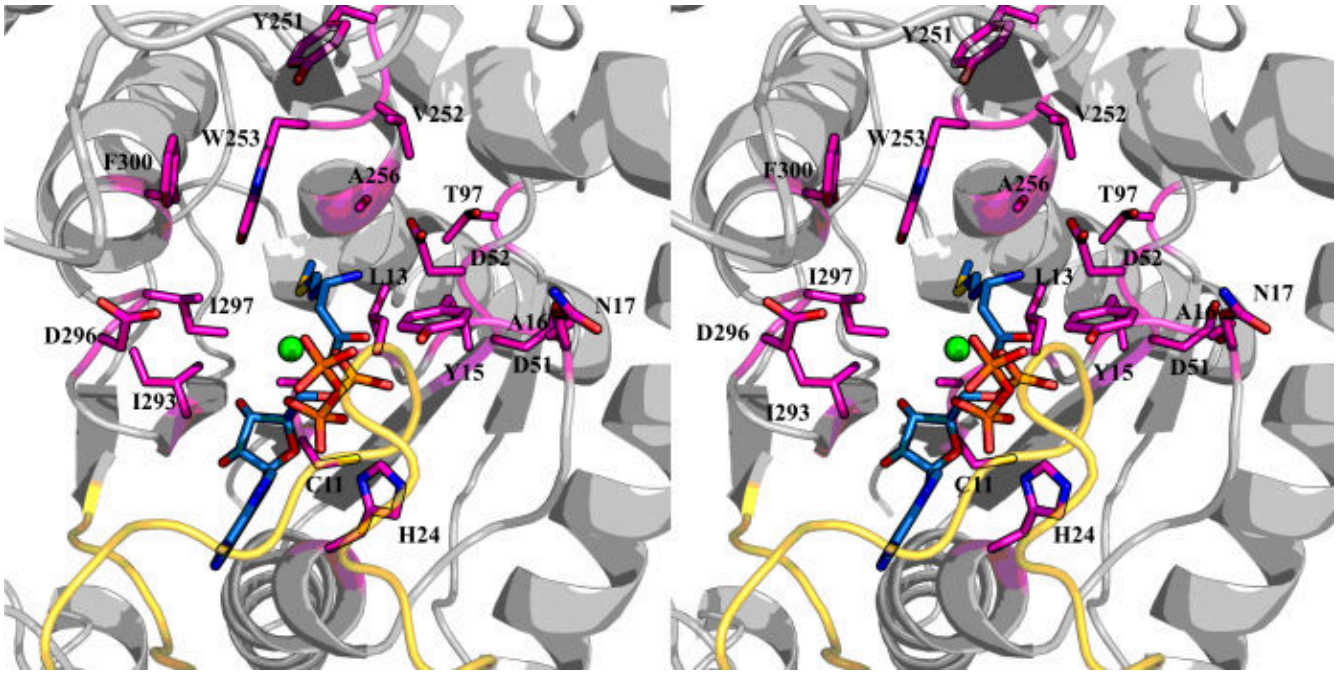


Figure 4.15: **Closed view (stereo) of MetRS binding site with RS:[β-Met:ATP]‡** ligand. The transition state RS:[β-Met:ATP]‡ is showed in blue, mutable positions are showed in magenta, and KMSKS loop in yellow.

is performed to compute stability and catalytic efficiency. We removed the unstable pairs of types. For each pair, the score is the average of the 10 best pairs of types sampled. The score of a quadruplet of positions is the average of pair scores.

Pair scores are used to selected the quartet of positions that are likely to produce effective variants. For β-Val, we computed the scores for the 3876 quadruplets of positions. Residue types sampled to build the pair scores is used to build the mutation spaces. For β-Met, the pair score is a bit different. Since positions 13 and 297 already provided active variants (see below) we allow LMCS types for 13 and IC for 297 positions. For these augmented pairs, we computed the average over the 100 best pairs of types.

To select quadruplets with β-Met we used a different approach. To build a given quadruplet, we start by a position i chosen among the 19 positions. Among the 18 positions left, we select j so the score is maximized. Now, we have two positions. Then, we select a third position k among the 17 position that maximize the pair score with j. Finally, we performed the same search for the 4^{th} position. We repeated the same procedure by starting with the 19 other positions. This gave 19 quartets of positions. We repeated the procedure twice, with the structure relaxed with either the MAC sequence or the wild type sequence for positions 13, 256, and 297.

Chapter 5

Engineering methionyl-tRNA synthetase for β amino acid activity: results

5.1 First search for active variants

We begin with a first round of predictions of variants for β-Met and β-Val activity. We investigated three positions selected manually: 13, 256, and 297. To design variants, we used the binding to the activation reaction products β-MetAMP and β-ValAMP. 20 variants were then selected for experiments. 11 were directly chosen from the sampling and 9 were derived from the mutation observed in the sampling. Among the 11 predicted variants, five have a weak but measurable catalytic efficiency. For three variants, the selectivity for α-Met is reduced by factors of 2-8. Details are in the next sections.

5.1.1 Affinity design for β-MetAMP and β-ValAMP

First, we modeled the binding of MetRS to β-MetAMP or β-ValAMP. Ligands were aligned in the binding site using the α-Met pose as described in [Opuu et al., 2020a]. In those models, the KMSKS loop is in the inactive conformation. For the APO system denoted *RS*, we removed the ligand from the binding site.

We considered the free energy changes RS $\rightarrow$ RS:β-MetAMP and RS $\rightarrow$ RS:β-ValAMP reactions. We used the FDBSA solvent with a dielectric constant of 80 for the water and 4 for the protein. Charge parameters were assigned from α-MetAMP model. Bonded and other

non-bonded parameters were assigned by analogy to existing parameters.

First, we flattened the apo sequence space with an adaptive MC simulation. Then, we applied the bias to the sampling of variants in the bound states with β-MetAMP and β-ValAMP. Variants were sampled directly according to their binding affinity for the products of the activation reaction. Table 5.1 shows the best variants for β-MetAMP and β-ValAMP binding affinity. Only variants with a stability loss below 3 kcal/mol with respect to the wild type sequence are shown.

Eleven variants were chosen for the experiments directly from the sampling based on the mutations observed in the predicted variants. Table 5.2 shows the bindings and catalytic efficiencies obtained experimentally. The selected variants are denoted with a $*$ of $**$ in table 5.1. Eleven variants have a measurable catalytic efficiency but only five were directly sampled in the MC simulation. However, four of the false positive are unstable variants (stability $>$ 3 kcal/mol). Here, the wild type variant LAI was not sampled, therefore, the affinities are estimated with respect to a slightly different variant, LAA. The catalytic efficiency of the tested variants were not better than the wild type for β-Met but three improved slightly the selectivity in favor of β-Met. These variants all have the mutation I297C.

For β-ValAMP, MAC is the best predicted variant with 6.3 kcal/mol gain in binding free energy compare to the wild type sequence. The I297C mutation appears in the six best β-Val variants. None of the 19 best variants have the A256S mutation which appeared in the best β-MetAMP predicted variants.

Tableau 5.1: **Best variants with a stability threshold of 3 kcal/mol sorted by the β-MetAMP binding (left) or β-ValAMP binding (right).** β-Met variants free energy binding are compared to the reference sequence LAA since LAI (the wild type sequence is not sampled here). Variants with a measurable catalytic efficiency are annotated with ** and other tested variants are annotated with *. Not sampled variants with β-MetAMP (respectively β-ValAMP) are assumed to have a free energy of binding $>$ 0.88 kcal/mol ($>$ 1.2 kcal/mol for β-ValAMP).

reference	LAI	LAA	LAI		LAI	LAA	LAI
variants	stability	β-Met	β-val	variants	stability	β-Met	β-val
MAA	2.01	-3.50	-1.93	MAC	2.28	-3.41	-6.29
MAC**	2.28	-3.41	-6.29	CAC	-7.87	-0.57	-4.22
MAV**	1.28	-2.38	-2.67	SAC	-4.00	0.17	-4.13
CSA	-6.34	-2.18	0.82	LAC	-0.32	0.88<	-3.80
HSA	2.65	-2.12	1.2<	AAC	-7.59	0.32	-3.33
CSC	-6.18	-2.11	-0.97	HAC	0.35	0.88<	-2.74
LSC	1.25	-2.06	-0.46	MAV	1.28	-2.38	-2.67
LSA	1.05	-1.97	1.2<	CAT	-5.65	0.88<	-2.53
SSA	-2.46	-1.79	-0.04	CAS	-6.18	-0.04	-2.49
ASA	-5.92	-1.78	1.2<	SAT	-1.72	0.88<	-2.38
CSS	-4.38	-1.61	1.2<	SAS	-2.35	0.88<	-2.25
HSC	2.77	-1.60	1.2<	LAS	1.33	0.88<	-1.93
CSV	-7.00	-1.50	1.2<	MAA	2.01	-3.50	-1.93
CST	-3.90	-1.46	1.2<	LAT	1.78	0.88<	-1.78
ASC	-5.76	-1.24	0.19	AAT	-5.40	0.88<	-1.70
SSC*	-2.21	-1.15	-0.83	AAS	-5.90	0.88<	-1.45
HSV	1.86	-1.10	1.2<	MAI	2.53	-0.42	-1.39
LSV	0.36	-0.92	1.2<	AAD	0.06	0.88<	-1.34
CAA	-8.10	-0.70	1.2<	CAD	-0.34	0.88<	-1.28
ASS	-4.06	-0.61	1.2<	CSC	-6.18	-2.11	-0.97
CAC**	-7.87	-0.57	-4.22	HAS	2.95	0.88<	-0.94
SSV	-3.21	-0.57	1.2<	HAT	2.53	0.88<	-0.89
SST	0.01	-0.42	1.2<	SSC	-2.21	-1.15	-0.83
MCC	1.94	-0.42	-0.60	MCC	1.94	-0.42	-0.60
MAI	2.53	-0.42	-1.39	CAV	-8.78	0.34	-0.58
HAA	2.17	-0.36	-0.33	LSC	1.25	-2.06	-0.46
TSA	-1.48	-0.33	1.2<	HAA	2.17	-0.36	-0.33
AAA	-7.84	-0.23	1.2<	SAV	-5.02	0.88<	-0.20
ASD	1.96	-0.23	0.45	SSA	-2.46	-1.79	-0.04
CSD	1.49	-0.21	0.64	HAV	1.38	0.88<	-0.04
SSS	-0.70	-0.10	1.2<	LAI	0.00	0.88<	-0.00
CAS	-6.18	-0.04	-2.49	MCV	1.02	0.88<	0.16
LAA*	-0.62	-0.00	1.2<	ASC	-5.76	-1.24	0.19
TSS	0.35	0.00	1.2<	CCS	-6.36	0.88<	0.33
LSI	1.64	0.11	1.2<	SCA	-4.29	0.88<	0.43
SAC**	-4.00	0.17	-4.13	ASD	1.96	-0.23	0.45
AAC	-7.59	0.32	-3.33	AAV	-8.57	0.88<	0.58
ASV	-6.54	0.33	1.2<	LAV	-1.33	0.88<	0.58
CAV**	-8.78	0.34	-0.58	CSD	1.49	-0.21	0.64
MST*	6.21	-3.97	-1.28				
MSV*	3.06	-3.95	0.64				
MAT*	4.56	-2.70	-4.55				
MSI*	4.36	-2.36	1.2<				

Tableau 5.2: **Experimental biding affinity and selectivity for variants of positions 13, 256, and 297 for β-Met.** ND = not determined (details about experiments are in [Nigro, 2019]). $\frac{\text{specificity}}{\text{specificity(WT)}}$ is the reduction of selectivity in favor of α-Met, relative to the wild type sequence LAI.

variants	K_M (β) (mM)	k_{cat} (β) ($10^{-3}s^{-1}$)	$\frac{k_{cat}}{K_M}$ (β) ($10^{-3}s^{-1}mM^{-1}$)	$\frac{k_{cat}}{K_M}$ (α) ($s^{-1}mM^{-1}$)	$\frac{k_{cat}/K_M(\alpha)}{k_{cat}/K_M(\beta)}$ (10^3)	$\frac{\text{specificity}}{\text{specificity(WT)}}$
CAC*	6.3	21	3.4	9.8	2.9	7.7
MAC*	7.7	16	2.1	14.0	6.7	3.3
SAC*	20	22	1.1	12.0	10.9	2.0
LAC	3.5	18	5.0	108.0	21.6	1.0
LAI	0.4	51	138	3073.0	22.3	1.0
LAT	6.2	5.2	0.83	26.0	31.3	0.7
MAV*	7.1	5.4	0.75	28.0	37.3	0.6
CAI	2.6	21	8.9	412.0	46.3	0.5
SAI	4.2	23	5.4	326.0	60.4	0.4
CAV*	38	13	0.35	23.0	65.7	0.3
LAV	11	8.0	0.74	66.0	89.2	0.2
SSI	7.5	$<$3.5	ND	70.0	ND	ND
MAT	8.2	$<$3.6	ND	4.9	ND	ND
MSI	8.4	$<$3.7	ND	27.0	ND	ND
MST	19	$<$5.7	ND	0.5	ND	ND
LAS	21	$<$6.1	ND	0.4	ND	ND
MSV	39	$<$10	ND	1.5	ND	ND
PAI	47	$<$11	ND	3.1	ND	ND
SSC*	60	$<$14	ND	4.5	ND	ND
LAA*	180	$<$38	ND	0.5	ND	ND

5.1.2 Residence time in molecular dynamic simulations

We investigated a few variants with molecular dynamics simulations in explicit solvent, to calculate the residence times of the β amino acids. First we model the complexes with the β amino acids instead of the adenylate ligands. Then, the complexes were truncated at 25 Å from the ligand and solvated in a large water box. Harmonic constraints were applied to non-hydrogen atoms in the layer close to the truncated region. When the ligand RMSD is above 4 Å compare to the initial pose, we considered that the ligand has become unbound.

Table 5.3 shows the residence times for selected variants. For β-Val, three variants have a residence time above 30 ns (LAI, LAS, and MSC). MSC is the most stable variant in the selection although it has the worst stability estimation. Among the best predicted binders with the MC simulations, only one has a residence time above 40 ns. For β-Met, four variants have a residence time of at least 60 ns (CSS, MAC, MAV, and MSV). MAC and MAV have a measurable catalytic efficiency. CSS and MSV are false positives. Among all the variants and complexes, the wild type system is the most stable.

Tableau 5.3: **Residence time per complex and ligand β-Met and β-Val.** `aff.` and `sta.` are the stability and the binding free energy estimated with the MC simulations.

ligands	variants	sta.	aff.	Simulation time (ns)	Residence time (ns)	RMSD Å
α-Met	`LAI` (native)			80	_	1.06
β-Val	`CAC`	-7.87	-4.22	60	1	7.37
	`LAC`	-0.32	-3.80	80	20	4.43
	`LAI`	0.00	-0.00	70	_	2.62
	`LAS`	1.33	-1.93	30	_	2.62
	`MAC`	2.28	-6.29	60	1	4.70
	`MAI`	2.53	-1.39	46	13	4.55
	`MAV`	1.28	-2.67	60	36	5.09
	`MSC`	4.06	-3.15	80	_	1.81
	`SAC`	-4.00	-4.13	80	14	7.32
β-Met	`CSS`	-4.38	-1.61	69	_	1.62
	`CSV`	-7.00	-1.50	30	3	4.17
	`LAC**`	-0.32	0.88<	30	18	4.19
	`LSC`	1.25	-2.06	30	18	4.26
	`MAC**`	2.28	-3.41	60	_	2.65
	`MAS`	3.92	-2.17	30	_	3.01
	`MAT*`	4.56	-2.70	30	5	5.02
	`MAV**`	1.28	-2.38	60	_	2.33
	`MSA`	3.76	-4.89	60	30	4.6
	`MSC`	4.06	-4.85	60	5	4.11
	`MSS`	5.77	-3.91	30	3	4.83
	`MST*`	6.21	-3.97	60	58	4.14
	`MSV*`	3.06	-3.95	60	_	3.67

5.1.3 Selection using catalytic efficiency

To investigate further those three positions 13, 256, and 297 and improve the β-Met activity, we searched for variants according to the catalytic efficiency. Here, we modeled the MetRS:ATP $\rightarrow$ RS:[β-Met:ATP]‡ binding. The binding free energy correspond to the catalytic efficiency. We used an aggressive ALF to compute the catalytic efficiency of 2474 variants with respect to the wild type sequence.

Figure 5.1 shows all the variants catalytic efficiencies. Two groups appear: one group of stable variants but with positive catalytic efficiency and a second group of improved catalytic efficiency but unstable. This can be an artefact of the fixed backbone approximation. 86 variants have a stability loss of 2 kcal/mol or less when compared to the wild type. 8 of the 11 active variants are in this region of sequence space, {MAV,MAC,SAI,CAV,CAI,LAC,LAT,LAV}.

Table 5.4 compares predicted and experimental values of catalytic efficiency. LAV and CAV are overestimated by 1.7 and 1.2 kcal/mol. CAI, SAI, and LAT (simple mutants) have an absolute error of 1 kcal/mol. CAC, SAC, and MAV are underestimated by 2 kcal/mol. MAC is the worst prediction with an error of 4 kcal/mol. The average absolute error is 1.9 kcal/mol and the highest errors are for MAC and MAV.

Tableau 5.4: **Stability, catalytic efficiency (with experimental values) compared to the wild type sequence LAI for positions 13, 256, and 297**

variants	stability	$\Delta G^{\ddagger}$(β-Met) predicted	$\Delta G^{\ddagger}$(β-Met) experiments
CAC	2.2	4.5	2.2
CAI	0.9	1.0	1.6
CAV	1.5	2.4	3.6
LAC	1.4	3.2	2.0
LAI	0.0	0.0	0.0
LAT	1.7	2.3	3.1
LAV	0.6	1.4	3.1
MAC	0.8	8.1	2.5
MAV	0.0	6.1	3.1
SAC	2.5	5.3	2.9
SAI	1.1	2.0	1.9

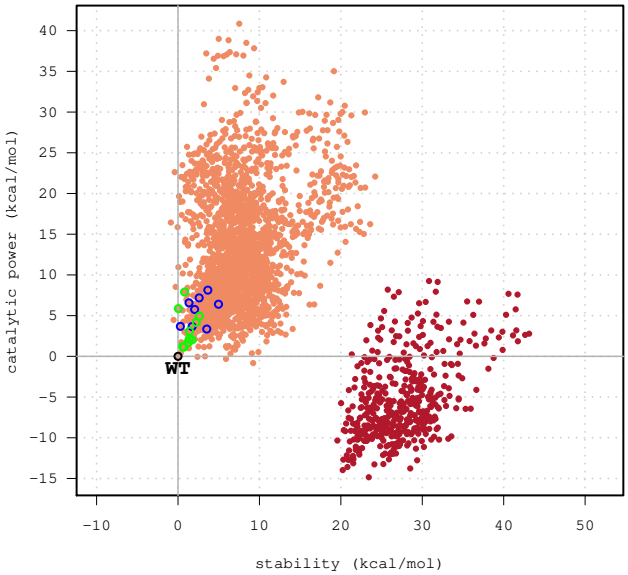


Figure 5.1: **Distribution of variants according to stability and catalytic efficiency for the positions 13, 256, and 297.** The reference sequence is LAI (Wt). Stable variants are shown in orange and least stable variants are shown in red. Variants with measurable activity are shown in green and blue dots are tested variants but without measurable activity.

5.2 Screening of pairs, formation of β-Met quadruplets

To improve the activity, we enlarged the search space to the whole active site, through a new screening approach. We modeled the binding MetRS:ATP $\rightarrow$ RS:[β-Met:ATP]‡. We considered 87 pairs of positions for which we computed the scores of pairs based on β-Met catalytic efficiency. From the screening we selected two quadruplets of positions in the wild type context and one quartet of positions in the MAC context. Talbes 5.5 and 5.6 show the selected positions.

Tableau 5.5: **Selected positions for the backbone relaxed in the wild type context.** In the top of the table are shown the variants selected with a stability threshold of 5 kcal/mol and an average score over the 100 best variants. In the bottom of the table are shown the selection of variants with a threshold of 10 kcal/mol and an average score over the best 200 variants. The selected four positions are annotated with a $*$. A mutation space is assigned to each position.

	Pos				Score	Space			
	a	b	c	d		a'	b'	c'	d'
	17	24	13	51	-2.3	EFMLQRTV	H	SQKMC	ACIKMLNSTV
	51	24	13	17	-2.1	ACKMLNS	H	SQKMC	EFMLQRTV
*	297	24	13	51	-2.1	ACSTV	QHIV	CHKMNQST	ACIKMLNSTV
	11	24	13	51	-1.9	AS	IHELV	CHKMNQST	ACIKMLNSTV
	24	13	51	11	-1.8	IHELV	CHKMNQST	ACIKMLNSTV	ADMNSTV
	52	24	13	51	-1.8	ACNSTV	H	SQKMC	ACIKMLNSTV
	252	24	13	51	-1.7	AHCST	H	SQKMC	ACIKMLNSTV
	296	24	13	51	-1.6	ACEFHJML NQSRTWVY	H	SQKMC	ACIKMLNSTV
	97	24	13	51	-1.6	ACS	H	SQKMC	ACIKMLNSTV
	12	24	13	51	-1.5	SC	IHMNTV	CHKMNQST	ACIKMLNSTV
	16	24	13	51	-1.5	SC	H	SQKMC	ACIKMLNSTV
	50	11	24	13	-1.3	CS	AS	IHELV	CHKMNQST
	293	11	24	13	-1.3	ATV	AST	IHELV	CHKMNQST
	17	24	13	51	-2.4	EFMLQRTV	H	CHKMQSH	ACIKMLNSTV
*	297	24	13	17	-2.3	ACSTV	EIHNQV	ACHKVMQSTH	EFMLQRTV
	51	24	13	17	-2.3	ACKMLNQST	H	CHKMQSH	EFMLQRTV
	52	24	13	17	-1.9	ACNSTV	H	CHKMQSH	EFMLQRTV
	11	24	13	17	-1.9	AST	IHELV	CHKMNQSTH	EFMLQRTV
	252	24	13	17	-1.8	ACHNST	H	CHKMQSH	EFMLQRTV
	24	13	17	297	-1.8	IHELV	CHKMNQSTH	EFMLQRTV	C
	296	24	13	17	-1.7	ACEFIHKJ MLNQSRTW HYV	H	CHKMQSH	EFMLQRTV
	16	24	13	17	-1.6	CST	H	CHKMQSH	EFMLQRTV
	97	24	13	17	-1.6	ACS	H	CHKMQSH	EFMLQRTV

Tableau 5.6: **Selected positions for the backbone relaxed in the MAC context.** In the top of the table are shown the variants selected with a stability threshold of 5 kcal/mol and an average score over the 100 best variants. The selected four positions are annotated with a *. A mutation space is assigned to each position.

Pos				Score	Space			
a	b	c	d		a'	b'	c'	d'
* 297	24	13	51	-6.25	CEDMLQSRV	SHCT	ACEIKMQSTV	H
51	24	13	297	-6.24	ACEIHMNQST	H	ACEIKMQSTV	C
				-6.24	V	H	ACEIKMQSTV	C
13	24	297	51	-6.23	ACIKMQSTV	SHCT	C	H
24	13	297	51	-6.22	ACHKMNQST	ACEIKMQSTV	C	H
252	24	13	297	-5.42	ACEDIKMLNQ	H	ACEIKMQSTV	C
				-5.42	SRT	H	ACEIKMQSTV	C
11	24	13	297	-5.40	ADIHNQSTV	HKV	ACEIKMQSTV	C
52	24	13	297	-5.22	ACEIHLNSTH	H	ACEIKMQSTV	C
				-5.22	V	H	ACEIKMQSTV	C
17	24	13	297	-5.20	ACEDFHKMLQ	H	ACEIKMQSTV	C
				-5.20	SRTWYV	H	ACEIKMQSTV	C
97	24	13	297	-5.03	ACDIHKMNQS	H	ACEIKMQSTV	C
				-5.03	RV	H	ACEIKMQSTV	C
296	24	13	297	-4.69	ACEFIHKMLN	H	ACEIKMQSTV	C
				-4.69	QSRTWVY	H	ACEIKMQSTV	C
50	24	13	297	-4.62	CEDNQSTHV	H	ACEIKMQSTV	C
16	24	13	297	-4.56	CSTV	H	ACEIKMQSTV	C
12	24	13	297	-4.29	CENSTV	AHSV	ACEIKMQSTV	C
256	24	13	297	-4.22	SNTDV	H	ACEIKMQSTV	C

Tableau 5.7: **Selected quadruplets of positions and the assigned mutation space.** NB. = the number of possible variants.

	positions	Initial types	Mutation space	Nb.
Q1	13	LEU	LEU CYS HIS LYS MET ASN GLN SER THR	2970
	24	HIS	GLN HIS ASN ILE VAL	
	51	ASP	ASP ALA CYS ILE LYS MET LEU ASN SER THR VAL	
	297	ILE	ILE ALA CYS SER THR VAL	
Q2	13	LEU	LEU ALA CYS HIS LYS VAL MET GLN SER THR	3564
	17	ASN	ASN GLU PHE MET LEU GLN ARG THR VAL	
	24	HIS	HIS GLU ILE ASN GLN VAL	
	297	ILE	ILE ALA CYS SER THR VAL	
Q3	13	MET	ALA CYS GLU ILE LYS MET GLN SER THR VAL	720
	24	HIS	SER HIS CYS THR	
	51	ASP	ASP HIS	
	297	CYS	CYS GLU ASP MET LEU GLN SER ARG VAL	

Q1 is the first selected quartet of positions 13, 24, 51, and 297 with a score of -2.1 kcal/mol. We used a stability threshold of 5 kcal/mol and a score of pairs averaged over the 100 best visited pairs of types. Two positions were among the three positions investigated earlier. Q1 contains 2970 different variants. Q2 is composed of positions 13, 17, 24, and 297 with a score of -2.3 kcal/mol averaged over the 200 best pairs of types sampled. It allows us to enlarge the search to less stable pairs of residues. Q3 is composed of positions 13, 14, 51, and 297. It was obtained using the backbone relaxed in the MAC context with a score of -6.25 kcal/mol (table 5.6). We used a stability threshold of 5 kcal/mol and averaged the pair scores over the 100 best pairs of types. Q3 contains 720 variants.

5.3 Design of β-Met quadruplets

For Q1, we sampled all the variants in MetRS:ATP and RS:[β-Met:ATP]‡ states at least 1000 times per variant. We estimated the catalytic efficiency for all the variants in Q1. Then, we ran a biased simulation of the complex MetRS:[α-Met:ATP]‡ for to calculate the β-Met selectivity. The variants with a stability loss compared to the wild type sequence are removed. Variants that do not improve the catalytic efficiency are also removed. From the 95 variants left, we removed the variants those with a selectivity of more than 2 kcal/mol against β-Met, to obtain 89 variants (table 5.8). The best variant is `TQAI` with a catalytic efficiency of -4.3 kcal/mol compared to the wild type sequence.

To show the composition and the co-occurrences of types for the 89 variants selected, we used

a circular logo representation (figure 5.2). We computed the co-occurrences of types t_i, t_j with: $c(t_i, t_j) = P(t_i, t_j) \times log(\frac{P(t_i,t_j)}{P(t_i)\times P(t_j)})$. An arc represents the co-occurrences if $c(t_i, t_j) \geq 0.03$. Positions 13 and 297 are populated by native types Leu and Ile. Mutation to small hydrophobic side chains are observed for positions 24 and 51 where the native types are His and Asp. Three pairs of correlated side chains appeared between the positions 13 and 24. Correlation is also observed between types Q13, V24, and I297 but no correlations were detected between the positions 297 and 51 and only two between the positions 13 and 297.

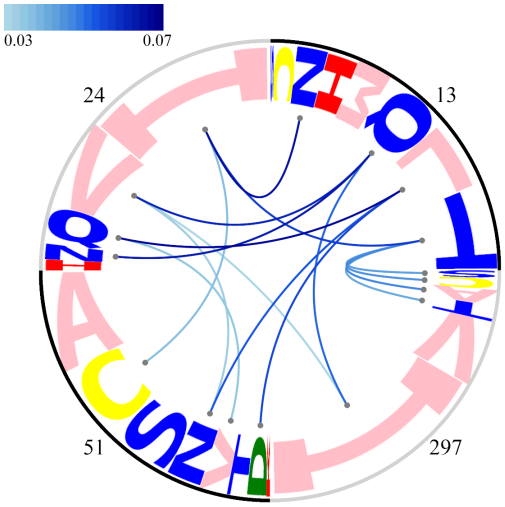


Figure 5.2: **Circular logo for selected Q1 sequences for the backbone relaxed in the wild type contexte.**

Q2 is composed of positions 13, 17, 24, and 297. We used the same protocol as for Q1. For MetRS:ATP, we sampled 3450/3564 sequences at least 1000 times. For the activated states MetRS:[α-Met:ATP]‡ and RS:[β-Met:ATP]‡, we sampled 3168 sequences. Then, we estimated stability, catalytic efficiency, and selectivity in favor of β-Met. We removed variants less stable than the wild type variant. We were left with 97 variants. Sequences with a loss of catalytic efficiency above 2 kcal/mol were removed. Table 5.9 shows the 81 variants selected. The most active variant is `TNII` with two mutations at positions 13 and 24. Figure 5.3 shows correlations between the positions 13, 24, and 51 but none with position 17, which stays in its native type.

Q3 is composed of positions 13, 24, 51, and 297 with the backbone relaxed in the context of the M_{13}, A_{256}, and C_{297} mutant (MAC). For both states MetRS:ATP and RS:[β-Met:ATP]‡, all the sequences were sampled at least 1000 times. Catalytic efficiency and stability are computed with respect to the variant MHDC (used for the relaxation of the backbone).

Tableau 5.8: **Q1 variants filtered with threshold for the stability (≤ 0), the catalytic efficiency (cat. eff. $\Delta G^{\ddagger}$(β-Met) ≤ 0), and selectivity (sel. $\Delta\Delta G^{\ddagger}$($\beta$-Met) ≤ 2).** The reference variant is the wild type LHDI.

seq.	sta.	cat. eff.	sel.	seq.	sta.	cat. eff.	sel.	seq.	sta.	cat. eff.	sel.
TQAI	-0.1	-4.8	-1.4	MICV	-0.3	-1.3	-4.7	TISA	-1.6	-0.4	-1.4
LQCI	-1.1	-4.2	-0.0	MIAV	-1.1	-1.3	-4.7	TISV	-3.8	-0.4	-1.9
LQSI	-0.8	-4.1	0.4	TVDI	-0.2	-1.2	0.0	QNAI	-2.9	-0.4	-4.2
CQAI	-0.0	-4.1	-1.5	QIVV	-0.6	-1.1	-2.3	HIVI	-1.7	-0.3	0.7
QICI	-0.4	-3.9	-3.6	CVNI	-1.0	-1.1	-0.6	HVAV	-0.6	-0.3	-2.3
LQAI	-2.1	-3.9	0.0	TISI	-5.2	-1.1	-1.2	TNNI	-1.5	-0.3	-0.7
QISI	-0.1	-3.8	-3.2	HITI	-0.6	-1.1	0.6	CVDI	-0.2	-0.3	0.0
QIAI	-1.3	-3.8	-3.8	HVCI	-0.9	-1.0	-1.3	NIAV	-2.7	-0.3	-2.2
LQAV	-0.6	-3.6	-0.8	TIAI	-6.2	-0.9	-1.4	QNSI	-1.8	-0.3	-3.5
TQVI	-0.8	-2.9	0.5	LHNI	-0.6	-0.9	-0.5	LVDI	-2.0	-0.3	1.6
HIAI	-0.8	-2.6	-1.5	HVAI	-1.9	-0.9	-1.4	MITI	-2.3	-0.3	-1.8
HICI	-0.1	-2.6	-1.6	TICI	-5.5	-0.9	-1.6	LICI	-7.4	-0.2	-0.3
LINI	-2.1	-2.5	1.1	HVSI	-0.7	-0.9	-0.9	TIAA	-2.8	-0.2	-1.9
QITI	-1.0	-2.4	-1.6	LVNI	-3.0	-0.9	0.9	LIAI	-8.2	-0.2	-0.3
LINV	-0.5	-2.3	0.2	NIAI	-4.1	-0.9	-1.6	NISV	-1.6	-0.2	-1.6
QVSI	-1.0	-2.1	-2.9	SVNI	-0.0	-0.9	-0.7	TICA	-1.9	-0.2	-1.8
QVAI	-2.3	-2.1	-3.5	NICI	-3.3	-0.9	-1.4	NIAT	-0.8	-0.2	-1.5
LIDI	-1.0	-2.1	1.4	LVKI	-0.9	-0.8	1.5	QVVI	-2.9	-0.2	-1.5
CQVI	-0.8	-2.0	0.4	TICC	-1.1	-0.7	-1.9	NICV	-1.9	-0.2	-2.1
QHAI	-0.0	-2.0	-5.2	TIAC	-1.9	-0.7	-1.8	CISI	-5.1	-0.2	-1.2
LQTV	-0.3	-2.0	1.3	NISI	-3.0	-0.7	-0.9	CICI	-5.4	-0.2	-1.6
QVCI	-1.5	-2.0	-3.4	QVTI	-1.9	-0.7	-1.6	MVAI	-3.5	-0.2	-3.8
TVNI	-0.9	-1.9	-0.3	TISC	-0.8	-0.6	-1.4	MVCI	-2.6	-0.1	-3.7
MIAI	-2.5	-1.8	-3.9	TICT	-2.2	-0.5	-2.0	LISI	-7.1	-0.1	0.3
QIVI	-2.0	-1.8	-1.5	TIST	-1.8	-0.5	-1.4	NIAA	-0.7	-0.1	-1.8
MISI	-1.3	-1.8	-3.3	TICV	-4.1	-0.5	-2.6	TIAS	-1.8	-0.0	-1.9
MICI	-1.6	-1.8	-3.8	TIAV	-5.0	-0.5	-2.5	TICS	-0.9	-0.0	-1.8
QVCV	-0.0	-1.5	-4.2	QNCI	-2.0	-0.4	-4.2	MVSI	-2.3	-0.0	-3.2
QVAV	-1.0	-1.5	-4.3	TIAT	-3.0	-0.4	-1.8	LHDI	0.0	0.0	0.0
LQVV	-1.4	-1.4	1.4	LVNV	-1.6	-0.4	0.1				

Once the catalytic efficiency and stability are estimated, we removed the variants with a loss of stability above 3 kcal/mol and a degradation of the catalytic efficiency. 37 variants were left. Since the variant MAC already showed an improvement in selectivity in favor of β-Met, we did not used the selectivity to select variants. Here, we want to improve the activity of variants compared to MAC.

Figure 5.4 shows that positions 13 and 51 are mainly populated with the native types Met and Asp. Only a few correlated position appeared in this analysis, and no correlations with the position 17.

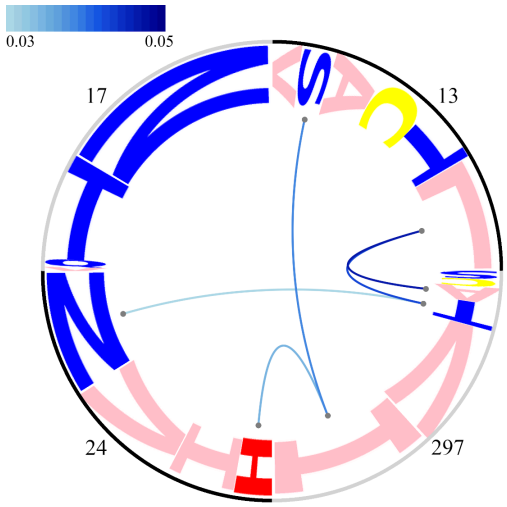


Figure 5.3: **Circular logo for selected Q2 sequences for the backbone relaxed in the wild type contexte.**

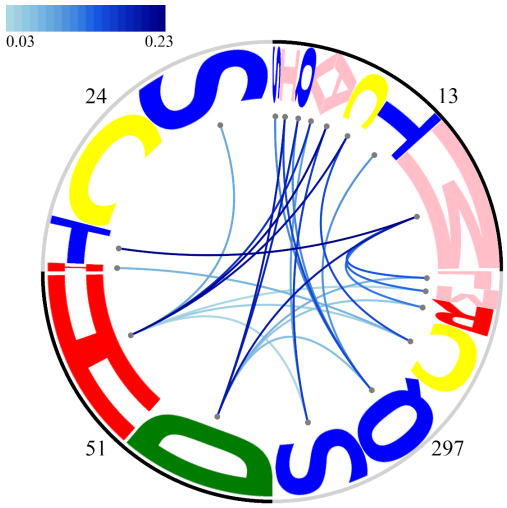


Figure 5.4: **Circular logo for selected Q3 sequences for the backbone relaxed in the MAC contexte.**

Tableau 5.9: **Q2 variants filtered with thresholds for the stability (≤ 3) and the catalytic efficiency (cat. eff. $\Delta G^{\ddagger}$(β-Met) ≤ 3).** The reference variant is the wild type LNIH.

seq.	sta.	cat. eff.	sel.	seq.	sta.	cat. eff.	sel.	seq.	sta.	cat. eff.	sel.
TNII	0.9	-3.0	0.2	CNHI	2.0	0.1	-1.3	VNHI	2.8	1.3	-1.8
TNIV	2.4	-2.5	-0.7	CNVV	1.3	0.1	-0.7	VNVV	2.1	1.3	-1.1
ANII	1.4	-2.2	0.4	ANVV	1.5	0.2	-0.4	LTHI	1.4	1.3	1.0
CNII	1.1	-2.1	0.3	LNVA	1.5	0.2	1.2	SNNI	0.4	1.5	-0.5
LNII	-0.9	-1.9	1.7	ANHI	2.2	0.3	-0.9	LNNC	1.9	1.5	0.9
TTII	2.3	-1.8	1.0	LNVV	-0.8	0.3	0.8	LNNV	-0.9	1.5	0.3
SNII	1.9	-1.8	0.0	SNVV	2.3	0.3	-0.8	CNNT	2.9	1.5	-0.5
LNIV	0.4	-1.6	0.8	TNNI	-0.6	0.3	-0.2	LNNT	1.0	1.5	0.8
ANIV	2.6	-1.4	-0.3	LNVT	1.3	0.3	1.4	ANNV	1.2	1.5	-0.8
CNIV	2.3	-1.4	-0.6	SNHI	2.9	0.4	-1.5	CNNV	0.9	1.6	-1.2
LNIT	2.4	-1.4	1.4	TTVV	2.7	0.4	0.4	VTVI	2.2	1.7	0.3
TNVI	-0.2	-1.4	0.0	LQNV	2.6	0.5	1.4	TTNV	2.3	1.8	-0.3
LNIA	2.6	-1.3	1.4	CTVI	1.4	0.6	0.9	LNNA	1.1	1.8	0.8
VNII	1.9	-1.1	-0.4	LNHV	1.3	0.7	-0.8	LTHV	2.7	1.8	0.1
ATII	2.6	-1.0	1.1	VNVI	0.7	0.8	-0.2	SNNV	1.8	1.8	-1.4
CTII	2.3	-0.9	1.0	ATVI	1.5	0.8	1.3	CTNI	1.0	2.0	0.4
LQVV	3.0	-0.9	2.0	LNVS	2.5	0.9	1.6	VNNI	0.5	2.1	-0.7
TNVV	1.2	-0.7	-0.7	TNNT	2.7	0.9	-0.4	LNNS	2.2	2.2	1.0
TNHI	1.8	-0.6	-1.4	TNNV	0.9	0.9	-1.0	TVVI	2.7	2.2	1.2
ANVI	0.3	-0.5	0.4	TTNI	1.1	1.0	0.4	ATNI	1.3	2.2	0.8
LTIV	1.7	-0.5	1.6	STVI	2.2	1.0	0.8	CTNV	2.5	2.4	-0.5
TTVI	1.4	-0.3	1.0	CNNI	-0.4	1.0	-0.4	STNI	1.8	2.6	0.5
CNVI	-0.0	-0.3	0.3	LNNI	-2.3	1.0	1.2	LTNV	0.3	2.6	1.0
LNVI	-2.2	-0.2	1.7	LTVV	0.8	1.0	1.5	VNNV	1.9	2.7	-1.5
SNVI	0.8	-0.2	-0.0	ANNI	-0.2	1.2	0.1	LTNT	2.3	2.7	1.8
LNHI	-0.0	0.0	0.0	ATVV	3.0	1.2	0.4	LTNA	2.5	2.8	1.5
LNVC	2.3	0.1	1.5	CTVV	2.7	1.2	0.1	ATNV	2.5	2.9	0.0

Tableau 5.10: **Q3 variants filtered with thresholds for the stability (≤ 3) and the catalytic efficiency (cat. eff. $\Delta G^{\ddagger}$(β-Met) ≤ 3).** The reference variant is the wild type MHDC.

seq.	sta.	$\Delta G^{\ddagger}$(β-Met)	seq.	sta.	$\Delta G^{\ddagger}$(β-Met)	seq.	sta.	$\Delta G^{\ddagger}$(β-Met)
MCHM	1.5	-7.6	TCHS	1.6	-4.0	TTDQ	0.7	-0.5
MSHM	1.2	-7.4	TSHS	1.3	-3.8	MCDL	1.5	-0.4
MTHC	1.8	-7.0	VCHS	1.8	-3.3	MCDR	-1.1	-0.4
MCHC	-0.1	-6.8	CCHS	1.3	-3.1	QCDQ	0.5	-0.3
MSHC	-0.5	-6.5	VSHS	1.4	-3.0	TCDQ	-1.3	-0.3
TSHC	2.0	-5.8	CSHS	1.1	-3.0	MSDL	1.3	-0.3
MTHS	1.2	-5.1	ACHS	1.5	-2.8	ICDQ	0.1	-0.3
MCHS	-0.9	-4.9	ASHS	1.1	-2.5	MSDR	-1.3	-0.2
CCHC	1.9	-4.9	MTDQ	-1.8	-1.2	QSDQ	0.2	-0.2
CSHC	1.7	-4.8	MCDQ	-3.8	-1.0	SCDQ	0.0	-0.1
MSHS	-1.2	-4.6	MSDQ	-4.1	-0.8	TSDQ	-1.7	-0.1
ASHC	1.8	-4.3	MTDR	1.0	-0.7	ISDQ	-0.2	-0.1
						MHDC	-0.0	0.0

5.4 Screening of pairs, formation of β-Val quadruplets

Next, we searched variants active for β-Val. We started from RS:[β-Met:ATP]‡ where we removed the [β-Met:ATP]‡ ligand and replaced it with [β-Val:ATP]‡. MetRS:ATP is the same structure as for β-Met catalytic efficiency estimation. We considered the same set of 19 positions and 87 pairs as before. For the score of pairs, we don't allow the additional mutations at positions 13 and 297. Score of pairs is computed with the $18 \times 18 = 324$ possible pair of types. We scored all the quadruplets of positions, and selected one. We investigated that quadruplet of positions to search for active variants with β-Val.

For the MC simulations of pairs, we used 5.10^6 step simulations for both adaptive and biased simulations. Then, the score was computed with the 10 best pairs of types sampled in each MC procedure. With these catalytic efficiency estimation for pairs, we scored all the 3876 quadruplets. Table 5.11 shows the 20 best quadruplets. We chose to investigate the positions 13, 16, 24, and 51 with a score of -4.77 kcal/mol. Three positions overlap Q1 and Q3 from the β-Met selection. It will allow us to draw some comparisons with β-Met. Finally, the mutation space is assigned according to the side chains sampled in the MC of pairs. We denoted this last selection of positions Q4.

Tableau 5.11: **20 best quadruplets of positions for β-Val with the backbone relaxed in the wild type context.** The stability threshold used is 5 kcal/mol and the average score is computed over the 10 best variants. The mutation space of each positions was assigned by the side chain sampled in the pair MCs. The selected four positions is denoted ∗.

Pos				Score	Space			
a	b	c	d		a'	b'	c'	d'
11	15	16	51	-5.06	ACEDKMSR	IKMQRTWVY	CDNST	CIMLSRTHV
*13	16	24	51	-4.77	ACHKMLNQ STD	CDNST	F	CIHMSTV
11	15	24	253	-4.54	ACDKMQSR TVA	AIKMNQRTW VY	F	ACDKMNSR
11	16	17	51	-4.53	ACEDKMSR	ACDNST	FHKMLNQRW YD	CIHVMST
11	15	51	256	-4.45	ACEDKMSR	RTWVI	VILQRH	DFHMNT
11	15	24	51	-4.44	ACEDKMQS RTV	AIKMNQRTW VY	F	RHL
11	15	51	253	-4.43	ACEDKMSR	IRTWV	RHL	ACDKMNSR
13	16	51	253	-4.43	CDIHKMLN QT	CDNST	CIHMSTV	ASK
13	15	16	51	-4.39	CIHKMLNQ TD	IKMQRTWVY	CDNST	CIHMLSRTV
13	17	24	51	-4.36	ACHKMLNS TD	FHKMLQRWY	F	H
13	15	24	51	-4.34	ACIHKMLN STD	AIKMNQRTW VY	F	HLR
12	16	51	252	-4.29	AS	CDNST	CIHMLSTWV	ACDFHNSTVD
13	16	17	51	-4.26	CDHKMLNQ T	ACDNST	FHKMLNQRW YD	CIHMSTV
11	15	16	24	-4.23	ACDKMQSR	AIKMNQRTVY	CS	F
11	15	16	24	-4.23	TV	AIKMNQRTVY	CS	F
12	16	17	51	-4.22	AS	ACDNST	FHKMLNQRW YD	CIHMLSTWV
13	16	51	297	-4.19	CDHKMLNQ T	CDNST	CIHMSTV	RYH
13	16	51	256	-4.17	CDHKMLNQ T	DCTSN	CIHMQSRTV	AEDFHMNQSD
11	15	17	51	-4.13	ACEDKMSR	IRTWV	FHKMLQRWY	RHL
11	15	17	24	-4.09	ACDKMQSR TV	AIKMNQRTVY	FKLQRW	F
11	15	24	256	-4.09	ACDKMQSR TV	AIKMNQRTVY	F	DHMNT

Tableau 5.12: **Q4 mutation space assigned for the RS:[β-Val:ATP]‡** selected positions

	Positions	Initial types	Mutation space
Q4	13	LEU	ALA CYS LYS MET LEU ASN GLN SER THR ASP HIS
	16	ALA	ALA CYS ASN SER THR ASP
	24	HIS	HIS PHE TYR
	51	ASP	CYS ILE LEU HIS MET SER THR VAL ASP

5.5 Design of β-Val quadruplets

Q4 is composed of 1782 sequences. All the sequences were sampled at least 1000 times in both states. However, no variant with aromatic mutation were sampled for position 24 with the wild type complex MetRS:[α-Met:ATP]‡. 661 variants sampled have a stability and catalytic efficiency improved compared with the wild type variant (figure 5.5). 301 of them were sampled with the complex MetRS:[α-Met:ATP]‡. Table 5.13 shows the catalytic efficiency, stability, and selectivity in favor of β-Val for the 120 best variants according to the predicted activity. The best variant is MTYV with a catalytic efficiency predicted at -8.5 kcal/mol while the most active variant also sampled in the wild type system is CTHV.

Figure 5.5 shows that position 13 is populated with hydrophobic side chains. Position 16 is mainly mutated from Ala into Thr or Cys. Position 24 is mutated from His into Phe. Position 51 is mutated from Asp into non-polar side chains (Val, Ile). Positions 13 and 51 don't seem to be correlated in this analysis. However, positions 16 and 24 have some correlations between the types Cys_{16} and Ala_{16} with Phe_{24}. The wild type variant is not among the variants predicted as highly active.

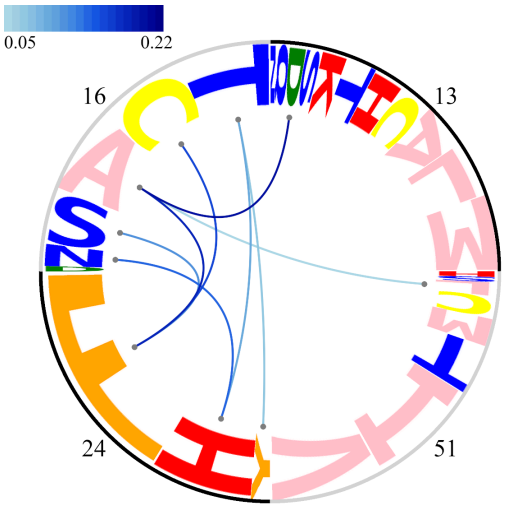


Figure 5.5: **Circular logo for selected Q4 sequences for the backbone relaxed in the wild type contexte.**

Tableau 5.13: **Q4 120 best variants filtered with threshold for the stability (≤ 0), the catalytic efficiency (cat. eff. $\Delta G^{\ddagger}$(β-Met) ≤ 0), and selectivity (sel. $\Delta\Delta G^{\ddagger}$($\beta$-Met) ≤ 0).** The reference variant is the wild type LAHD. Here, the wild type variant is not among the 120 best variants.

seq.	sta.	cat. eff.	sel.	seq.	sta.	cat. eff.	sel.	seq.	sta.	cat. eff.	sel.
MTYV	-0.6	-8.5	ND	DCFI	-0.2	-7.1	ND	SCFI	-1.3	-5.8	ND
LTYV	-0.0	-8.5	ND	TNHI	-0.3	-6.8	-34.1	MSFV	-2.5	-5.7	ND
MTYI	-0.9	-8.4	ND	HCFV	-0.0	-6.6	ND	LSFT	-0.6	-5.7	ND
CTHV	-1.1	-8.4	-21.9	HCFI	-0.2	-6.6	ND	HAFT	-0.5	-5.7	ND
CTHI	-1.4	-8.4	-28.4	TTHM	-0.3	-6.5	-20.6	LSFV	-1.9	-5.7	ND
HTHI	-0.4	-8.3	-30.6	TNHV	-0.1	-6.5	-25.1	ASFV	-1.5	-5.7	ND
LTYI	-0.5	-8.3	ND	CCFI	-1.6	-6.4	ND	QCFI	-1.0	-5.6	ND
HTHV	-0.3	-8.3	-24.2	HSFV	-0.2	-6.4	ND	HAFI	-2.2	-5.6	ND
ATYI	-0.1	-8.2	ND	CCFT	-0.1	-6.4	ND	QCFV	-0.7	-5.6	ND
LTHV	-1.9	-8.2	-19.1	CCFV	-1.3	-6.4	ND	LSFI	-2.2	-5.6	ND
MTHI	-0.4	-8.2	-26.3	DAFV	-1.9	-6.2	ND	ASFT	-0.3	-5.6	ND
KTHI	-0.2	-8.2	-31.6	DAFT	-0.7	-6.2	ND	MSFI	-2.9	-5.6	ND
MTHT	-1.4	-8.2	-25.7	KCFV	-0.5	-6.2	ND	MCFS	-0.5	-5.5	ND
MTHV	-2.6	-8.1	-27.6	LCFV	-2.2	-6.2	ND	MSFC	-0.5	-5.5	ND
LTHT	-0.8	-8.1	-17.4	MCFI	-3.0	-6.2	ND	HAFC	-0.8	-5.4	ND
ATHT	-0.5	-8.0	-20.5	DAFI	-2.1	-6.2	ND	ASFI	-2.0	-5.4	ND
LTHI	-2.4	-8.0	-25.7	CSFV	-0.9	-6.2	ND	DAFM	-0.4	-5.4	ND
ATHV	-1.7	-7.9	-22.2	MCFT	-1.5	-6.2	ND	LCFM	-0.7	-5.4	ND
STHI	-1.0	-7.9	-29.2	MAFH	-0.5	-6.2	ND	CAFV	-3.5	-5.4	ND
STHV	-0.7	-7.8	-22.5	KCFI	-0.7	-6.1	ND	NAFV	-1.0	-5.4	ND
ATHI	-2.0	-7.8	-28.7	CAFL	-0.0	-6.1	ND	TCFV	-1.9	-5.4	ND
MTHC	-0.7	-7.7	-25.6	LCFC	-0.2	-6.1	ND	NAFI	-1.3	-5.3	ND
MNHI	-1.3	-7.6	-41.1	LAFL	-0.8	-6.1	ND	CAFI	-3.8	-5.3	ND
LNHI	-0.7	-7.6	-33.0	LCFT	-1.0	-6.0	ND	TCFT	-0.6	-5.3	ND
MDHV	-0.0	-7.6	-26.1	LCFI	-2.7	-6.0	ND	SSFV	-0.7	-5.3	ND
LTHM	-0.4	-7.6	-19.5	HSFI	-0.7	-6.0	ND	SSFI	-1.1	-5.3	ND
LNHV	-0.4	-7.5	-24.0	ACFV	-1.9	-6.0	ND	CAFT	-2.3	-5.3	ND
QTHV	-0.5	-7.5	-23.7	ACFT	-0.6	-6.0	ND	KAFT	-1.1	-5.3	ND
QTHI	-1.0	-7.5	-30.4	KSFI	-0.1	-6.0	ND	TAFL	-0.4	-5.3	ND
MTHS	-0.3	-7.5	-24.4	MCFV	-2.9	-6.0	ND	KAFV	-2.4	-5.2	ND
MNHV	-1.0	-7.5	-32.4	DAFC	-0.0	-6.0	ND	MCFM	-1.4	-5.2	ND
TTYI	-0.2	-7.5	ND	CSFI	-1.3	-5.9	ND	TCFI	-2.3	-5.2	ND
MDHI	-0.3	-7.4	-33.9	ACFI	-2.3	-5.9	ND	KAFC	-0.4	-5.2	ND
ANHI	-0.5	-7.3	-36.0	KSFV	-0.0	-5.9	ND	QSFV	-0.4	-5.2	ND
TTHI	-1.9	-7.3	-27.0	AAFL	-0.5	-5.9	ND	ACFM	-0.4	-5.2	ND
ANHV	-0.1	-7.3	-27.1	HAFV	-1.7	-5.9	ND	MCYH	-0.5	-5.2	ND
TTHT	-0.5	-7.2	-18.7	MAFL	-1.4	-5.8	ND	LAFT	-3.1	-5.1	ND
MTHM	-1.1	-7.2	-27.5	MCFC	-0.9	-5.8	ND	CAFC	-1.6	-5.1	ND
ATHM	-0.2	-7.2	-22.0	MSFT	-1.0	-5.8	ND	LAFV	-4.4	-5.1	ND
TTHV	-1.6	-7.1	-20.3	SCFV	-1.0	-5.8	ND	MAFV	-5.0	-5.1	ND

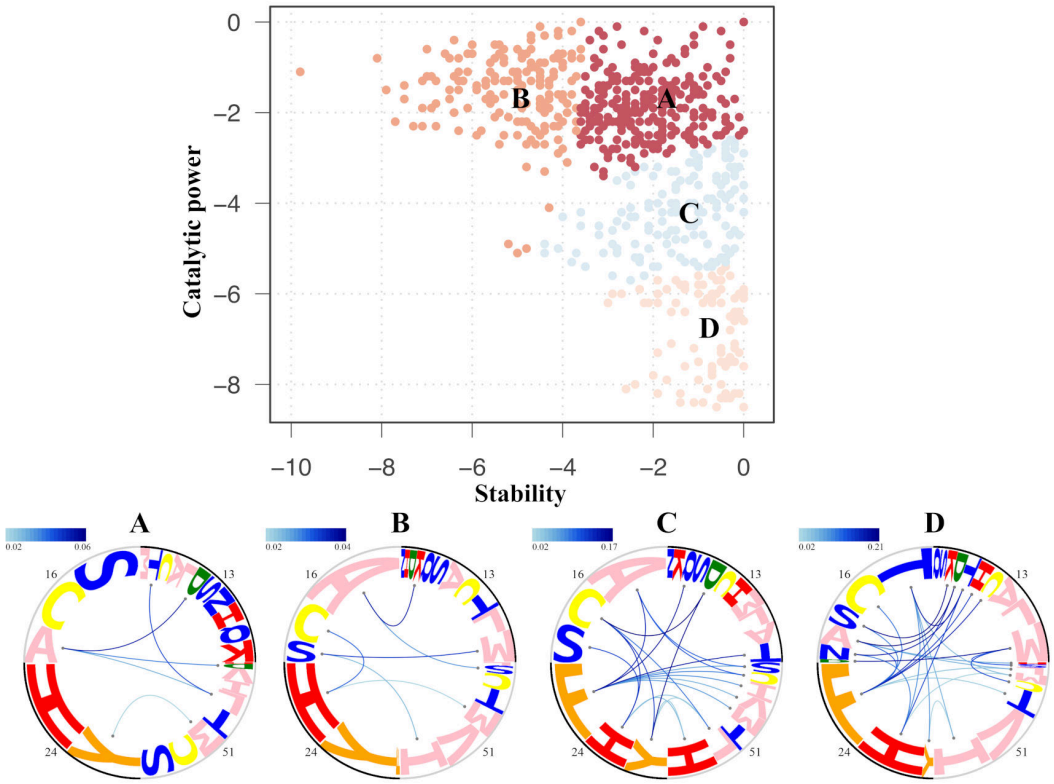


Figure 5.6: **661 variants from Q4 for the activity of β-Val.** In the top are shown the catalytic efficiency and stability of variants grouped into 4 groups A-D. In the bottom are shown the four group logos.

Figure 5.6 shows the distribution of variants with respect to the stability and the catalytic efficiency grouped into four sets A, B, C, and D. For each group, we show the composition and the correlation between positions with a logo representation. Group A is the closest to the wild type sequence LAHD. Variants at position 16 and 24 are populated by homologous side chains ACS and HF. Group B is the most stable. Positions are populated mainly by wild type side chains except for position 51. Group C contains the most active variants compare to the wild type where position 24 is mutated into Phe. This mutation seems to favor native side chains for the position 16 and 13. Finally, group D contains even more active variants. Position 13 is populated by non-polar side chains MLA; 16 is mutated into small, slightly polar side chains TCS; 51 is mutated into small side chains IVTC, and Phe or His populate position 24.

5.6 Concluding discussion

For the search of active MetRS variants for β-Met and β-Val, we first considered the triplet of positions 13, 256, and 297. We applied a selection based on the affinity for β-MetAMP and β-ValAMP. At this point, we cannot conclude on the β-Val prediction power. For β-Met, we produced a set of predictions in which eleven were tested experimentally. Five have a weak but measurable catalytic activity. The structural models used here take into account the strictly APO state and the complex with the reaction product.

Two of six variants validated experimentally showed residence times for β-Met above 60 ns in molecular dynamic simulations. Three predicted variants have reduced slightly the selectivity in favor of α-Met. To improve the β-Met activity, we used the transition state [β-Met:ATP]‡ and the catalytic efficiency. The wild type variant is predicted as the most active variant among the stable ones. The use of native rotamers could explain partially this result. Using a stability threshold of 2 kcal/mol, we recovered eight of the eleven experimentally active variants. The absolute deviation from experiments is 1.9 kcal/mol. These points show that the modelling of the catalytic efficiency is able to discriminate active variants for this system. However, the initial triplet of positions does not yield a higher activity.

The design of active MetRS variants for β amino acids raise other issues. First, we used two structural models where the KMSKS loop is in the active conformation to compute the catalytic efficiency. In addition, one ATP and one Mg^{2+} are modelled in the active site for both states. We assumed that the conformation change, the binding of ATP and Mg^{2+} are identical for all variants. This hypothesis seems reasonable since we were able to reproduce semi-quantitatively the catalytic efficiency of half of the variants. Underestimated variants have the mutation L13M known to deform the backbone. These variants are not well modeled with the rigid backbone approximation.

The ligand pose is another issue for such an application. There is no known structure for the complex MetRS with a β amino acid and ATP. We guessed the β amino acid and ATP poses based on α-Met and β-Met experimental complexes. This assumes that we search for variants where the reaction coordinate is identical ([Crepin et al., 2003, Banik and Nandi, 2010, Zurek et al., 2004]). Designs for β-Met activity showed the same issues as for α-Met. L13M and I297C showed a loss of accuracy for the predictions. However, 8 out of 11 variants were

experimentally active. Therefore, we assume that the hypotheses for the fragments poses are satisfying for this system.

Another computational problem is the choice of parameters for the MC simulations. As for simulated annealing, those parameters are system specific. To converge to a satisfying bias potential, we iteratively optimized the bias with multiple MC simulations. Using a stop condition based on the increment, one can refine the bias potential with a small dependency on parameters.

Another crucial point is the choice of the positions allowed to mutate. For a robust estimation with ALF, one has to produce a sufficient sampling. We empirically determined a visiting threshold of 1000 times. Also, model error may accumulate when a lot of mutations are performed. To reduce the number of positions to investigate, we introduced a selection strategy based on a score of pairs of positions. It represents the contribution of a given pair to the catalytic efficiency. For the search of MetRS variants active with β-Met and β-Val, we restrained the search to groups of four positions simultaneously. To score a quadruplet of positions, we computed the average of pair scores. We chose three quartets for β-Met denoted Q1, Q2, and Q3 and one for the complex β-Val denoted Q4.

From the chosen quartets, we predicted 89 + 81 + 37 active variants for β-Met where we applied different stability and activity thresholds to limit the number of variants to analyze. The predictions showed that mutations to small and slightly polar side chain (Asn, Thr, Ile, and Val) at positions 17 and 24 may improve β-Met activity. Positions 13 and 297 are mainly populated by their native types Leu and Ile. The choice of these positions could be an artefact of the residual mutation space we allowed at those positions. Q3 was constructed with the backbone relaxed in the context of the sequence MAC. It favored the mutation D51H.

For β-Val, we obtained 661 variants where the best catalytic efficiency is estimated at -8 kcal/mol. Since we don't have experimental values to confirm or disprove the prediction, we can't conclude on its performance. However, we observed that H24F is especially favorable for the activation of β-Val in the predictions.

Chapter 6

Design of PDZ pairs with overlapping coding

A gene is encoded by a stretch of DNA. In principle, the same stretch can encode more than one protein: up to six in theory, if all available reading frames are used. The encoding of gene pairs represents a goal in biotechnology. Indeed, the progress in genome engineering raises the question of the confinement of these genes in modified organisms, and overlapping genes have reduced drift. In addition, overlapping coding schemes allow a size reduction, for the design of compact genomes. Also, the evolutionary processes associated with overlapping genes may explain the rapid appearance of new viral proteins by the process called overprinting ([Williams, 1978]).

Here we present an engineering project of pairs of structured proteins with an overlapping coding constraint. We considered five PDZ domains with known structures. We designed pairs of homologous sequences with their DNA sequences completely overlapping. Below, we first recall the results known to date on overlapping genes and the algorithm we used for this work. Indeed, for the production of protein pairs with an overlapping coding scheme, we used a dynamic programming algorithm we developed earlier ([Opuu et al., 2017]), able to derive the overlapping sequences that maximize similarity to a given protein pair. Details of the earlier study are provided in appendix A. Next, we present the production and characterization protocol. We consider here all five overlapping reading frames. To characterize the pairs of produced sequences we used analyses based on similarity, physical and structural properties. Also, we estimated disorder from the Iupred software ([Mészáros et al., 2018]) to determine the

fraction of potentially disordered residues. Finally, we selected three pairs among the designs with satisfying scores for numerical validation by molecular dynamics. One pair had stable structures in a simulation of 500 ns long. Another pair was stable for 3 μs.

6.1 Biological context

Genetic information is stored in the form of sequences of nucleotide triplets, or codons. Each codon is associated with one amino acid. The translation machinery reads the sequence of codons to create the corresponding protein sequence. Within one DNA sequence, there are in fact six possible reading frames (figure 6.1). So, with respect to a first sequence X, there are five overlapping phases in which to encode another protein Y, as shown in figure 6.1. This possibility has been exploited by different areas of Life, especially viruses.

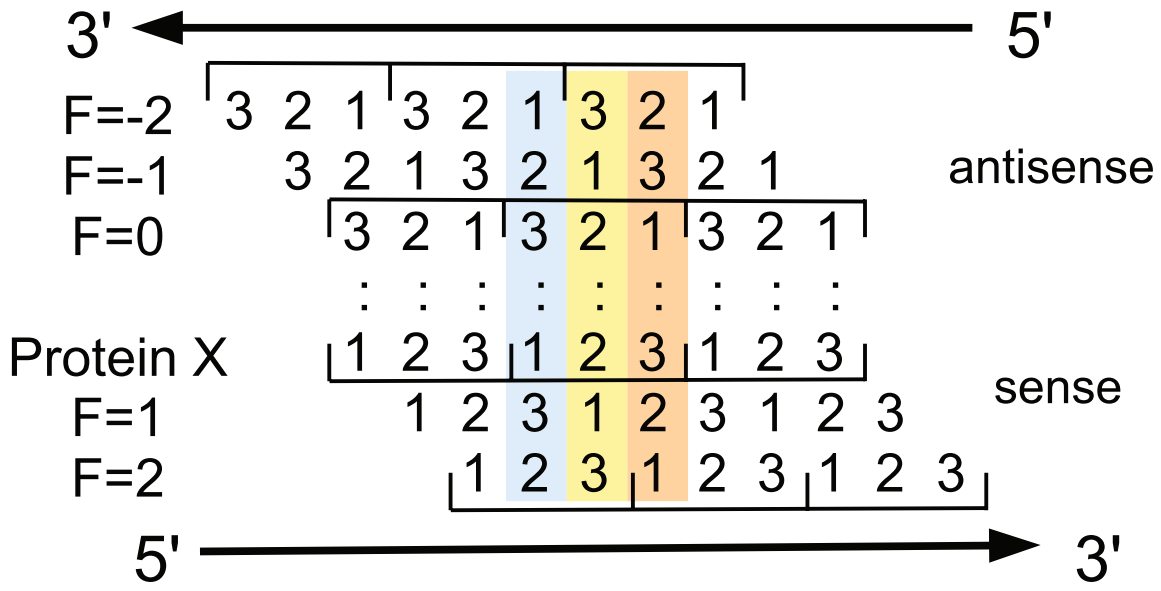


Figure 6.1: **The six possibles reading frames and the five overlapping phases F of protein Y with respect to protein X.**

6.1.1 Natural examples of overlapping codings

There are several examples of overlapping genes in nature and especially in viruses. However, this type of coding scheme has also been detected in Eukaryotes ([Pavesi et al., 2018]). An example of overlapping genes in humans is the anti-oncogene pair INK4a and ARF ([Ouelle et al., 1995]).

These two genes produce proteins involved in aging and cell death. The proteins have 156 and 132 amino acids respectively. The overlapping region is 198 nucleotides long. Relative to INK4a, the ARF gene is encoded in the phase F = 1 (figure 6.1). For the INK4a protein, there are several structures in the PDB (`1A5E`, `1BI7`, `1DC2`, `2A5E`) with a scaffold composed of alpha-helices. For the ARF protein, there is a crystallographic structure (1HN3) from mouse.

A second example in the human genome is the pair of Gs subunit αXL and αAlex ([Abramowitz et al., 20 This example has a large overlapping region of 1884 nucleotides *i.e.* 628 amino acids. The αAlex subunit is encoded in the phase F = 1 relative to the αXL sequence. There is no known structure for theses proteins. For the Alex subunit, there are only certain regions with known structures in the PDB.

A final striking example is the presence of two overlapping regions involving three genes in the SARS-COV-2 virus from the COVID-19 pandemic. The three Uniprot codes associated with these proteins are `NCAP_SARS2`, `ORF9B_SARS2`, `Y14_SARS2`. The protein `NCAP_SARS2` has a known structure. This protein is a component of the viral capsid. The two overlapping pairs are `ORF9B_SARS2/NCAP_SARS2` and `NCAP_SARS2/Y14_SARS2`. Thus, the two proteins `ORF9B_SARS2` and `Y14_SARS2` overlap the capsid protein. The overlapping regions are respectively 96 and 73 amino acids long, and the two non-annotated sequences are completely embedded in the capsid protein-coding region.

6.1.2 Biotechnological applications

Overlapping coding genes have several technical advantages. The first is genome compaction. The compaction of the genome is a factor in viral selection due to the fixed size of the capsids. There is a clear interest in compaction for the production of artificial genomes. In gene therapy, overlapping encoding gives stability to genomes. Indeed, overlapping genes are less sensitive to genetic drift since, in such encoding, a modification of the DNA sequence can introduce a deleterious mutation in two proteins. Therefore, the modified organism is less likely to accumulate mutations. Thus, overlapping genes could be a bio-confinement strategy for the safe use of modified organisms.

One bio-confinement strategy involves the overlap of a gene of interest with a deleterious gene ([Blazejewski et al., 2019]). In such a case, the deleterious gene will kill the organism to which the pair of genes is transferred except under special environmental conditions. For

example, a cytotoxic gene is encoded in one overlapping phase of a given gene of interest. Thus, the organism only survives in a medium containing the anti-toxin. A recent study ([Blazejewski et al., 2019]) showed the effectiveness of this strategy. This study allowed the design of two overlapping protein pairs validated experimentally. The design method consisted of two main steps. The first one used a dynamic programming algorithm allowing the search for initial solutions. This algorithm is equivalent to the one we proposed earlier in [Opuu et al., 2017]. The second step optimized correlations between positions, using a stochastic heuristic. This second step was deemed crucial for the designs produced.

6.1.3 Evolutionary hypothesis

Overlapping encoding could be used for the acquisition of new genes in viruses. The derivation of a new gene from an existing one is known as *overprinting* ([Williams, 1978]). This scenario breaks down into two stages: first, a new gene appears in an overlapping phase of an existing gene and produces a potentially functional protein. Next, a duplication of the overlapping sequences occurs. Eventually, only one gene remains active in each copy. For genes whose coding has remained overlapping, the overlapping encoding may ensure coordination between genes whose functions are linked ([Krakauer, 2000]).

One study ([Kovacs et al., 2010]) showed a correlation between overlapping encoding and disordered proteins. However, this study was limited to 67 human genes with overlapping regions at least 35 amino acids long. A disordered protein or region is defined by its ability to adopt several conformations under physiological conditions. *A contrario*, an ordered structure under physiological conditions folds into a native conformation. Using the IUPRED software ([Mészáros et al., 2018]) for disorder predictions, the study found a correlation with overlapping coding regions.

Another recent study showed that it is possible to encode two aminoacyl-tRNA synthetase homologs on opposite DNA strands ([Martinez-Rodriguez et al., 2015]). This study designed two structured and stable proteins with measured catalytic activity. For this study, no numerical method was used.

6.2 Material and methods

To design pairs of homologous sequences based on selected PDZ domains we use a three-step approach. First, we design solutions guided by sequence similarity. Next, the solutions are characterized by several metrics based on the recognition of sequences or physicochemical properties. Finally, we select a few pairs for numerical validation with long MD simulations. Experimental tests are underway.

6.2.1 Selected proteins

For this study, we have chosen sequences from the PDZ domain family. These are small globular domains of about 90 amino acids. The secondary structure includes 5-6 β sheets and two α helices. It is an interesting test candidate since it has been investigated heavily with numerical studies and has well-known properties. We reported a recent complete redesign study of Tiam1 and CASK ([Mignon et al., 2017, Opuu et al., 2020b]). Those studies were based on PDZ domain three-dimensional structures, but did not include any coding constraint. Here, we focused on five PDZ domains: CASK, DLG2, NHREF, Tiam1, and Grip1. Here, no structural information is involved in the design of the sequence pairs. However, 3D structures will be analyzed in a second phase.

Tableau 6.1: **List of sequences and structures used.** For each structure, we have the size, the protein name, and the isoelectric point calculated by Propka.

PDB	Organism	Size	Name	Pi
1KWA	*Homo sapiens*	88	Cask	10.23
4GVD	*Homo sapiens*	94	Tiam1	5.84
1G9O	*Homo sapiens*	91	NHRF1	6.82
1N7E	*Rattus norvegicus*	95	Grip1	9.44
2BYG	*Homo sapiens*	97	DLG2	8.03

6.2.2 Overlapping pairs design algorithm

Finding an overlapping coding for an arbitrary pair of sequences is usually impossible. Here, for a pair of protein sequences (X, Y) we determine the pair of maximally similar homologs (X', Y') such that (X', Y') can be coded by overlapping sequences. To achieve this goal, we use a dynamic programming algorithm we introduced earlier ([Opuu et al., 2017]).

6.2.2.1 Algorithm specifications

Let (X, Y) be a pair of protein sequences. We first choose the overlapping region in which amino acids will be paired between the two sequences and we choose the phase F ∈ {-2, -1, 0, 1, 2} in which Y will be encoded with respect to X. For this overlapping region, we use the Blosum62 similarity matrix B ([Henikoff and Henikoff, 1992]) to determine the similarity between the pairs of amino acids X_i, X'_i and Y_j, Y'_j. The similarity score of X', Y' then takes the form:

$$S(X', Y') = \sum_i p_i B(X_i, X'_i) + \sum_j q_j B(Y_j, Y'_j) \tag{6.1}$$

We have included weights p_i and q_j depend on the position and reflect the conservation in an alignment. Weights are derived from the entropy H_i of alignment positions: $p_i = exp(-H_i)$. Here, the entropy is based on a reduced alphabet: {LVIMC}, {FYW}, {G}, {ASTP}, {EDNQ} and {KRH}.

First, we consider the phase F = 0 where the codons are in register on opposite strands. This phase is a special case, since X' and Y' codons fully overlap. Maximizing the score corresponds to choosing from 64 possible codons the one that contributes the most to $S(X', Y')$ for each position.

For the other phases, each X' codon overlaps with two Y' codons, so a different method is necessary. First, we consider the phase F = -2. In figure 6.2 A, we represent the nucleotide sequence of the overlapping region for X and Y. We denote $c_X(k)$ and $c_Y(k)$ the codons at position k of the overlapping region. These two codons define the quartet Q_k, a quadruplet of nucleotides. Note that the Q_{k+1} quartet shares its 5' end with the quartet Q_k: $Q_{k+1}(1) = Q_k(4)$. The nucleotide sequence of the overlapping region can be re-expressed as a linked list of quartets (figure 6.2 B).

This reformulation into a linked sequence of quartets allows the use of a dynamic programming approach illustrated in figure 6.3. Let N be the size of the overlapping region. For a position k, one has 256 possible quartets divided into groups of 64 according to the nucleotide at their 3' end: $Q_k(4) \in$ {A, C, G, T}. This last nucleotide defines the quartet state: $\mathcal{S}(Q_k) = Q_k(4)$. Let $s(Q_j)$ be the contribution to S for the couple of codons $c_X(k)$, $c_Y(k)$. At the first overlapping position Q_0, we choose a quartet per group that maximises the scoring function $s(Q_0)$. These four optimal quartets are then stored in the first column of a $4 \times N$ table M. At

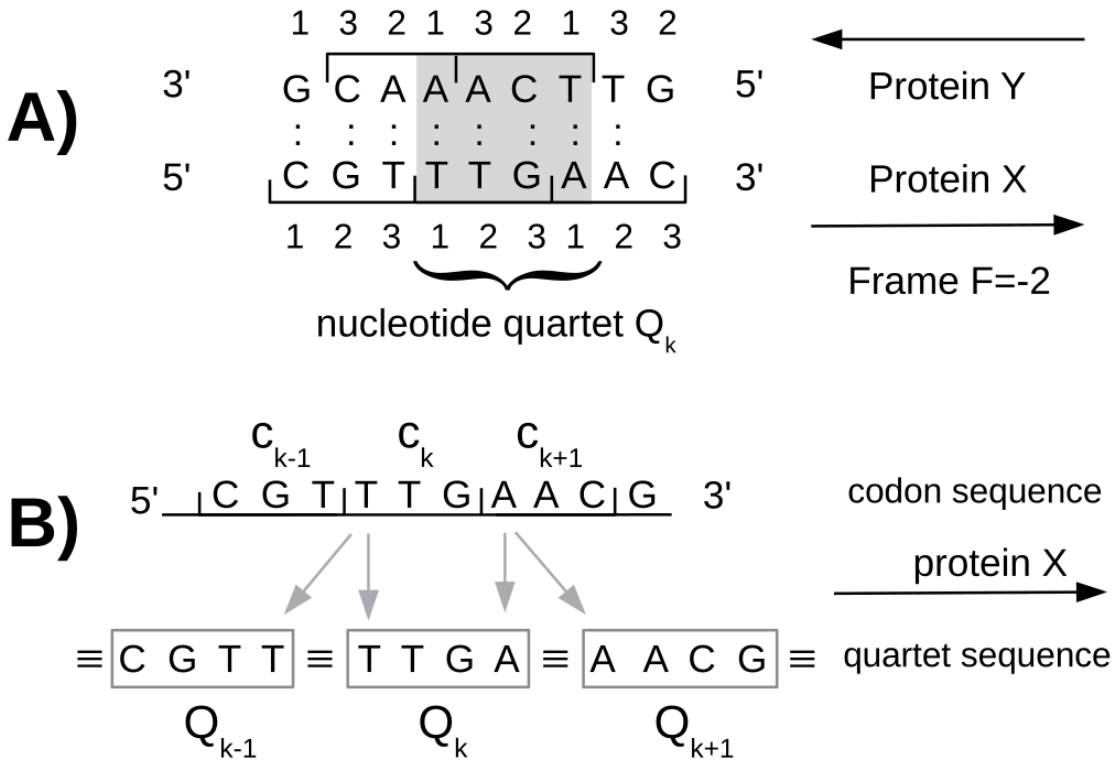


Figure 6.2: **Formulation of the overlapping region of X and Y into a linked quartet list (figure [Opuu et al., 2017]).** A) A segment of DNA with a quartet in gray. B) The same segment represented as a linked sequence of quartets

position k, we assume that the four optimal quartets of the position k-1 are known. For each of the 256 possible quartets, the score is added to the quartet Q_{k-1} if $Q_k(1) = Q_{k-1}(4)$. Thus, we use the following recursion:

$$M(j;\nu) = \max_{Q_j \in \nu} \begin{cases} s(Q_j) + M(j-1;\nu' \equiv Q_j(1)) & \text{if } j > 0 \\ s(Q_j) & \text{if } j = 0 \end{cases} \quad (6.2)$$

Among the 64 quartets whose state is ν ($Q_j(4)$ ν), we choose the one that maximizes the score M(k-1; ν '$\equiv$ $Q_j(1)$) where ν is the state to which Q_j is linked. When the end of the overlapping region is reached, one can carry out standard backtracking to obtain the quartet sequence *i.e.* the DNA sequence encoding X' and Y' whose score is maximum.

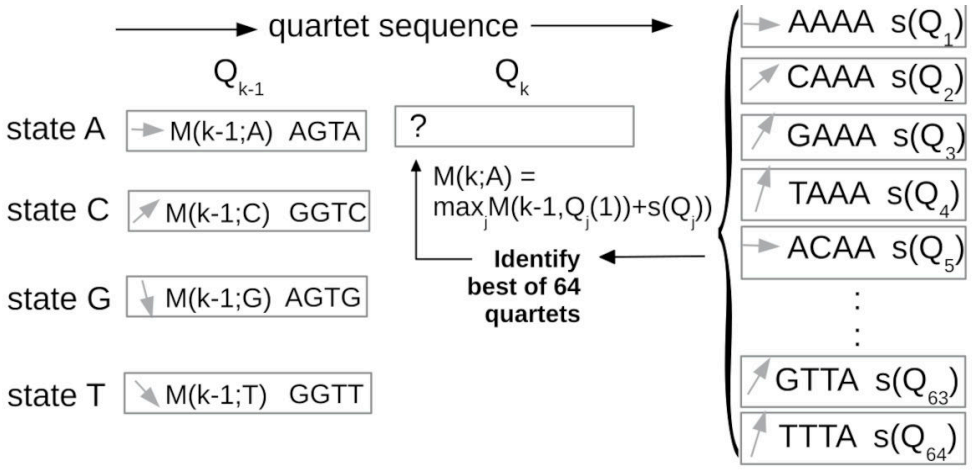


Figure 6.3: **Algorithm for overlapping pair design ([Opuu et al., 2017]).** Representation of the $4 \times N$ dynamic programming table centered on positions k and k-1, two consecutive positions in the sequence of quartets. Each box of a column k-1 represents the optimal quartets for the four states $\nu \in$ {A, C, T, G}. For each optimal quartet Q_{k-1}, we represented by an arrow the pointer to the quartet Q_{k-2} to which it is linked. To choose the optimal quartet for the state ν = A at position k, we choose the one where M(k, ν = A) is the highest among the 64 quartets ending with A.

This approach is generalizable to all the phases F $\in$ {-2, -1, 1, 2}. Indeed, the quartet is the minimum unit of nucleotides to define a pair of overlapping codons ([Lèbre and Gascuel, 2017]). We implemented this method in an iterative way (algorithm 1). The program takes two protein sequences X, Y defining the overlapping region, the size of the region N and the phase F. We produce a DNA sequence where the two sequences X' and Y' are encoded. For this implementation, the complexity is $O(N)$ where N is the size of the overlap.

```
Data: X, Y, N, F
Result: X', Y', DNA
// create an N × 256 table to store the quartets and their scores
table ← create_table(N, 256);
// create an N × 4 table to store the optimal quartets and their scores
find_best ← create_table(N, 4);
// create a list of quartet
l_quartet ← create_quartet_list(N, 256);
// loop in all positions
for i ∈ 1 → N do
    // loop in all possible quartets
    for j ∈ 1 → 256 do
        // the special case of the first position
        if i = 0 then
            table[i,j] ← blosum_score(l_quartet[j], X[i], Y[i]);
        else
            // Get the best quartet of the corresponding state
            mi ← find_best[i − 1, l_quartet[j]];
            // add the current score to the one saved in the table
            table[i,j] ← blosum_score(l_quartet[j], X[i], Y[i]) + table[mi, i-1];
        end
    end
    // Saved the best score for each state in a table
    foreach nuc ∈ A, C, T, G do
        find_best[i, nuc] ← max(table, nuc);
    end
end
// get the best DNA sequence using a backtracking
DNA ← max(table);
// Translate the DNA into two protein sequences
X', Y' ← translation(DNA,);
```

Algorithm 1: Optimization algorithm for the design of X' and Y', two overlapping protein sequences based on X and Y

6.2.2.2 Proof by induction

The algorithm used here provides an exact solution. We use the dynamic programming scheme to deduce a proof by induction. First, we define the concept of an optimal quartet. An optimal quartet $Q_i(S)$ is one of the 64 quartets in state $\nu \in \{A, C, T, G\}$ maximizing the contribution of the score M(i ν) at position i.

1. **Initialization P(1):** For the first position, we choose the four quartets maximizing the score s(Q, ν) for the four different states. Therefore, we obtain four optimal quartets for the first position of the overlapping region. Thus, for the position n = 1, we know the best quartets of the previous position.

2. **Induction hypothesis P(n):** At the position n of the overlapping region, we know the four best quartets of the position each state ν.

3. **Inheritance P(n + 1):** Assuming $P(n)$, we have 4 optimal quartets at position n. The choice of the optimal quartets for the n + 1 position is trivial. Simply connect each of the 256 possible quartets to the corresponding optimal quartet. Since a quartet is only linked to the consecutive positions, it is the only solution to have an optimal quartet. Therefore, we obtain the optimal quartets for the position $n + 1$. We note here that the essential point is to keep the 4 states optimal along the table.

6.2.2.3 The genetic code degeneracy

The coding possibilities depend on the degeneracy of the genetic code. An amino acid may be translated from one or more codons (Table 6.2). In addition, the coding constrains depend on the phase F. Table 6.3 shows the number of pairs of amino acids that can be encoded in an overlapping manner. The reading frame F = -2 is the most flexible and F = -1 is the least.

Tableau 6.2: **Degeneracy of the genetic code: number of codons per amino acid type.**

type	nb codons	type	nb codons	type	nb codons
SER	6	GLY	4	ASP	2
ARG	6	ILE	3	GLU	2
LEU	6	STO	3	PHE	2
THR	4	ASN	2	TYR	2
PRO	4	LYS	2	CYS	2
VAL	4	HIS	2	MET	1
ALA	4	GLN	2	TRP	1

Tableau 6.3: **Number of unique overlapping residue pairs per phase with and without the pairs containing the stop codon.**

phase F	with stop codons	without stop codons
-2	212	196
-1	42	35
0	56	52
1	90	80
2	90	80

6.2.2.4 Production pairs exploiting systematic offsets

We considered five PDZ sequences and 15 sequence pairs. For each pair, we considered the five overlapping phases F. For each triplet (X, Y, F), we considered different offsets between the sequence X and Y as shown in Figure 6.4. We authorized a shift of 10% of the longest sequence in both directions (Figure 6.4). On average, for each triplet (X, Y, F) we produced 21 designs (X ', Y'). In total, we produced 1,715 pairs of designed sequences (X ', Y'). It is important to note the coding symmetries between different overlapping phases. In the phases F $\in$ {-2, -1, 0}, the pairs (X, Y) and (Y, X) are equivalent. In phase F = 1, (X, Y) is equivalent to (Y, X) in phase F = 2. These elements of symmetry simplify the pair production.

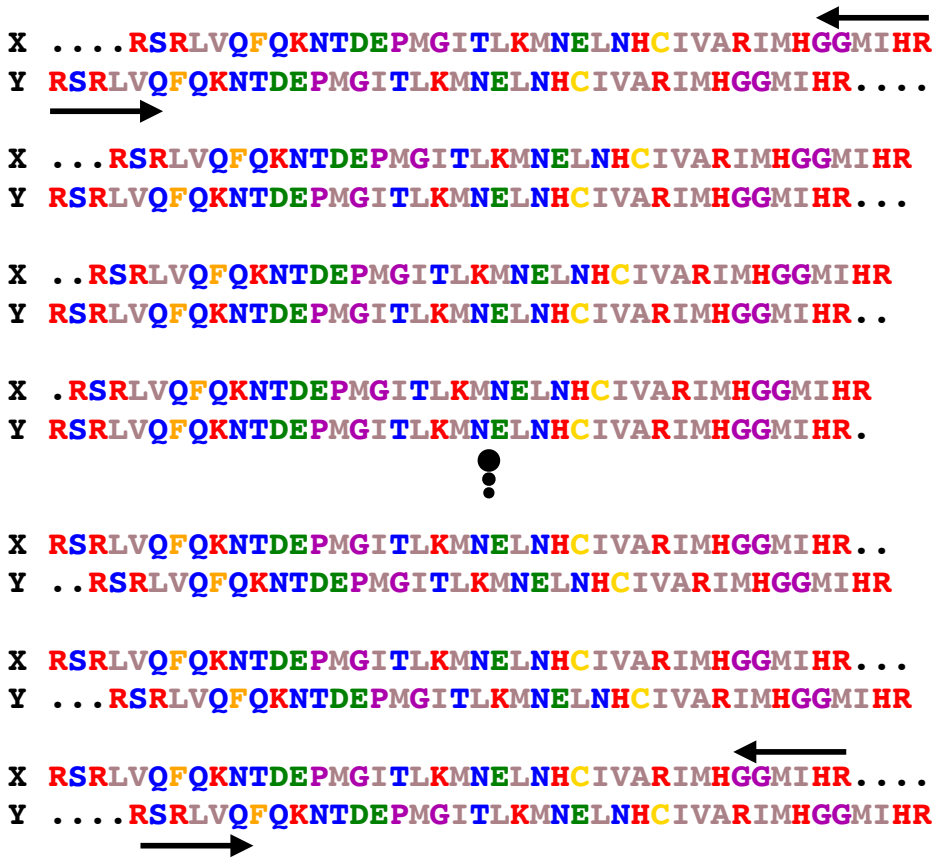


Figure 6.4: **Representation of sequence pair shifts tested.**

6.2.3 Designed sequence characterization

We introduce here a step of manual characterization and filtering. The purpose of this filter is to select a restricted set of sequence pairs to test by molecular dynamics and experimentally.

6.2.3.1 Evolutionary and structural properties

We assume that the pairs of sequences (X', Y') produced here are homologous to the initial sequences (X, Y). One way to support this is to build the inverse relationship with a Blast analysis. Thus, a Blast search starting from X' should find X as the closest sequence.

Optimization of an overlapping region using similarities does not take into account structural information. Therefore, we apply here a negative design filter using the software *Superfamily* ([Gough et al., 2001]). *Superfamily* matches sequences to a SCOP structural family using hidden Markov models. Searches with X', Y' should return the PDZ family.

One experimental issue is protein solubility. To promote soluble sequences, we estimate the isoelectric point (Pi) using the software Propka ([Olsson et al., 2011]). Pi is the pH at which the protein is neutral. If Pi is close to the physiological pH, proteins tend to precipitate. Pi depends on the composition and structure of the protein. Also, we calculated the net charge of each protein. Indeed, charged residues could destabilize folding if there are too many because of electrostatic repulsion.

Finally, mutations creating internal cavities could destabilize folding and are excluded. To detect cavities, we built models using Scwrl ([Krivov et al., 2009]) for all designed pairs (X', Y') using the structures of the initial sequences (X, Y). For each structure, we searched for cavities with the McVol software ([Till and Ullmann, 2009]).

6.2.3.2 Disorder measurement

It has been shown that overlapping coding regions often involve intrinsically disordered protein regions ([Kovacs et al., 2010]). Here, we used the software Iupred ([Mészáros et al., 2018]) to predict disordered regions. We used the protocol established in a previous study ([Kovacs et al., 2010]). A disorder score was assigned to each position of each designed sequence. We applied a threshold of 0.4 to assign the disordered nature of each position and determined for each sequence the percentage of disordered positions.

To assign disorder scores, Iupred uses an energy function based on a coarse-grained model. The score for position k takes the following form:

$$e_i^k = \sum_{1 \leq j \leq 20} M_{ij} C_j \tag{6.3}$$

e_i^k is the disorder score for position i with type k. C_j is the probability to have the types j and j at the neighboring positions. M_{ij} is an energy term obtained empirically. The score is then normalized to be included in the [0, 1] interval. The 0.4 threshold represents the average score by position in the Disprot database ([Hatos et al., 2020]).

6.2.4 Molecular dynamics protocol

The structural stability of the best pairs is estimated by MD. We start by rebuilding the three-dimensional structure of each sequence using the reference structure. Then, the folded structures are simulated in an explicit solvent for at least 500 ns. These simulations allow us to test the global stability as well as local stability of secondary structures. Also, it allows us to compare the behavior of designed sequences with a wild-type PDZ. However, they do not determine whether these structures are capable of folding since we started the simulations with folded structures.

We rebuilt each structure by applying the designed sequence onto the experimental structure with the Scwrl4 software. The protonation state of histidines was predicted using Propka and visual inspection. Then, the structure was solvated in an octahedral water box using the CHARMM GUI ([Jo et al., 2008]). The system was neutralized by adding counter ions (Na^+, Cl^-).

Simulations were done at ambient temperature and pressure with Langevin dynamics and a Nosé Hoover piston ([Martyna et al., 1994, Feller et al., 1995]) , and under periodic boundary condition. The long-range electrostatic interactions were handled by Particle Mesh Ewald ([Becker et al., 2001]).

To analyze simulations, we used the global RMSD to the average MD structure. We removed some highly flexible regions such as the $\beta_2 - \beta_3$ loop. We also calculated the contact lifetimes between positions using their center of mass and a contact threshold of 6.5 Å. For this contact map, we don't consider positions less than four amino acids away in the sequence.

6.2.5 *Ab initio* structure prediction

For the variants validated by MD, we tested folding by an *ab initio* structure prediction from their protein sequence. This test, although less stringent, aims to support the robustness of designs. We used the Robetta server ([Kim et al., 2004]) for these predictions. The server takes as input a sequence in fasta format and provides a set of structural models. The first step in the procedure is to create two libraries of three-dimensional fragments from known proteins. To obtain these fragments Robetta performs a Blast search to identify similar domains. Then, using a Monte Carlo approach, the fragments are assembled into 10,000 structures. Fragment based decoys are then filtered with a coarse-grained energy function ([Bonneau et al., 2002]). Finally, side chain conformations are refined using a Monte Carlo based optimization with the all-atom Rosetta energy function ([Bonneau et al., 2002]).

6.3 Results

6.3.1 Overlapping pair designs

We produced 1,715 homologous sequences pairs with an average of 21 different overlapping regions per pair and phase. Table 6.4 shows the composition of wild-type sequences of the RP55 alignment of the PDZ family in the Pfam database and of the designed sequences. We observe that {LEU, ARG, SER} are over-represented in the designed sequences (column PDZ 6.4) by at least 1% compared to Pfam. The types {ALA, GLY, VAL, ASP} are underrepresented. The over-represented types are coded by 6 codons.

Table 6.5 gives Blast and Superfamily scores for the designed sequences. The Superfamily score allows us to detect sequences that may fold into a non-PDZ structure. The sequences based on the Nhrf1 domain give the best Blast scores. We see that the sequences based on Dlg2 have the highest average log score (-$\log_{10}$(Evalue)) with 17.2 points (Table 6.5). The Tiam1 domain produces sequences on average less reliable with 6.4 points.

The Blast scores confirm that phase F = -2 is the most favorable. For the Superfamily scores, the differences are weaker. Nevertheless, we do find phases F = -1 and 0 as least favorable. However, there are a few isolated sequences in these phases with a score greater than 40 (figure 6.5).

Tableau 6.4: **Amino acid composition of wild-type sequences (Pfam and PDZ) and designed pairs.** (A) Composition of the Pfam RP55 alignment. (B) Average composition of five selected PDZ domains. (C) Average composition of the designed sequences.

type	PFAM[a]	PDZ[b]	Cask[c]	Dlg2[c]	Tiam1[c]	Grip1[c]	Nhrf1[c]
ALA	6.6	4.9	2.4	4.5	6.9	6.4	4.4
ARG	5.3	7.5	10.3	5.2	6.7	6.0	9.4
ASN	4.1	3.9	4.5	4.1	4.4	2.4	4.1
ASP	5.6	3.2	2.5	3.1	4.3	2.8	3.6
CYS	0.9	1.4	1.7	1.0	0.9	1.0	2.6
GLN	3.9	3.3	4.9	2.9	2.0	3.0	3.9
GLU	6.1	4.7	4.4	4.7	4.6	3.8	6.0
GLY	11.6	8.8	7.0	10.0	8.0	10.0	8.9
HIS	2.2	2.6	3.5	2.0	2.3	2.0	3.1
ILE	7.9	6.4	7.9	6.5	4.5	9.0	4.2
LEU	9.8	12.2	11.2	11.7	13.6	12.4	12.2
LYS	5.6	4.7	4.5	5.3	4.3	4.7	4.6
MET	1.7	1.9	3.5	1.6	1.9	1.1	1.5
PHE	2.3	2.1	2.3	3.1	2.7	1.1	1.2
PRO	3.3	4.4	3.6	3.6	3.3	5.4	5.9
SER	6.5	11.9	10.8	11.3	14.7	14.0	8.9
THR	4.8	6.0	6.2	6.5	5.5	7.0	4.6
TRP	0.3	0.4	0.2	0.5	0.4	0.4	0.4
TYR	1.2	2.5	1.3	3.0	3.2	2.5	2.7
VAL	10.1	7.0	7.1	9.4	5.8	4.8	7.9

Tableau 6.5: **Average Blast and Superfamily scores for designed seqeuences.**

	$\log_{10}$(Evalue)			$\log_{10}$(Evalue)	
PDZ	Blast	Superfamily	Phase F	Blast	Superfamily
Cask	24.6	11.2	-2	28.5	14.0
dlg2	25.0	17.2	-1	19.4	9.3
grip1	25.0	12.5	0	23.2	11.3
nhrf1	28.9	15.5	1	27.7	14.3
tiam1	22.8	6.5	2	27.5	14.1

Regarding the similarity to PDZ family members in the PFAM database, similarities are only favorable for the Cask domain. Comparing these similarities to those obtained for the complete redesign of the Cask domain (with no overlapping constraints) ([Opuu et al., 2020b]), the scores are lower (figure 6.6). In addition, we studied the effect of omitting the conservation weights (figure 6.6, curves denoted OG). It appears that the weights slightly, but systematically improve the final similarity. As showed in table 6.6, the scores for the phases F=0 and F=-1 are the least favorable. Note that we calculate the similarity for the entire sequence.

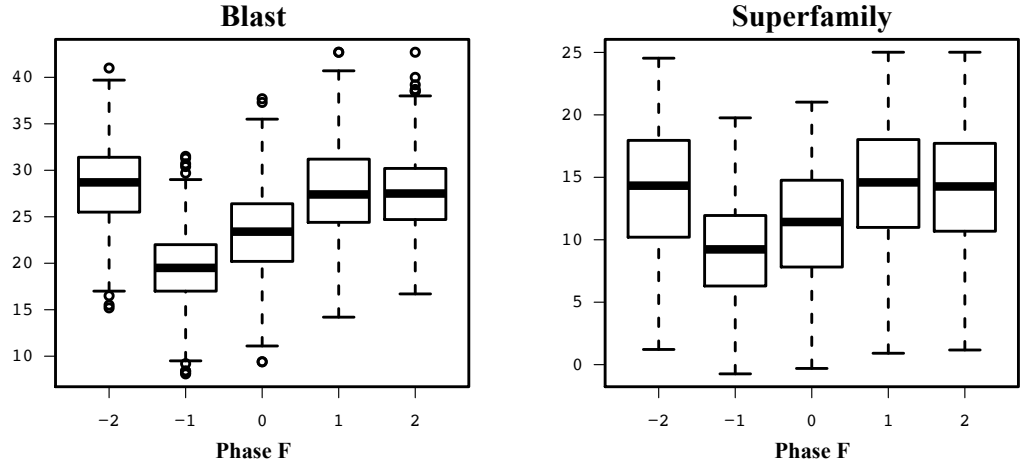


Figure 6.5: **Blast et Superfamily scores per phase F.**

Tableau 6.6: **Pfam similarity scores using Blosum40 substitution matrix.**

PDZ	Pfam score	Phase F	Pfam score
Cask	8.45	-2	-15.70
Dlg2	-14.33	-1	-40.46
Grip1	-34.10	0	-31.67
Nhrf1	-22.55	1	-13.24
Tiam1	-57.38	2	-13.64

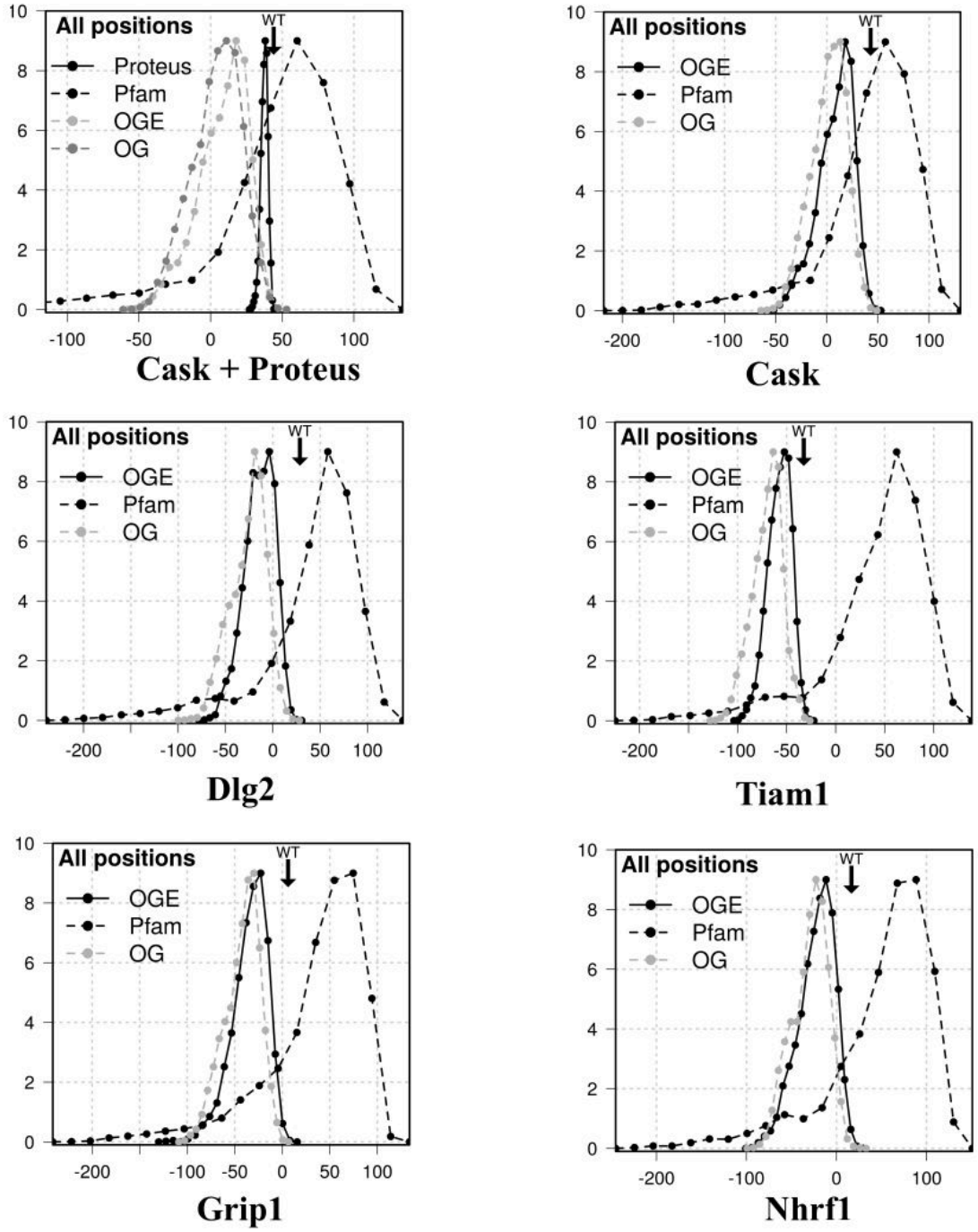


Figure 6.6: **Histograms of Pfam similarity scores per PDZ domain** The curves denoted **Pfam** represent the Pfam vs. Pfam similarity scores. The curves denoted **OG** mean overlapping sequence similarities produced without weighting compared to Pfam. Curves denoted **OGE** mean overlapping sequence similarities produced with weighting compared to Pfam. The curve denoted **Proteus** compares the designed Cask sequences from ([Opuu et al., 2020b]) compare to Pfam.

Finally, we analyzed the tendency for disorder of the designed pairs. We calculated the proportion of positions per sequence to have an Iupred score less than 0.4 ([Kovacs et al., 2010]). The designed sequences are not predicted to be disordered on average. Indeed, over 65% of the designs have at least 80% of non-disordered positions. Moreover, fewer than 3% of designs are predicted to have 90% disordered positions. Figure 6.7 shows that phases F=1 and F=2 seem to produce the most disordered sequences. The Dlg2 domain produced the least disordered sequences.

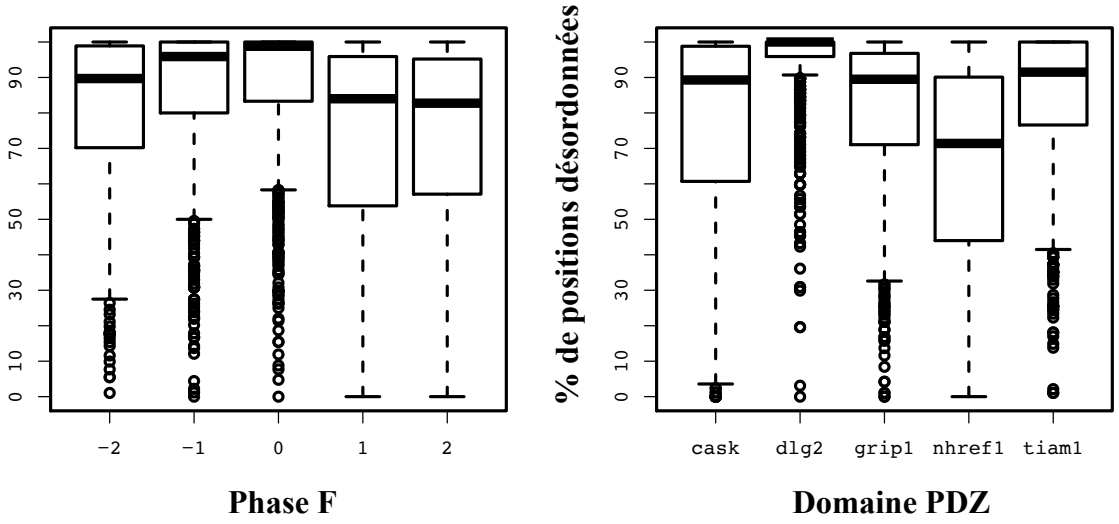


Figure 6.7: **Percent of disorder positions.** A threshold of 0.4 Iupred score is used to determine the disorder propensity.

6.3.2 Pairs selected for MD

For each designed sequence, we constructed three-dimensional models using the experimental structures of the initial PDZ domains and computed various properties. First, we looked at buried cavities. 796 pairs among the 1715 have no cavity. Among the 796 pairs, four had a Pfam similarity score ≥ 25 (see table 6.7). Of these four pairs, we explored the stability of the first three (C1, C2, and C3). These pairs have a low number of drastic mutations (1-4 per sequence). However, the sequence X' in the pair C1 has a Pi very close to the physiological pH (table 6.7).

Tableau 6.7: **Properties of the four selected pairs.** (A) the number of residues for the shift which define the overlapping region. (b) The Blast score on the logarithmic scale. (c) the mean Pfam similarity score. (d) The Superfamily score on the logarithmic scale. (e) The match length detected by Superfamily. (f) The number of drastic mutations. (g) The overall net charge.

	PDZ		shift[a]	phase F	Blast[b]		Pfam[c]		Superfamily[d]	
id	X	Y			X	Y	X	Y	X	Y
C1	Cask	Cask	6	-2	30.0	29.4	31.0	27.3	35.9	33.5
C2	Cask	Cask	0	-2	26.2	25.0	31.2	28.9	35.2	35.4
C3	Cask	Cask	8	-2	24.5	26.2	26.1	34.9	31.1	35.1
C4	Cask	Cask	7	0	24.3	24.7	25.9	28.0	37.6	37.3
	PDZ		match[e]		Pi		Drastic[f]		Charge[g]	
id	X	Y	X	Y	X	Y	X	Y	X	Y
C1	Cask	Cask	79	79	7.4	10.0	1	1	2	4
C2	Cask	Cask	80	80	11.3	11.3	1	1	7	8
C3	Cask	Cask	80	76	12.0	12.0	4	3	8	9
C4	Cask	Cask	80	80	11.4	11.6	4	4	10	10

Figure 6.8 shows the C1 overlapping region with the two Cask sequences encoded in the phase F = -2 with an offset of 6 residues. We denote this pair (X1, Y1). The C-ter ends are not constrained by the overlapping encoding. X1 has 34 mutations relative to Cask; Y1 has 36. Figure 6.9 shows the overlapping region of pair C2=(X2, Y2), two Cask domains encoded in the phase F = -2 with an offset of one residue. X2 has 31 mutations and Y2 has 32. Figure 6.10 shows the overlapping region of pair C3=(X3, Y3) with two Cask domains encoded in the phase F = -2 with an offset of 8 residues. X3 has 28 mutations and Y3 has 27. This pair is the least mutated among the three pairs chosen for MD simulations (figure 6.11).

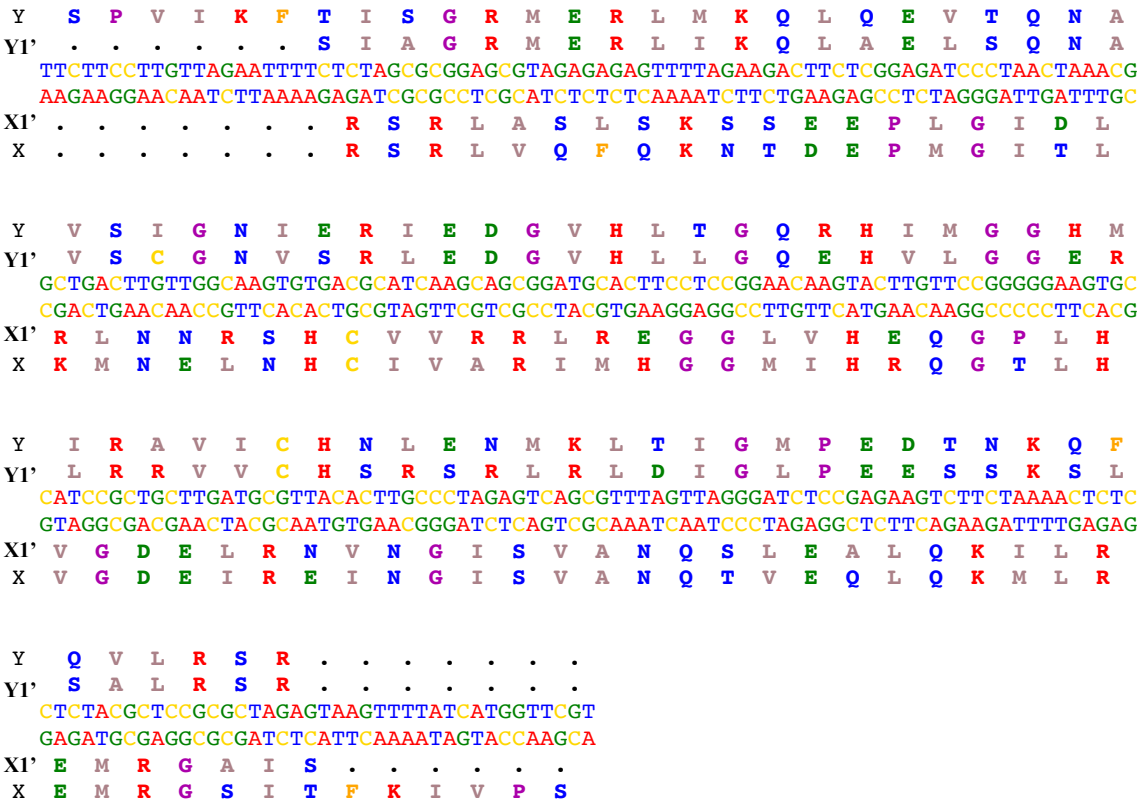


Figure 6.8: **C1 pair of overlapping sequences Cask = X Cask = Y in phase F = -2 with a 6 residues shift.**

Figure 6.9: **C2 pair of overlapping sequences Cask = X Cask = Y in phase F = -2 with a 1 residues shift.**

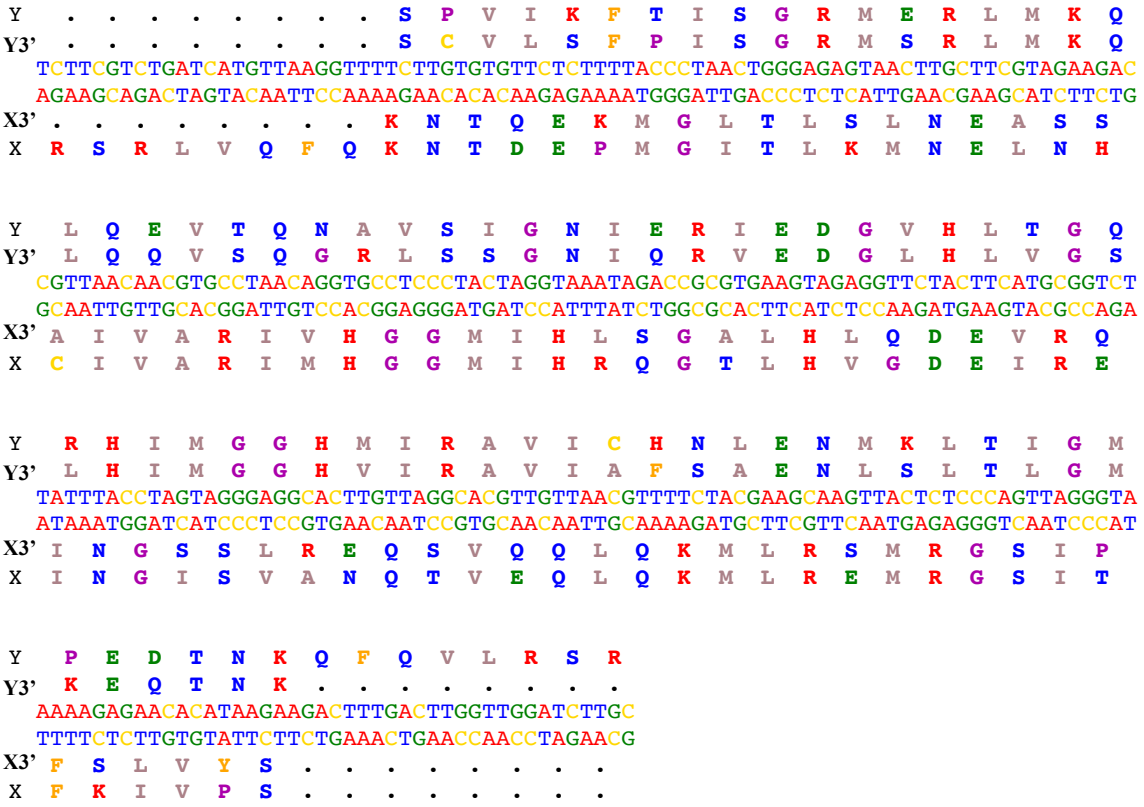


Figure 6.10: **C3 pair of overlapping sequences Cask = X Cask = Y in phase F = -2 with a 8 residues shift.**

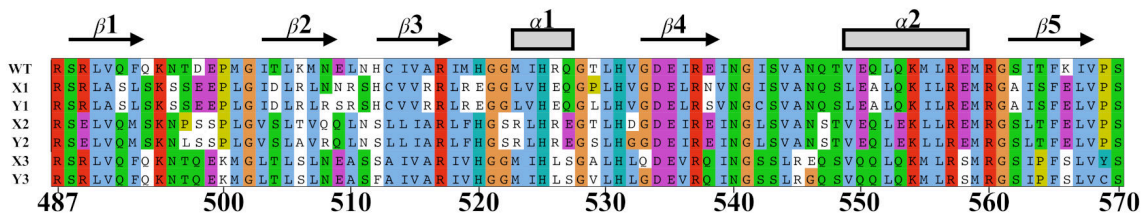


Figure 6.11: **All selected sequence pairs for molecular dynamic simulations.**

6.3.3 Molecular dynamic validation

We selected three overlapping pais (C1, C2, and C3) for molecular dynamic investigations. For the C1 pair, we performed 500 ns simulations. The average structures of both sequences are well reproduced when compared to the WT (figure 6.12, panel A). Next, we computed the RMSD to the average structure found in the simulation. The X1 RMSD evolution during the simulation shows that the sequence is stable with an average RMSD of 1.4 Å (figure 6.12, panel C). The secondary structure analyses performed with DSSP confirmed that X1 reproduced well the PDZ fold. However, Y1 sequence is slightly less stable. As shown in figure 6.12, β_2 and β_3 sheets are slightly degraded. This instability appeared after the first 100 ns of simulation and the RMSD seems to grow after 200 ns of simulation. This RMSD profile may be the signal of the nascent unfolding of Y1 structure.

For C2 pair, both designs showed instabilities in 500 ns simulations. X2 is the most stable design here, as shown by the RMSD (figure 6.13, panel C) and the conservation of secondary structures (figure 6.13, panel A and B). However, the β_2, β_3, and α_2 secondary structures are slightly degraded. For Y2, the structure did not reproduce well the secondary structures, according to DSSP. The frames displayed and the RMSD show that Y2 is more flexible and may unfold soon.

Next, we analyzed C3 with 3 μs simulations, the longest test. The C3 pair is the most stable. As described by the frames sampled during the simulation, X3 and Y3 conformations are close to the average structure (figure 6.14). This observation is supported by the DSSP and RMSD analyses, with an average RMSD around 1.2 Å.

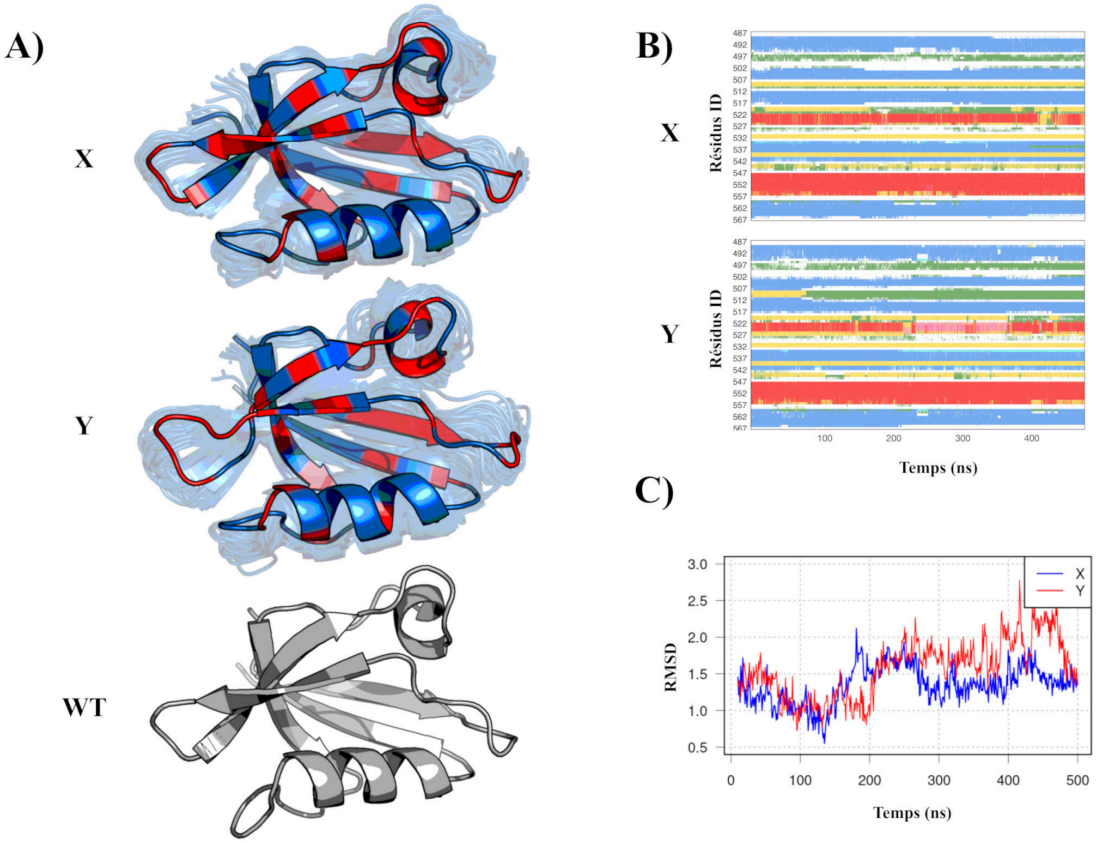


Figure 6.12: **Simulation results for X1 and Y1 sequences of the pair C1.** Panel A shows the average structures of X and Y compared (a few sampled frames are displayed in transparency) to the WT (grey) with the mutated position in red. Panel B shows the secondary structure analyses performed with DSSP. Panel C shows the evolution of the RMSD to the average structure found in the simulation.

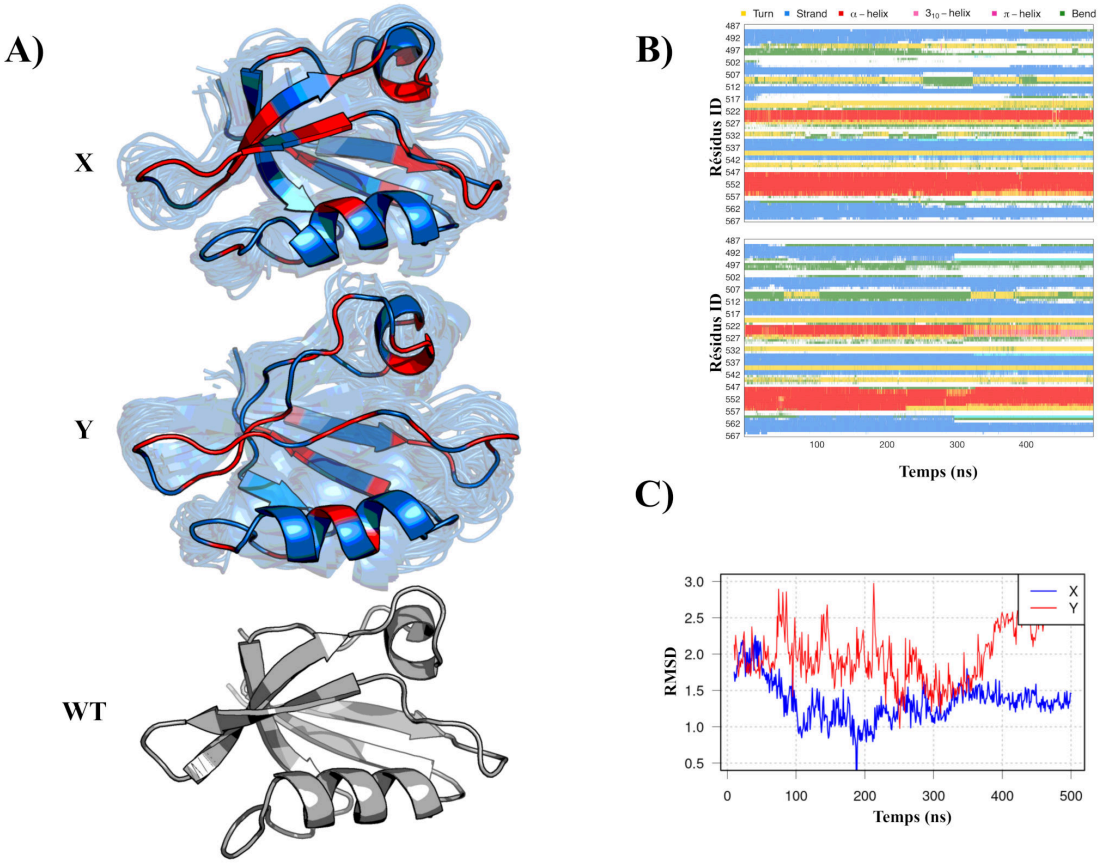


Figure 6.13: **Simulation results for X2 and Y2 sequences of the pair C2.** Panel A shows the average structures of X and Y (a few sampled frames are displayed in transparency) compared to the WT (grey) with the mutated position in red. Panel B shows the secondary structure analyses performed with DSSP. Panel C shows the evolution of the RMSD to the average structure found in the simulation.

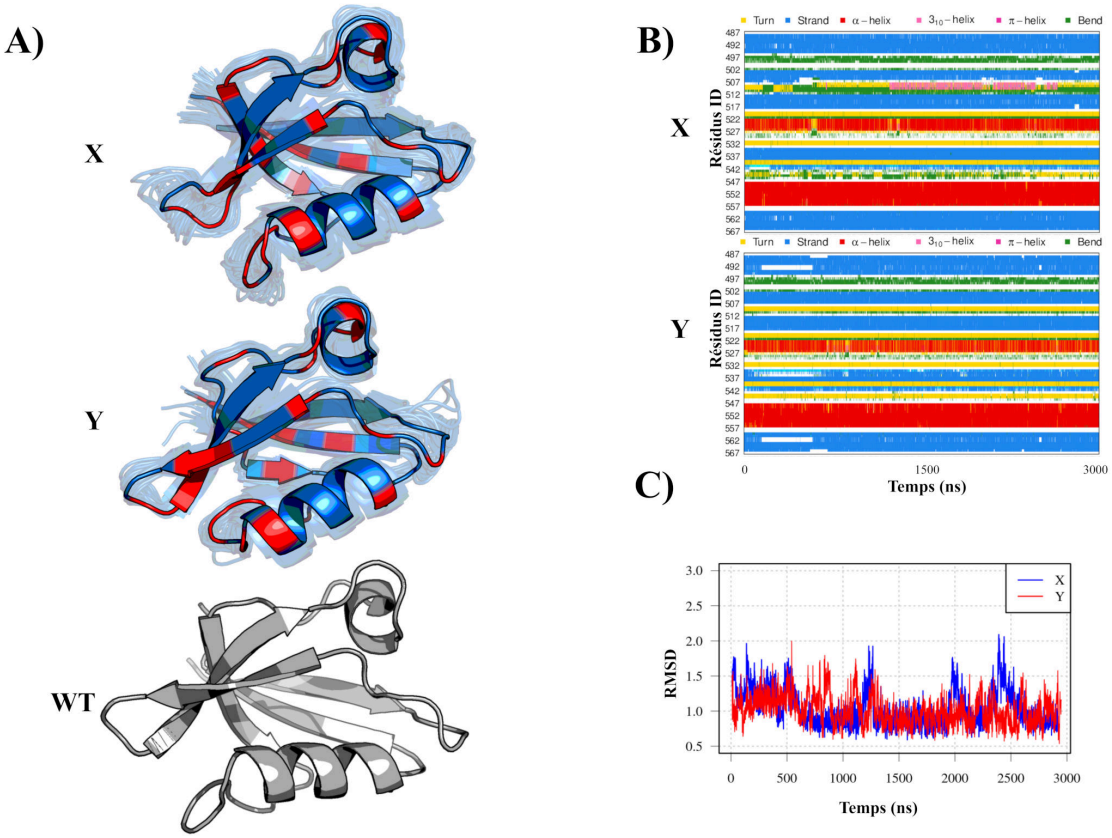


Figure 6.14: **Simulation results for X3 and Y3 sequences of the pair C3.** Panel A shows the average structures of X and Y (a few sampled frames are displayed in transparency) compared to the WT (grey) with the mutated position in red. Panel B shows the secondary structure analyses performed with DSSP. Panel C shows the evolution of the RMSD to the average structure found in the simulation.

For Y2, we extended the analyses to the study of contact lifetimes. The goal of this analysis is to detect missing native contacts which may explain the instabilities observed. First, we computed the contact lifetimes for the WT simulation. Then, we computed the ones from the Y2 simulation. Figure 6.15 shows the log lifetime ratios for Y2 and WT proteins. 503-518 is one striking missing contact (red circle in figure 6.15). When we compare both average structures, there is a shift in the aligned structures (left, red circle in figure 6.15). The mutation M519F may destabilize Y2 although the mutation cost in Blsoum62 score is only 0.

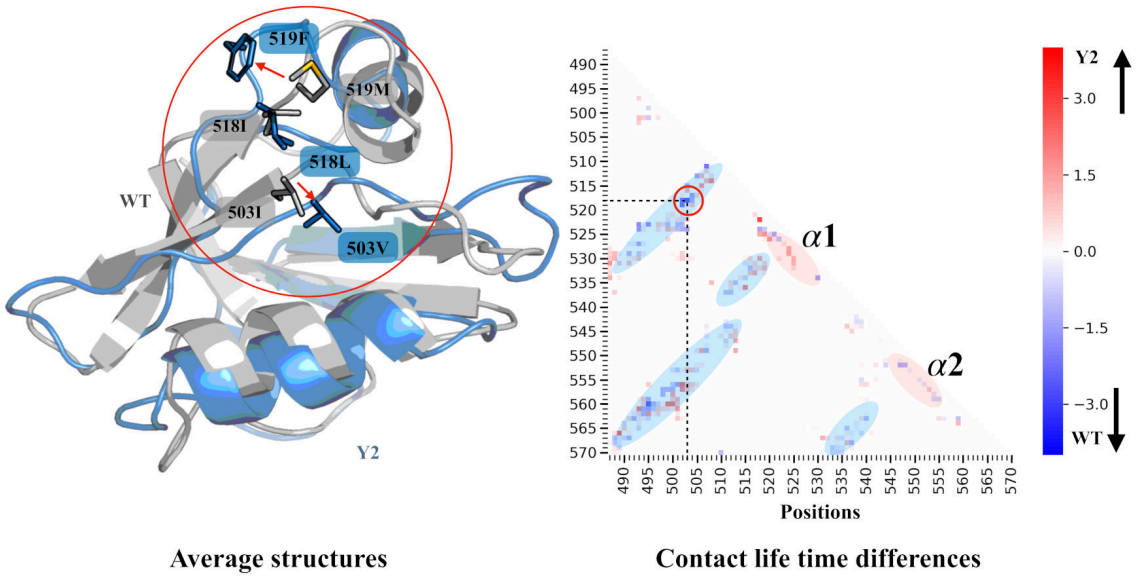


Figure 6.15: **Comparison of lifetime contacts between the wild type and Y2.** On the left are average structures alignment with Y2 in blue and WT in grey. On the right side is the contact lifetime differences between WT and Y2. Red values indicate longer contacts for Y2. Blue values indicate longer contacts for the WT.

6.3.4 *Ab initio* structure prediction for the C3 pair

We performed another test to support the designed pair C3. We made an *ab initio* structure prediction using the Robetta server ([Kim et al., 2004]). As shown in figure 6.16, the two sequences processed by Robetta were folded into a PDZ type conformation. We can recognize the secondary structures of PDZ family folds, the five β sheets, and the two α helices.

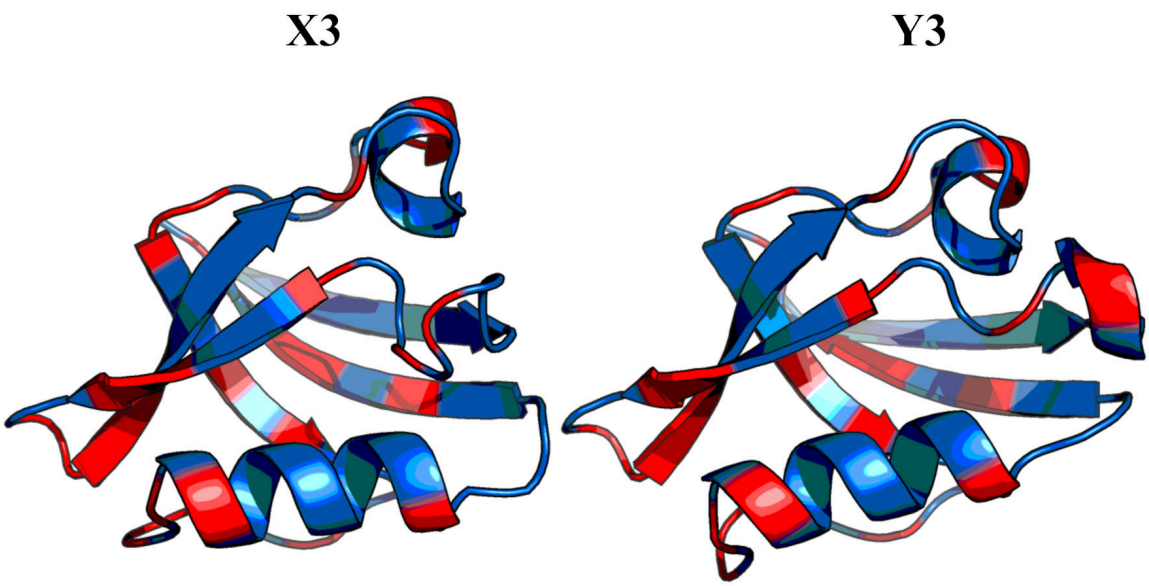


Figure 6.16: **Results of structure prediction using the *ab initio* method implemented in Robetta server for the C3 pair.**

6.4 Concluding discussions

To create compact genomes and bio-confinement strategies, we explored overlapping coding. Indeed, this type of coding helps contain mutations, since a mutation in an overlapping region results in two protein mutations. Nevertheless, finding an overlapping scheme for an arbitrary pair of sequences (X, Y) is in general impossible. The approach we used here allows us to produce a pair X', Y' homologous to X and Y'.

We generated 1715 pairs of overlapping coded sequences based on five PDZ domains. From a compositional point of view, the sequences produced are slightly enriched in LEU, ARG, and SER types (from 1% to 5% enrichment) compared to PFAM. This enrichment can be explained by the code degeneracy for these types. Indeed, these types are the most degenerate with 6 codons.

The effect of composition bias appears to be systematic for overlapping genes ([Rancurel et al., 2009]). We do not observe an over-representation of disordered regions in the designed sequences. Therefore, we conclude that the correlation between overlapping regions and disorder is not systematic. The disorder of naturally occurring overlapping genes observed in the human genome may not be due to the overlap constraint.

The dynamic programming algorithm described here allows us to produce homologous pairs of proteins with an overlapping encoding scheme. To validate the homology relationship between X $\rightarrow$ X' (Y $\rightarrow$ Y'), we used Blast. Indeed, using the same threshold of $\log_{10}$(Evalue) = 10 as in previous work ([Opuu et al., 2017]), we have 1593 out of 1715 pairs of homologous to their reference sequence.

We confirmed some results of the previous work ([Opuu et al., 2017]) especially in terms of phase performance. Indeed, the pairs produced in phase F = -2 have more favorable scores. *A contrario*, the pairs produced with phases 0 or -1 have the lowest scores. This difference can be explained by the flexibility of the coding in phase F = -2. Indeed, it is possible to overlap 196 pairs of different amino acids while the phase F = 0 allows only 52 pairs.

We show here that it is possible to design a structured protein pair in an overlapping coding scheme. Among the 1715 designed pairs, we selected three where Blast, Superfamily, and Pfam similarities scores were high. For these three pairs, we produced MD trajectories. The C2 pair shows partial instability and the β-2, and β-3 sheets vanished for one sequence. The pairs C1 was stable over trajectory of 500 ns. The pair C3 was stable over trajectories of 3 μs. Furthermore, we have shown that the *ab initio* structure prediction recovered the PDZ fold for C3.

C2 instability could be a result of a design method using only a position-based score. Indeed, a recent study ([Blazejewski et al., 2019]) shows that the addition of the correlations between positions is beneficial. This approach could take into account collisions between distant positions in the sequence. Besides, the lifetime of contacts between positions has shown that although the cost of some mutations is weak from an evolutionary point of view (blosum score), the interaction network can be greatly affected.

To take into account the correlations between positions, one can add to the cost function a weight in the form of the logarithm of the joint probabilities taken from an alignment. However, this approach assumes that we have an alignment of sequences rich enough to describe the correlations in a relevant way. A more satisfactory approach is to use structural information. We could thus apply the approach described in ([Blazejewski et al., 2019]) for the optimization of correlations using a *physics-based* energy function. At this point, we also looking forward to experimental tests for the pairs C1 and C3 (G. Travé, personal communication).

6.5 Design perspectives

We introduce several perspectives for the dynamic programming algorithm. First, we discuss the use of quintuplets instead of nucleotide quartets. This reformulation makes it possible to use the full potential of the overlapping encoding, so that one can encode up to six sequences in a single DNA section. Then, we discuss the generalization of the algorithm for overlapping designs by deriving the dynamic programming scheme used for pairwise sequence alignment. This generalization allows a more flexible optimization but also allows insertions and deletions. It constitutes a path to whole genome compaction.

6.5.1 Quintuplets of nucleotides

In this work, we used a formulation in sequences of linked quadruplets, or quartets. In fact, this formulation is much more flexible. It allows the encoding of up to four sequences on the same overlapping region. Moreover, this generalization was used in the previous study ([Opuu et al., 2017]) for the design of triplets where three sequences are embedded in the same DNA section. For those designs, we used the following scoring function:

$$S(X', Y', Z') = \sum p_i B(X_i, X'_i) + \sum q_j B(Y_j, Y'_j) + \sum t_k B(Z_k, Z'_k) \tag{6.4}$$

Similarly, we introduce a sequence of linked quintuplets, in order to use up to six reading frames (figure 6.17). Indeed, as shown in Figure 6.17, each quintuplet potentially includes six different codons. Each quintuplet is linked by the pairs of nucleotides at each end. The only change for this generalization is the number of optimal quintuplets. Now, for each column of the dynamic programming table, we have 16 states instead of 4, representing the 16 possible pairs of end nucleotides.

To illustrate this generalization, we have tried to encode six times the Cask sequence in the same coding region (Figure 6.18). For the six reading frames X, Y, Z, U, V, and W, we obtain respectively the similarity scores 119, 86, 118, 88, 97, and 114. For the identity percent, we obtain respectively 32.1, 27.3, 32.1, 28.5, 28.5 and 30.9 %. These identity scores are on average half of those obtained by dual coding.

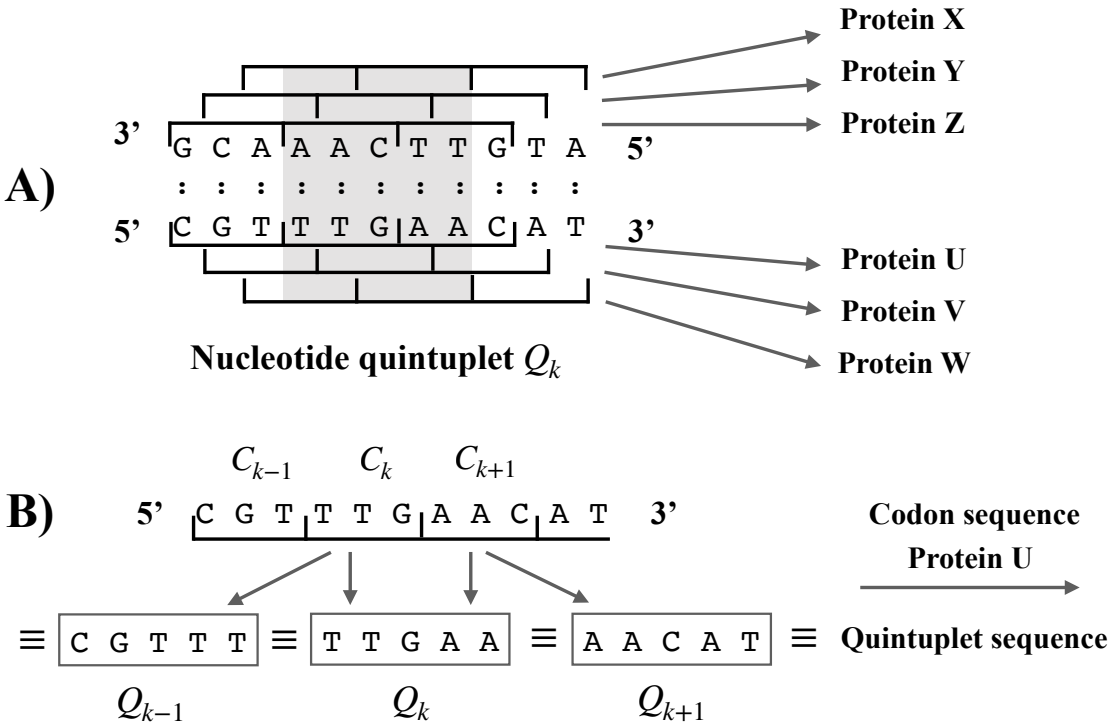


Figure 6.17: **Reformulation of the overlapping region of six protein sequences into a linked list of quintuplets.** A) A DNA segment with a shaded quintuplet. B) The same segment represented as a linked sequence of quintuplets

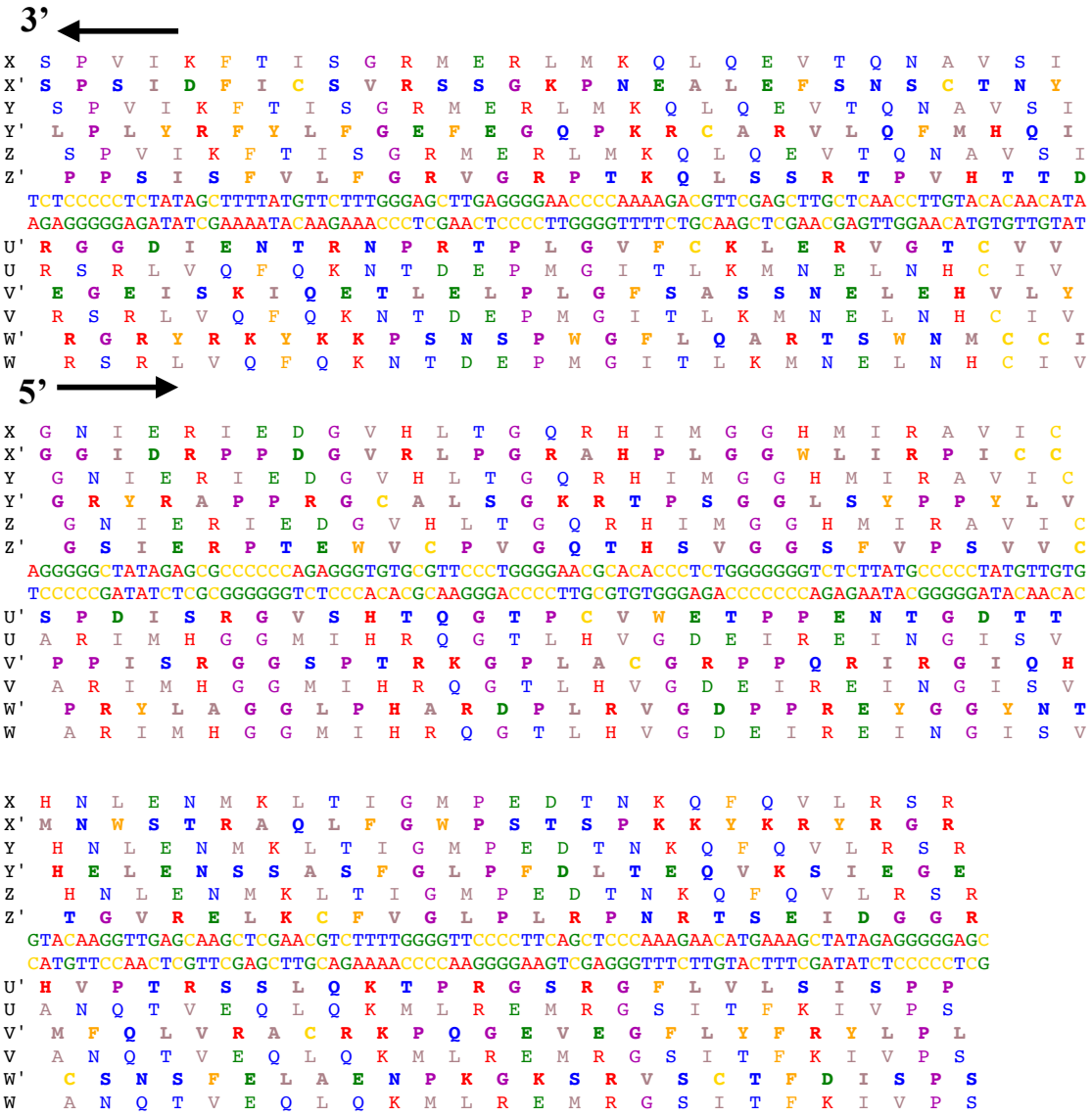


Figure 6.18: **Overlapping encoding of six PDZ Cask domains denoted X, Y, Z, U, V, and W**

6.5.2 From Smith & Watermann to overlapping designs

One problem encountered when designing pairs is the optimal choice of the overlapping region for a pair of sequences. The choice of this region has so far been either arbitrary or based on exhaustive explorations of the shift between the two sequences. We propose here an approach based on the local sequence alignment algorithm illustrated in figure 6.19.

Let X and Y be a pair of protein sequences of size N and P. Let M be a $N \times P$ dynamic programming table. For each pair of positions i, j of X and Y, we have a vector of four elements defining the four possible states {A, C, T, G}. For the first pair of positions i = 0 and j = 0, we choose an optimal quartet per group among the 64 possible. For other pairs of positions i and j, we assume that the optimal quartets of cell M(i-1; j-1), M(i-1; j) and M(i; j-1) are known. We add the score of 256 quartets Q_{ij} to the three optimal quartets of the previous boxes to choose the four optimal quartets of the positions i, j. Thus, we have the following dynamic programming scheme:

$$M(j;i;\nu) = \underset{Q_{ij}\in\nu}{max} \begin{cases} s(Q_j) + M(i-1;j-1;\nu' \equiv Q_{ij}(1)), \\ M(i;j-1;\nu' \equiv Q_{ij}(1)) - W, \\ M(i-1;j;\nu' \equiv Q_{ij}(1)) - W, \\ 0 \end{cases} \quad \text{if } i > 1, j > 1 \quad (6.5)$$

W is the gap penalty. When the dynamic programming table is completed, we perform a backtracking from the quartet with the highest score in the table and finish when the score is zero. We notice that if we use the same reading frame, we fall back on an algorithm similar to the local alignment established by Smith and Watermann ([Smith and Waterman, 1981]). Thus, it is possible to introduce modifications such as insertion or deletion which may represent in some cases non-coding introns.

To illustrate this approach, we searched for the largest overlapping regions in a set of natural sequence pairs. For a given sequence pair, we are looking for the largest pair of sub-sequences that can be encoded in an overlapping fashion. First, we created a database of 500 representative domains chosen randomly from SCOP superfamilies ([Conte, 2000]). The sequence sizes range from 19 to 690 amino acids with an average of 164 amino acids. To guide the search for the

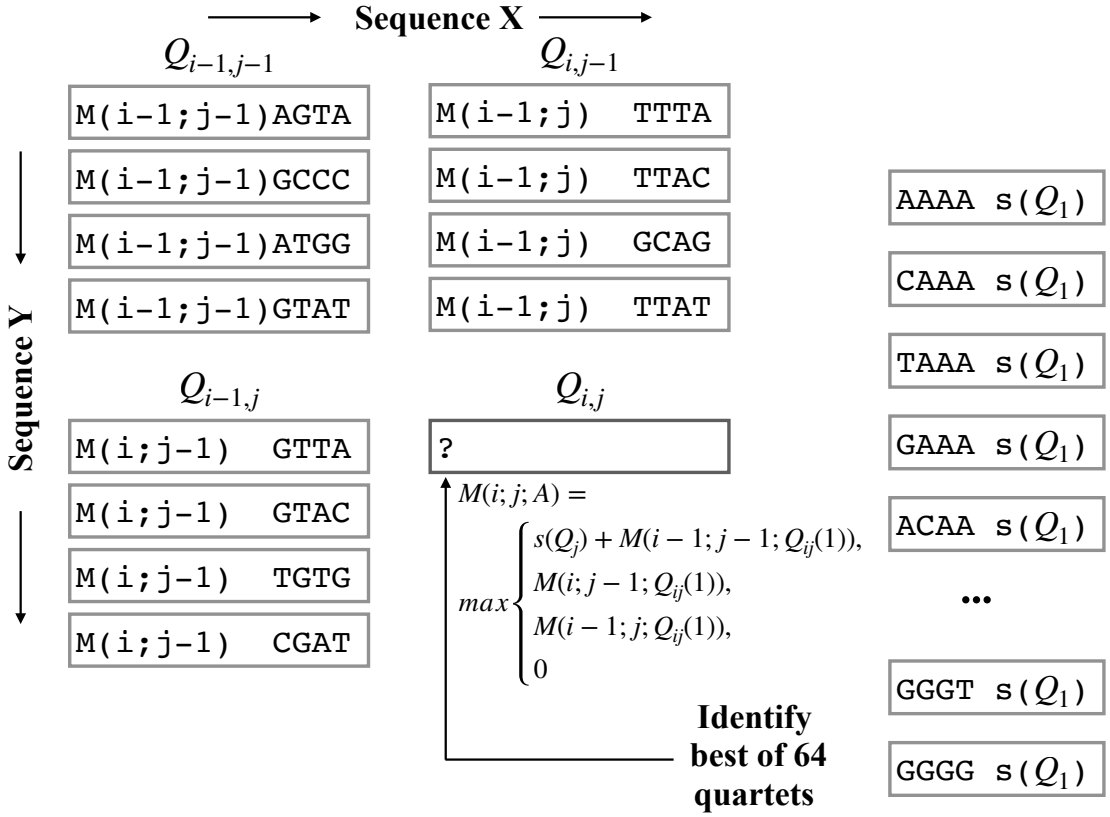


Figure 6.19: **Diagram representing the overlapping local alignment algorithm deduce from the pairwise local alignment algorithm.** Representation of the dynamic programming table $4 \times N - by - P$ centered on the cells $M(i;j), M(i-1;j-1), M(i-1;j), M(i;j-1)$ which contain 4 optimal quartets. To choose the optimal quartet for the state $\nu = A$ for (i,j), we choose the one that maximizes M(i; j; ν = A) among the 64 quartets ending with A.

maximum size, we used the following scoring function:

$$f(X_i, Y_j) = \begin{cases} 1 & \text{if } X_i = X'_i \wedge Y_i = Y'_i \\ -\infty & \text{else} \end{cases} \tag{6.6}$$

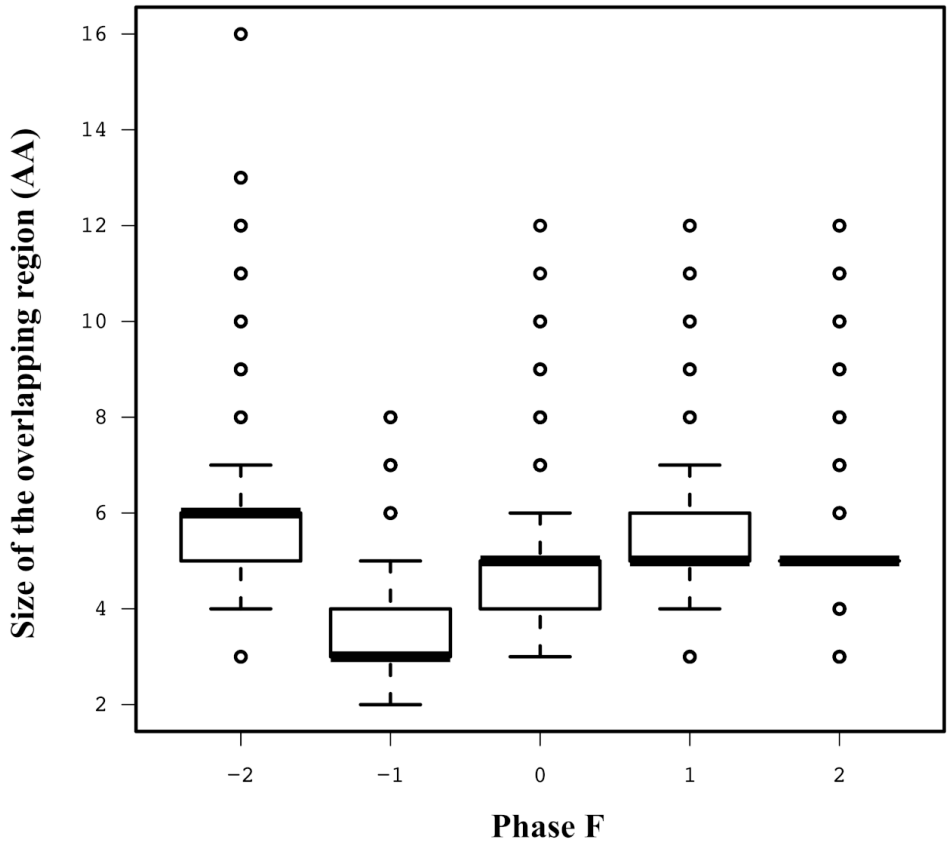


Figure 6.20: **Distribution of the maximum lengths for overlapping regions in natural sequences.**

This approach recently allowed us to recover the two overlapping genes in SARS-COV-2 genome. It has the potential to detect proteins that appeared recently from overprinting. Also, this approach is more flexible, as it allows the addition of introns/deletions/insertions in the form of gaps. In addition, one can use the developments made on multiple alignment algorithms to compress a whole genome. A tree-based multiple alignment method can be used to guide the overlapping encoding of all the sequences from a single genome. The last point that we want to raise here is about the detection of interactions between genes. Indeed, it was proposed that new genes that appeared by *overprinting* interact with the initial gene. Thus, the design of a successful sequence pair may be a signal of overprinting or interaction.

Conclusions

To design molecular systems, we start from naturally optimized components, namely proteins. Proteins can act as structural components, information transporters, or catalysts. It is then possible to create molecular systems composed of one or more proteins capable of complex tasks. To design such new systems, we used natural proteins as templates, with known three-dimensional structures and functions. Design methods use the paradigm that links three-dimensional structures to biological functions. Here, we studied the possibility to design new protein domains, PDZ domains. Because of their role in protein/protein recognition, PDZ domain designs are interesting for the engineering of metabolic pathways. Then, we studied the methodological aspects of biocatalysts design through the redesign of MetRS. Finally, we introduced a protocol to design overlapping PDZ genes, in view of bio-confinement strategies.

For the structural aspect of proteins, we report here the complete redesign of a PDZ domain. We modeled the folded state of the Cask PDZ domain with a *physics-based* energy function combining molecular mechanics, continuum electrostatics, and Monte Carlo sampling. We described the unfolded state with an empirical model. We produced three variants for experimental validations. Circular dichroism experiments showed conservation of the secondary structures and 1D-NMR spectra were typical of folded structures for all three variants. Two had detectable binding for natural PDZ binders.

This complete redesign of a PDZ domain is encouraging for the use of *physics-based* approaches since it does not suffer from biological data dependencies or biases. Indeed, biological data form a biased ensemble because of the nature of experiments by which they are produced or the sampling trends in databases. In addition, physical model parameters are transferable and can be interpreted with ease.

For the catalytic aspect, we redesigned the active site of an enzyme involved in the translation machinery, methionyl-tRNA synthetase (MetRS). First, we studied Met recognition MetRS

through a design method that favors variants that bind Met. It allowed us to produce 17 variants that were all found to be active experimentally. Then, we extended for the first time the method to transition state bindings, so the variants can be sampled according to their catalytic efficiency.

Although we used rather rigorous and non-empirical approaches, we were not able to detect variants with a higher activity than the wild type enzyme. The size of the mutation/conformation space implies the use of approximations for the solvent and the protein flexibility. In addition, the physical model has some additional approximations. Also, we used a parsimony strategy in which we assumed that a small number of mutations could produce a highly active variant. Indeed, modeling errors might accumulate with a larger number of mutations. However, it might be necessary to mutate more positions so one can surpass the optimization of natural proteins.

Next, we considered the incorporation of unnatural amino acids into proteins. Expansion of the genetic code can enrich the properties of designed systems. One example is the use of β amino acids to enrich the backbone geometries and allow protease resistance. This is of interest for the design of stable peptides *in vivo*. In this work, we searched for MetRS variants that can activate β amino acids such as β-Met or β-Val. First, we considered three positions. We used the adaptive MC approach to sample variants according to their β-Met binding affinity. 20 variants were selected for experiments of which 11 were directly obtained from the simulations and 7 had a good predicted stability. Five of the seven stable variants were found to have a measurable catalytic activity for β-Met. For all variants, catalytic efficiencies were weak, however, three had slightly improved selectivity in favor of β-Met, by factors of 2-8. Then, we introduced a screening method that allowed us to scan the whole active site and select positions according to catalytic efficiency. We selected and investigated four quadruplets of positions for the activation of β ligands. At this point, we produced variants predicted as stable and with a gain of activity. A few of these variants are under experimental testing (Y. Mechulam, E. Schmitt, personal communication)

The modeling of the considered states is crucial in this approach. The catalytic efficiency is a subtle equilibrium between a sequence of complex states. Therefore, we introduced a few hypotheses about the contribution of ATP binding and the KMSKS loop conformational change. Since we don't have all the structural information for β amino acids, we made additional

hypotheses for the ligands placements. However, the collaboration with the experimental part will allow us to refine the predictions and improve the calculation models.

Finally, we studied overlapping coding schemes for pairs of PDZ domains. This coding would allow genome compaction and would reduce genetic drift. We produced almost 2000 pairs of overlapping pairs based on five PDZ domains. Three were selected at the end of a filtering process based on physicochemical and evolutionary properties. We investigated these three pairs further with molecular dynamics simulations. One pair was found stable with a simulation of 500 ns. A second pair was found stable for 3 μs. Experimental testing of these pairs is underway (G. Travé, personal communication). This work suggests that filtering on structural features may be sufficient to produce stable pairs. However, explicit correlation optimization may improve the quality of the predictions further. Also, we showed that one can extend these methods to detected natural overlapping pairs in existing proteomes.

We covered many aspects of CPD. However, we are still redesigning existing proteins. CPD also allows the design of new proteins *de novo* without starting from a structural template ([Huang et al., 2016]). New proteins with new backbones and biochemical functions absent in nature can then be designed.

To reduce the problem complexity, we used the fixed backbone approximation where only side chains remain flexible. Other approaches take explicitly into account flexibility. The first uses local "backrub" changes in the backbone geometry ([Smith and Kortemme, 2008]). The search for sequences is then accompanied by local backbone adjustments. Another uses structural fragments obtained from the decomposition of existing structures stored in the PDB ([Mackenzie et al., 2016]). Fragments are grouped into a library that is capable of reproducing the geometries of experimentally observed structures. Assembling those fragments allows taking into account local adjustment in the backbone ([Mackenzie and Grigoryan, 2017]). Proteus proposes a multi-backbone approach where the flexibility is modeled by a pre-defined ensemble of structures, produced by MD simulation for example. Sequences can then populate the available backbone geometries using a hybrid MC algorithm ([Druart et al., 2016]).

We mainly investigated physics-based CPD. However, statistical potentials are also important for CPD. The recent rise of deep-learning ([LeCun et al., 2015]) allowed new approaches to push further the use of biological data. Progress was made by that type of approach for structure predictions ([Senior et al., 2020]), and similar methods are now applied to CPD

([Ingraham et al., 2019]). Also, our team started to applied machine learning methods to the selection of positions to redesign.

This work provides a consistent set of methodological tools for the design of molecular systems using CPD approaches. The design of new proteins may allow the design of a system composed of one or more of these compounds and could be incorporated into living organisms.

Bibliography

[Abramowitz et al., 2004] Abramowitz, J., Grenet, D., Birnbaumer, M., Torres, H. N., and Birnbaumer, L. (2004). Xl s, the extra-long form of the -subunit of the gs g protein, is significantly longer than suspected, and so is its companion alex. *Proceedings of the National Academy of Sciences*, 101(22):8366–8371.

[Amacher et al., 2014] Amacher, J. F., Zhao, R., Spaller, M. R., and Madden, D. R. (2014). Chemically modified peptide scaffolds target the cftr-associated ligand pdz domain. *PLoS ONE*, 9(8):e103650.

[Archontis and Simonson, 2005] Archontis, G. and Simonson, T. (2005). A residue-pairwise generalized born scheme suitable for protein design calculations. *The Journal of Physical Chemistry B*, 109(47):22667–22673.

[Banik and Nandi, 2010] Banik, S. D. and Nandi, N. (2010). Aminoacylation reaction in the histidyl-trna synthetase: Fidelity mechanism of the activation step. *The Journal of Physical Chemistry B*, 114(6):2301–2311.

[Barducci et al., 2008] Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical Review Letters*, 100(2):020603.

[Basdevant et al., 2006] Basdevant, N., Weinstein, H., and Ceruso, M. (2006). Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: Pdz domains, a case study. *Journal of the American Chemical Society*, 128(39):12766–12777.

[Baumann et al., 2019] Baumann, T., Hauf, M., Richter, F., Albers, S., Möglich, A., Ignatova, Z., and Budisa, N. (2019). Computational aminoacyl-trna synthetase library design for photocaged tyrosine. *International Journal of Molecular Sciences*, 20(9):2343.

[Becker et al., 2001] Becker, O., Jr, A. M., Roux, B., and Watanabe, M. (2001). *chapter 4*, page nil. Computational Biochemistry and Biophysics. CRC Press.

[Bhattacherjee and Wallin, 2013] Bhattacherjee, A. and Wallin, S. (2013). Exploring protein-peptide binding specificity through computational peptide screening. *PLoS Computational Biology*, 9(10):e1003277.

[Blazejewski et al., 2019] Blazejewski, T., Ho, H.-I., and Wang, H. H. (2019). Synthetic sequence entanglement augments stability and containment of genetic information in cells. *Science*, 365(6453):595–598.

[Blöchliger et al., 2015] Blöchliger, N., Xu, M., and Caflisch, A. (2015). Peptide binding to a pdz domain by electrostatic steering via nonnative salt bridges. *Biophysical Journal*, 108(9):2362–2370.

[Bonneau et al., 2002] Bonneau, R., Strauss, C. E., Rohl, C. A., Chivian, D., Bradley, P., Malmström, L., Robertson, T., and Baker, D. (2002). De novo prediction of three-dimensional structures for major protein families. *Journal of Molecular Biology*, 322(1):65–78.

[Briggs and Haldane, 1925] Briggs, G. E. and Haldane, J. B. S. (1925). A note on the kinetics of enzyme action. *Biochemical Journal*, 19(2):338–339.

[Busch et al., 2008] Busch, M. S. A., Lopes, A., Mignon, D., and Simonson, T. (2008). Computational protein design: Software implementation, parameter optimization, and performance of a simple model. *Journal of Computational Chemistry*, 29(7):1092–1102.

[Chen et al., 2007] Chen, X., Longgood, J. C., Michnoff, C., Wei, S., Frantz, D. E., and Bezprozvanny, L. (2007). High-throughput screen for small molecule inhibitors of mint1-pdz domains. *ASSAY and Drug Development Technologies*, 5(6):769–784.

[Conte, 2000] Conte, L. L. (2000). Scop: a structural classification of proteins database. *Nucleic Acids Research*, 28(1):257–259.

[Cornell et al., 1995] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197.

[Cornell et al., 1996] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic moleculesj.am.chem.soc.1995,117, 5179-5197. *Journal of the American Chemical Society*, 118(9):2309–2309.

[Crepin et al., 2003] Crepin, T., Schmitt, E., Mechulam, Y., Sampson, P. B., Vaughan, M. D., Honek, J. F., and Blanquet, S. (2003). Use of analogues of methionine and methionyl adenylate to sample conformational changes during catalysis in escherichia coli methionyl-trna synthetase. *Journal of Molecular Biology*, 332(1):59–72.

[Cusack et al., 1990] Cusack, S., Berthet-Colominas, C., Härtlein, M., Nassar, N., and Leberman, R. (1990). A second class of synthetase structure revealed by x-ray analysis of escherichia coli seryl-trna synthetase at 2.5 å. *Nature*, 347(6290):249–255.

[Daura et al., 2001] Daura, X., Gademann, K., Schäfer, H., Jaun, B., Seebach, D., and van Gunsteren, W. F. (2001). he β-peptide hairpin in solution: Conformational study of a β-hexapeptide in methanol by nmr spectroscopy and md simulation. *Journal of the American Chemical Society*, 123(10):2393–2404.

[Denessiouk and Johnson, 2000] Denessiouk, K. A. and Johnson, M. S. (2000). When fold is not important: a common structural framework for adenine and amp binding in 12 unrelated protein families. *Proteins: Structure, Function, and Genetics*, 38(3):310–326.

[Denessiouk and Johnson, 2003] Denessiouk, K. A. and Johnson, M. S. (2003). "acceptor-donor-acceptor" motifs recognize the watson-crick, hoogsteen and sugar "donor-acceptor-donor" edges of adenine and adenosine-containing ligands. *Journal of Molecular Biology*, 333(5):1025–1043.

[Denessiouk et al., 2001] Denessiouk, K. A., Rantanen, V.-V., and Johnson, M. S. (2001). Adenine recognition: a motif present in atp-, coa-, nad-, nadp-, and fad-dependent proteins. *Proteins: Structure, Function, and Genetics*, 44(3):282–291.

[Druart et al., 2016] Druart, K., Bigot, J., Audit, E., and Simonson, T. (2016). A hybrid monte carlo scheme for multibackbone protein design. *Journal of Chemical Theory and Computation*, 12(12):6035–6048.

[Feller et al., 1995] Feller, S. E., Zhang, Y., Pastor, R. W., and Brooks, B. R. (1995). Constant pressure molecular dynamics simulation: the langevin piston method. *The Journal of Chemical Physics*, 103(11):4613–4621.

[First and Fersht, 1995] First, E. A. and Fersht, A. R. (1995). Analysis of the role of the kmsks loop in the catalytic mechanism of the tyrosyl-trna synthetase using multimutant cycles. *Biochemistry*, 34(15):5030–5043.

[Garcia-Viloca, 2004] Garcia-Viloca, M. (2004). How enzymes work: Analysis by modern rate theory and computer simulations. *Science*, 303(5655):186–195.

[General, 2010] General, I. J. (2010). A note on the standard state's binding free energy. *Journal of Chemical Theory and Computation*, 6(8):2520–2524.

[Gough et al., 2001] Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4):903–919.

[Hallen et al., 2018] Hallen, M. A., Martin, J. W., Ojewole, A., Jou, J. D., Lowegard, A. U., Frenkel, M. S., Gainza, P., Nisonoff, H. M., Mukund, A., Wang, S., Holt, G. T., Zhou, D., Dowd, E., and Donald, B. R. (2018). Osprey 3.0: Open-source protein redesign for you, with powerful new features. *Journal of Computational Chemistry*, 39(30):2494–2507.

[Hastings, 1970] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

[Hatos et al., 2020] Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G. I., Bevilacqua, M., Chasapi, A., et al. (2020). Disprot: intrinsic protein disorder annotation in 2020. *Nucleic Acids Research*, 48(D1):D269–D276.

[Hawkins et al., 1995] Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1995). Pairwise solute descreening of solute charges from a dielectric medium. *Chemical Physics Letters*, 246(1-2):122–129.

[Henikoff and Henikoff, 1992] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919.

[Hsueh, 2009] Hsueh, Y.-P. (2009). Calcium/calmodulin-dependent serine protein kinase and mental retardation. *Annals of Neurology*, 66(4):438–443.

[Hsueh et al., 2000] Hsueh, Y.-P., Wang, T.-F., Yang, F.-C., and Sheng, M. (2000). Nuclear translocation and transcription regulation by the membrane-associated guanylate kinase cask/lin-2. *Nature*, 404(6775):298–302.

[Huang et al., 2016] Huang, P.-S., Boyken, S. E., and Baker, D. (2016). The coming of age of de novo protein design. *Nature*, 537(7620):320–327.

[Ibba and Söll, 2000] Ibba, M. and Söll, D. (2000). Aminoacyl-trna synthesis. *Annual Review of Biochemistry*, 69(1):617–650.

[Ingraham et al., 2019] Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. (2019). Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, pages 15820–15831.

[Jencks, 1987] Jencks, W. P. (1987). *Catalysis in chemistry and enzymology*. Courier Corporation.

[Jindal et al., 2017] Jindal, G., Ramachandran, B., Bora, R. P., and Warshel, A. (2017). Exploring the development of ground-state destabilization and transition-state stabilization in two directed evolution paths of kemp eliminases. *ACS Catalysis*, 7(5):3301–3305.

[Jo et al., 2008] Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008). Charmm-gui: a web-based graphical user interface for charmm. *Journal of Computational Chemistry*, 29(11):1859–1865.

[Jorgensen et al., 1983] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935.

[Kaiser et al., 2020] Kaiser, F., Krautwurst, S., Salentin, S., Haupt, V. J., Leberecht, C., Bittrich, S., Labudde, D., and Schroeder, M. (2020). The structural basis of the genetic code: Amino acid recognition by aminoacyl-trna synthetases. *BioRxiv*, page 606459.

[Khan et al., 2011] Khan, S., Sharma, A., Jamwal, A., Sharma, V., Pole, A. K., Thakur, K. K., and Sharma, A. (2011). Uneven spread of cis- and trans-editing aminoacyl-trna synthetase domains within translational compartments of p. falciparum. *Scientific Reports*, 1(1):188.

[Kim et al., 2004] Kim, D. E., Chivian, D., and Baker, D. (2004). Protein structure prediction and analysis using the robetta server. *Nucleic Acids Research*, 32(Web Server):W526–W531.

[Kovacs et al., 2010] Kovacs, E., Tompa, P., Liliom, K., and Kalmar, L. (2010). Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proceedings of the National Academy of Sciences*, 107(12):5429–5434.

[Kowal et al., 2001] Kowal, A. K., Kohrer, C., and RajBhandary, U. L. (2001). Twenty-first aminoacyl-trna synthetase-suppressor trna pairs for possible use in site-specific incorporation of amino acid analogues into proteins in eukaryotes and in eubacteria. *Proceedings of the National Academy of Sciences*, 98(5):2268–2273.

[Krakauer, 2000] Krakauer, D. C. (2000). Stability and evolution of overlapping genes. *Evolution*, 54(3):731–739.

[Krivov et al., 2009] Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with scwrl4. *Proteins: Structure, Function, and Bioinformatics*, 77(4):778–795.

[Larson et al., 2011] Larson, E. T., Kim, J. E., Zucker, F. H., Kelley, A., Mueller, N., Napuli, A. J., Verlinde, C. L., Fan, E., Buckner, F. S., Voorhis, W. C. V., Merritt, E. A., and Hol, W. G. (2011). Structure of leishmania major methionyl-trna synthetase in complex with intermediate products methionyladenylate and pyrophosphate. *Biochimie*, 93(3):570–582.

[Lazaridis and Karplus, 1999] Lazaridis, T. and Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins: Structure, Function, and Genetics*, 35(2):133–152.

[Leatherbarrow et al., 1985] Leatherbarrow, R. J., Fersht, A. R., and Winter, G. (1985). Transition-state stabilization in the mechanism of tyrosyl-trna synthetase revealed by protein engineering. *Proceedings of the National Academy of Sciences*, 82(23):7840–7844.

[Lèbre and Gascuel, 2017] Lèbre, S. and Gascuel, O. (2017). The combinatorics of overlapping genes. *Journal of Theoretical Biology*, 415(nil):90–101.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

[Ling et al., 2009] Ling, J., Reynolds, N., and Ibba, M. (2009). Aminoacyl-trna synthesis and translational quality control. *Annual Review of Microbiology*, 63(1):61–78.

[Liu and Schultz, 2010] Liu, C. C. and Schultz, P. G. (2010). Adding new chemistries to the genetic code. *Annual Review of Biochemistry*, 79(1):413–444.

[Mackenzie and Grigoryan, 2017] Mackenzie, C. O. and Grigoryan, G. (2017). Protein structural motifs in prediction and design. *Current Opinion in Structural Biology*, 44(nil):161–167.

[Mackenzie et al., 2016] Mackenzie, C. O., Zhou, J., and Grigoryan, G. (2016). Tertiary alphabet for the observable protein structural universe. *Proceedings of the National Academy of Sciences*, 113(47):E7438–E7447.

[Marti et al., 2004] Marti, S., Roca, M., Andres, J., Moliner, V., Silla, E., Tunon, I., and Bertran, J. (2004). Theoretical insights in enzyme catalysis. *ChemInform*, 35(19):nil.

[Martinez-Rodriguez et al., 2015] Martinez-Rodriguez, L., Erdogan, O., Jimenez-Rodriguez, M., Gonzalez-Rivera, K., Williams, T., Li, L., Weinreb, V., Collier, M., Chandrasekaran, S. N., Ambroggio, X., Kuhlman, B., and Carter, C. W. (2015). Functional class i and ii amino acid-activating enzymes can be coded by opposite strands of the same gene. *Journal of Biological Chemistry*, 290(32):19710–19725.

[Martyna et al., 1994] Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, 101(5):4177–4189.

[Mészáros et al., 2018] Mészáros, B., Erdős, G., and Dosztányi, Z. (2018). Iupred2a: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, 46(W1):W329–W337.

[Michael et al., 2017] Michael, E., Polydorides, S., Simonson, T., and Archontis, G. (2017). Simple models for nonpolar solvation: Parameterization and testing. *Journal of Computational Chemistry*, nil(nil):nil.

[Mignon et al., 2020] Mignon, D., Druart, K., Opuu, V., Polydorides, S., Villa, F., Gaillard, T., Michael, E., Archontis, G., and Simonson, T. (2020). Proteus software for physics-based protein design. *bioRxiv*.

[Mignon et al., 2017] Mignon, D., Panel, N., Chen, X., Fuentes, E. J., and Simonson, T. (2017). Computational design of the tiam1 pdz domain and its ligand binding. *Journal of Chemical Theory and Computation*, 13(5):2271–2289.

[Mignon and Simonson, 2016] Mignon, D. and Simonson, T. (2016). Comparing three stochastic search algorithms for computational protein design: Monte carlo, replica exchange monte carlo, and a multistart, steepest-descent heuristic. *Journal of Computational Chemistry*, 37(19):1781–1793.

[Nigro, 2019] Nigro, G. (2019). *Incorporation de la beta alanine dans des polypeptides*. PhD thesis, Université Paris-Saclay.

[Nigro et al., 2020] Nigro, G., Bourcier, S., Lazennec-Schurdevin, C., Schmitt, E., Marlière, P., and Mechulam, Y. (2020). Use of β3-methionine as an amino acid substrate of escherichia coli methionyl-trna synthetase. *Journal of Structural Biology*, 209(2):107435.

[Olsson et al., 2011] Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011). Propka3: Consistent treatment of internal and surface residues in empirical pka predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537.

[Opuu et al., 2020a] Opuu, V., Nigro, G., Gaillard, T., Schmitt, E., Mechulam, Y., and Simonson, T. (2020a). Adaptive landscape flattening allows the design of both enzyme: Substrate binding and catalytic power. *PLOS Computational Biology*, 16(1):e1007600.

[Opuu et al., 2017] Opuu, V., Silvert, M., and Simonson, T. (2017). Computational design of fully overlapping coding schemes for protein pairs and triplets. *Scientific Reports*, 7(1):15873.

[Opuu et al., 2020b] Opuu, V., Sun, Y. J., Hou, T., Panel, N., Fuentes, E. J., and Simonson, T. (2020b). A physics-based energy function allows the computational redesign of a pdz domain. *Scientific Reports*, 10(1):11150.

[Organisation, 2015] Organisation, W. H. (2015). World malaria report. geneva.

[Ouelle et al., 1995] Ouelle, D. E., Zindy, F., Ashmun, R. A., and Sherr, C. J. (1995). Alternative reading frames of the ink4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell*, 83(6):993–1000.

[Pavesi et al., 2018] Pavesi, A., Vianelli, A., Chirico, N., Bao, Y., Blinkova, O., Belshaw, R., Firth, A., and Karlin, D. (2018). Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLOS ONE*, 13(10):e0202513.

[Phillips et al., 2005] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. (2005). Scalable molecular dynamics with namd. *Journal of computational chemistry*, 26(16):1781–1802.

[Pokala and Handel, 2005] Pokala, N. and Handel, T. M. (2005). Energy functions for protein design: Adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *Journal of Molecular Biology*, 347(1):203–227.

[Rancurel et al., 2009] Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R., and Karlin, D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *Journal of Virology*, 83(20):10719–10736.

[Richter et al., 2011] Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S., and Baker, D. (2011). De novo enzyme design using rosetta3. *PLoS ONE*, 6(5):e19230.

[Rocklin et al., 2017] Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Chevalier, A., Arrowsmith,

C. H., and Baker, D. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175.

[Salentin et al., 2015] Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., and Schroeder, M. (2015). Plip: Fully automated protein-ligand interaction profiler. *Nucleic Acids Research*, 43(W1):W443–W447.

[Schaefer and Karplus, 1996] Schaefer, M. and Karplus, M. (1996). A comprehensive analytical treatment of continuum electrostatics. *The Journal of Physical Chemistry*, 100(5):1578–1599.

[Schmitt et al., 1994] Schmitt, E., Meinnel, T., Blanquet, S., and Mechulam, Y. (1994). Methionyl-trna synthetase needs an intact and mobile 332kmsks336 motif in catalysis of methionyl adenylate formation. *Journal of Molecular Biology*, 242(4):566–577.

[Schmitt et al., 1995] Schmitt, E., Panvert, M., Blanquet, S., and Mechulam, Y. (1995). Transition state stabilization by the 'high' motif of class i aminoacyl-trna synthetases: the case ofescherichia colimethionyl-trna synthetase. *Nucleic Acids Research*, 23(23):4793–4798.

[Senior et al., 2020] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710.

[Shapovalov and Dunbrack, 2011] Shapovalov, M. V. and Dunbrack, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858.

[Simonson, 2001] Simonson, T. (2001). *chapter 9*, page nil. Computational Biochemistry and Biophysics. CRC Press.

[Simonson, 2019] Simonson, T. (2019). The proteus software for computational protein design. *Ecole Polytechnique, Paris: https://proteus. polytechnique. fr*.

[Simonson et al., 2013] Simonson, T., Gaillard, T., Mignon, D., am Busch, M. S., Lopes, A., Amara, N., Polydorides, S., Sedano, A., Druart, K., and Archontis, G. (2013). Computational protein design: The proteus software and selected applications. *Journal of Computational Chemistry*, 34(28):2472–2484.

[Simonson et al., 2016] Simonson, T., Ye-Lehmann, S., Palmai, Z., Amara, N., Wydau-Dematteis, S., Bigan, E., Druart, K., Moch, C., and Plateau, P. (2016). Redesigning the stereospecificity of tyrosyl-trna synthetase. *Proteins: Structure, Function, and Bioinformatics*, 84(2):240–253.

[Smith and Kortemme, 2008] Smith, C. A. and Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of Molecular Biology*, 380(4):742–756.

[Smith and Kortemme, 2010] Smith, C. A. and Kortemme, T. (2010). Structure-based prediction of the peptide sequence space recognized by natural and synthetic pdz domains. *Journal of Molecular Biology*, 402(2):460–474.

[Smith and Waterman, 1981] Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

[Street and Mayo, 1998] Street, A. G. and Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Folding and Design*, 3(4):253–258.

[Tanrikulu et al., 2009] Tanrikulu, I. C., Schmitt, E., Mechulam, Y., Goddard, W. A., and Tirrell, D. A. (2009). Discovery of escherichia coli methionyl-trna synthetase mutants for efficient labeling of proteins with azidonorleucine in vivo. *Proceedings of the National Academy of Sciences*, 106(36):15285–15290.

[Thorsen et al., 2009] Thorsen, T. S., Madsen, K. L., Rebola, N., Rathje, M., Anggono, V., Bach, A., Moreira, I. S., Stuhr-Hansen, N., Dyhring, T., Peters, D., Beuming, T., Huganir, R., Weinstein, H., Mulle, C., Stromgaard, K., Ronn, L. C. B., and Gether, U. (2009). Identification of a small-molecule inhibitor of the pick1 pdz domain that inhibits hippocampal ltp and ltd. *Proceedings of the National Academy of Sciences*, 107(1):413–418.

[Tian et al., 2019] Tian, C., Kasavajhala, K., Belfon, K. A. A., Raguette, L., Huang, H., Migues, A. N., Bickel, J., Wang, Y., Pincay, J., Wu, Q., and Simmerling, C. (2019). Ff19sb: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of Chemical Theory and Computation*, 16(1):528–552.

[Till and Ullmann, 2009] Till, M. S. and Ullmann, G. M. (2009). Mcvol - a program for calculating protein volumes and identifying cavities by a monte carlo algorithm. *Journal of Molecular Modeling*, 16(3):419–429.

[Tournier et al., 2020] Tournier, V., Topham, C. M., Gilles, A., David, B., Folgoas, C., Moya-Leclair, E., Kamionka, E., Desrousseaux, M.-L., Texier, H., Gavalda, S., Cot, M., Guémard, E., Dalibey, M., Nomme, J., Cioci, G., Barbe, S., Chateau, M., André, I., Duquesne, S., and Marty, A. (2020). An engineered pet depolymerase to break down and recycle plastic bottles. *Nature*, 580(7802):216–219.

[Traoré et al., 2013] Traoré, S., Allouche, D., André, I., de Givry, S., Katsirelos, G., Schiex, T., and Barbe, S. (2013). A new framework for computational protein design through cost function network optimization. *Bioinformatics*, 29(17):2129–2136.

[Tuffery et al., 1997] Tuffery, P., Etchebest, C., and Hazout, S. (1997). Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Engineering Design and Selection*, 10(4):361–372.

[Villa et al., 2017] Villa, F., Mignon, D., Polydorides, S., and Simonson, T. (2017). Comparing pairwise-additive and many-body generalized born models for acid/base calculations and protein design. *Journal of Computational Chemistry*, 38(28):2396–2410.

[Villa et al., 2018] Villa, F., Panel, N., Chen, X., and Simonson, T. (2018). Adaptive landscape flattening in amino acid sequence space for the computational design of protein:peptide binding. *The Journal of Chemical Physics*, 149(7):072302.

[Villa and Simonson, 2018] Villa, F. and Simonson, T. (2018). Protein pka's from adaptive landscape flattening instead of constant-ph simulations. *Journal of Chemical Theory and Computation*, 14(12):6714–6721.

[Vondenhoff and Aerschot, 2012] Vondenhoff, G. H. M. and Aerschot, A. V. (2012). Cheminform abstract: Aminoacyl-trna synthetase inhibitors as potential antibiotics. *ChemInform*, 43(8):no–no.

[Warshel, 1978] Warshel, A. (1978). Energetics of enzyme catalysis. *Proceedings of the National Academy of Sciences*, 75(11):5250–5254.

[Williams, 1978] Williams, G. C. (1978). Evolution of living organisms: Evidence for a new theory of transformation. pierre-p. grassé. *The Quarterly Review of Biology*, 53(4):444–444.

[Xie and Schultz, 2006] Xie, J. and Schultz, P. G. (2006). A chemical toolkit for proteins - an expanded genetic code. *Nature Reviews Molecular Cell Biology*, 7(10):775–782.

[Zurek et al., 2004] Zurek, J., Bowman, A. L., Sokalski, W. A., and Mulholland, A. J. (2004). Mm and qm/mm modeling of threonyl-trna synthetase: Model testing and simulations. *Structural Chemistry*, 15(5):405–414.

DOCTORAL SCHOOL

Titre : Dessin computationel de protéines et d'enzymes

Mots clés : Ingénierie de protéines, mécanique moléculaire, interactions protéine ligand, gènes chevauchants

Résumé : Nous proposons un ensemble de méthodes pour la conception de systèmes moléculaires. Notre stratégie consiste à utiliser comme modèle des machines naturellement optimisées, les protéines. Les protéines peuvent être des briques structurales, des transporteurs d'informations ou des catalyseurs chimiques. Nous utilisons ici des approches computationnelles, complémentaires aux voies expérimentales, pour concevoir de tels systèmes.

Nous avons d'abord entièrement redessiné un domaine PDZ impliqué dans des voies métaboliques. Nous utilisons une approche *physics-based* basée sur la mécanique moléculaire, un modèle de solvant implicite et un échantillonnage Monte Carlo. Parmi plusieurs milliers de variants prédits pour adopter le repliement PDZ, trois ont été sélectionnés et montrent un repliement correct. Deux ont une affinité détectable pour les ligands peptidiques naturels.

Nous avons ensuite re-dessiné le site actif de l'enzyme méthionyl-ARNt synthétase (MetRS). En utilisant un algorithme de type Monte Carlo adaptatif, nous avons sélectionné des variants pour l'affinité MetRS/méthionine (Met). Sur 17 variants testés expérimentalement, 17 sont actifs. La méthode a été ensuite appliquée à l'état de transition pour sélectionner des variants directement sur leur efficacité catalytique.

Nous avons étudié la possibilité de modifier la MetRS pour étendre son activité aux acides aminés β, afin d'étendre le code génétique. Ces acides aminés non-naturels permettraient d'enrichir le répertoire structural des protéines. 20 variants MetRS obtenus à partir de prédictions d'affinité MetRS/β-Met ont été testés. Aucun n'augmente l'activité mais trois ont amélioré la sélectivité en faveur de la β-Met. Nous avons implémenté une méthode de sélection de positions d'intérêt et production de variants pour β-Met et β-Val. Une vingtaine de prédictions sont en cours de tests expérimentaux.

Enfin, la modification de protéines *in vivo* pose la question de leur dérive génétique. Nous introduisons ici une méthode de conception de paires de gènes chevauchants pour des domaines PDZ. Ce codage permettrait de limiter la dérive génétique. Nous avons produit près de 2000 paires de domaines PDZ au codage chevauchant, dont une a été validées par 3 microsecondes de dynamique moléculaire. Des tests expérimentaux sont en cours.

Title : Computational design of proteins and enzymes

Keywords : Protein engineering, molecular mechanics, protein/ligand binding, overlapping genes

Abstract :

We propose a set of methods to design molecular systems. We start from naturally optimized components, namely proteins. Proteins can act as structural components, information transporters, or catalysts. We use computational methods to complement experiments and design protein systems.

First, we fully redesigned a PDZ domain involved in metabolic pathways. We used a physics-based approach combining molecular mechanics, continuum electrostatics, and Monte Carlo sampling. Thousands of variants predicted to adopt the PDZ fold were selected. Three were validated experimentally. Two showed binding of the natural peptide ligand.

Next, we redesigned the active site of the methionyl-tRNA synthetase enzyme (MetRS). We used an adaptive Monte Carlo method to select variants for methionine (Met) binding. Out of 17 predicted variants that were tested experimentally, 17 were found to be active. We extended the method to transition state binding to select mutants directly according to their catalytic power.

We redesigned the MetRS binding site to obtain activity towards two β-amino acids, in order to expand the genetic code. These unnatural amino acids can enhance the structural repertoire of proteins. 20 predicted mutants were tested. Although none had increased β-Met activity, three had a gain in selectivity for β-Met. We then implemented a method to select optimal positions for design and applied it to β-Met and β-Val. Around 20 variants are being experimental tested.

Finally, *in vivo* protein modifications raise the question of their eventual drift away from the original design. We introduce here a design approach for overlapping genes coding PDZ domains. This overlap would reduce genetic drift and provide bio-confinement. We computationally produced almost 2000 pairs of overlapping PDZ domains. One was validated by 3 microsecond molecular dynamic simulations. Experiments are underway.

Institut Polytechnique de Paris
91120 Palaiseau, France