This is the post-print version of the following review: Magi Meconi, G; Sasselli, I; Bianco, V; Onuchic, J; Coluzza, I., Key aspects of the past 30 Years of protein design, *Rep. Prog. Phys.* **2022**.

The final publication is available at IOP Science through DOI: 10.1088/1361-6633/ac78ef

Reports on Progress in Physics

ACCEPTED MANUSCRIPT

Key aspects of the past 30 Years of protein design

To cite this article before publication: Giulia Magi Meconi *et al* 2022 *Rep. Prog. Phys.* in press https://doi.org/10.1088/1361-6633/ac78ef

Manuscript version: Accepted Manuscript

Accepted Manuscript is “the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an ‘Accepted Manuscript’ watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors”

This Accepted Manuscript is **© 2022 IOP Publishing Ltd**.

During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.
As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript is available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by-nc-nd/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions will likely be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 165.124.164.104 on 17/06/2022 at 07:02

IOP Publishing Journal Title
Journal **XX** (XXXX) XXXXXX https://doi.org/XXXX/XXXX

Key aspects of the past 30 Years of protein design

Giulia Magi Meconi[1], Ivan R. Sasselli[1] , Valentino Bianco[2], Jose Onuchic[3] and Ivan Coluzza[4,5]

[1] Computational Biophysics Lab. Center for Cooperative Research in Biomaterials (CIC biomaGUNE), Basque Research and Technology Alliance (BRTA), Paseo de Miramon 182, 20014, Donostia-San Sebastián, Spain.
[2] Onena Medicines, San Sebastian, Spain
[3] Center for Theoretical Biological Physics, Department of Physics & Astronomy, Department of Chemistry, Department of Biosciences, Rice University, Houston, Texas 77251, United States
[4] BCMaterials, Basque Center for Materials, Applications and Nanostructures, Bld. Martina Casiano, UPV/EHU Science Park, Barrio Sarriena s/n, 48940 Leioa, Spain
[5] Basque Foundation for Science, IKERBASQUE, 48009, Bilbao, Spain

E-mail: icoluzza@cicbiomagune.es

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Proteins are the workhorse of life. They are the building infrastructure of living systems; they are the most efficient molecular machines known, and their enzymatic activity is still unmatched in versatility by any artificial system. Perhaps proteins' most remarkable feature is their modularity. The large amount of information required to specify each protein's function is analogically encoded with an alphabet of just ~20 letters. The protein folding problem is how to encode all such information in a sequence of 20 letters. In this review, we go through the last 30 years of research to summarize the state of the art and highlight some applications related to fundamental problems of protein evolution.

Keywords: Protein Design, Heteropolymers, Coarse-graining, Protein Folding, Evolution.

1. Introduction

Proteins are one of the most versatile modular assembling systems in nature. A remarkable feature of proteins is their alphabet of just ~20 letters[1–4]. The use of such a limited set has the advantage that new target structures can be designed (*e.g.,* through evolution) by just changing the orders of the elements along the chain. Moreover, by degrading chains that do not fulfil their purpose, waste in the form of isolated residues can be efficiently recycled for new chains. Incidentally, this is why living organisms can eat each other and use their building blocks for themselves. Encoding the protein function and structure in the sequence is known as *protein design*.

Protein design is a scientific problem that has been one of the most interdisciplinary research fields of the past 30 years. Unfortunately, protein design remains one of the major challenges across biology, physics, and chemistry disciplines. The implications of solving such a problem are enormous and branch into material science, drug design, evolution and even cryptography. For instance, in drug design, an effective computational method to design protein-based ligands for biological targets, such as viruses bacterial or tumour cells, could significantly boost the development of new therapies with reduced side effects. In material science, self-assembly is a highly desired property, and, soon, artificial proteins could represent a new class of designable self-assembling materials. The scope of this review is to describe the state of the art in computational protein design methods and give the reader the information necessary to outline what to expect from this field in the near future.

The design of proteins belongs to the so-called "inverse folding problems" (IFPs). IFPs consist in the search for amino acid sequences whose lowest free energy state (*i.e.,* the native structure) coincides with a given target conformation. Protein design theory has roots in the statistical models of heteropolymers freezing transition [5–10]. Currently, there are several computational methodologies that, in some cases, give remarkable successful results in solving the IFPs. The advent of computational protein evolution (another name for protein design) [6,11–25] opens the possibility to address fundamental questions about the nature of the amino acid alphabet [26–29]. Protein design searches for protein sequences capable of folding into a given backbone conformation. The search is usually done by point mutations while keeping the backbone structure fixed. In addition to several applications to medicine [13,15,30–32] and material science [33–36], protein design offers the possibility to explore fundamental problems of protein evolution.

2. State of the art in protein design: Rosetta

There are many protein design software available [37–44]. Among the freely usable for academic use, the Rosetta package is one of the most recognised and has shown the largest variety of successful applications. Finally, Rosetta offers both design and structure prediction that allows testing the consistency of the prediction within the same package. That is why in this review, we will focus on Rosetta.

Rosetta is a biomolecular modelling software package originally developed for protein structure prediction and protein folding [37–41]. However, over the last two decades, the modelling suite extended its applications to different tasks such as protein-protein docking [45,46], protein-ligand docking [47–55], protein design, loop modelling [15,56–59] and the incorporation of nuclear magnetic resonance (NMR) spectroscopy data [60–67]. Additionally, several protocols have been developed for the interpretation of a wide range of chemical and biological macromolecular systems. This group includes the modelling of interactions with peptides [58,68–77] and nucleic acids [78–86], the antibody modelling [80,87–94] and design [32,95–98], the modelling of membrane proteins [99–102], carbohydrates [103,104] and metalloproteins [49].

The computational protein design consists of searching for amino acid sequences that adopt predefined folded structures and functions. The design methods have two fundamental components: a sampling algorithm to explore the extensive amino acid sequence and conformational space accessible to the protein [95] and a score energy function to rank the solutions [105].

Rosetta Design's exploration of the vast space of possible sequences is guided by using the Monte Carlo simulated annealing algorithm. The heuristic method finds the solution space randomly: every residue mutation to another one is done at a random position. The sampled solutions are accepted/rejected using the Metropolis criterion: the solution is accepted if its energy decreases with respect to the original conformation; whenever the energy increases, the new conformation has a small probability to be accepted (P = e^(- (E_{new} - E_{orig})/T))) [106,107].

The all-atom Rosetta energy function [108] is the potential employed for the energy estimation of the design solutions and it was originally created for the protein design [107,109].

$$\Delta E_{total} = E_{vdW} + E_{hbond} + E_{elec} + E_{disulf} + E_{solv} + E_{BBtorsion} + E_{rotamer} + E_{ref} \quad (1)$$

The potential is a weighted linear combination of physics-based and statistical energy terms: (a) E_{vdW} a 6-12 Lennard -Jones potential for van der Waals forces that favours the close-packed residues; (b) E_{hbond} an explicit orientation-dependence hydrogen-bonding potential; (c) E_{elec} an electrostatic potential between charged residues that includes an additional term representing the probability of observing two amino acids close to each other in the protein structure; (d) E_{disulf} disulfide bond energy; (e) E_{solv} a solvation approximation that favours the hydrophobic amino acids to pack in the interior of the proteins and the polar amino acids to point outward; (f) $E_{BBtorsion}$ backbone torsional angle potential; (g) $E_{rotamer}$ sidechain rotamer energy; (h) E_{ref} unfolded-state reference energy. A comprehensive overview of the full-atomistic score function is contained in the article of Alford et al. [108], where are all the mathematical and physical energy-function details are documented. This potential is essential because all energy terms are pairwise decomposable. Instead of estimating all the interactions among the atoms, the total number of energy contributions is restricted to $\frac{1}{2}$N(N - 1), where N is the number of atoms in the systems. In that way, the approximation considers only the pairwise terms involving the targeted residue, subjected to a mutation or a conformational change during the protein design. Thus, it allows a fast-computational implementation of the energy contributions, which is fundamental for the rapid performance of the Metropolis Monte Carlo (MCM) sampling simulations used by Rosetta during the protein design.

The search of the enormous conformational sequence space guided by the MCM algorithm is typically restricted by reducing the degrees of freedom during the design simulations.

As a first approximation, the flag “fixbb” is a Rosetta fixed backbone design application [49,107,109] in which the backbone is maintained fixed. At the same time, side-chain identities and conformations are allowed to vary during the sequence design [11,110]. The number of residues side-chain conformations is discretised through the Dunbrack rotamer library [111–113]. The rotamer is a side chain conformation described by its values of internal dihedral angles. The rotamers libraries gather, for each residue, a discrete number of values for these torsional angles. These collected rotamers are usually the most frequent and the most energetically favourable. The torsional angle side chains can be backbone independent, f and y backbone angles dependent, or secondary structure-dependent (the rotamer frequencies change considering a-helix or b-sheet motifs). The fixed backbone design is helpful for computational efficiency but is not adequate to sample the sequence space because it does not sample the backbone conformational space. Therefore, it limits the chance to optimize the functional interactions. Hence, the mutation is highly constrained and cannot guarantee that the new sequence will fold into the desired backbone conformation.

The backbone flexibility is a crucial feature for the characterization of natural proteins and the backbone adjustment to accommodate sidechain mutations occurring during the design [114,115]. Rosetta software used several strategies to deal with the backbone flexibility.

(I) The first strategy consists of generating large backbone conformations using short backbone fragments taken from previously solved protein. The fragment-based approach has been used for *de novo* protein design (design without a template structure) and *de novo* backbone folds or function design. SEWING [116] protocol generates *de novo* backbones by assembling large sub-structures of protein (typical helical building blocks). During the backbone design, the method allows the user to incorporate particular features, such as ligand binding sites for the ligand-binding protein design and functional motifs like protein-binding peptides for protein interface design [117].

RosettaRemodel [118] is a versatile approach for protein design, in which the new protein structure is built by sticking together protein fragments or small segments of native protein structures. The secondary structure of the desired protein is specified in a blueprint file. The executable consists of 3 main steps: backbone remodel, sequence design and a final minimization step. RosettaRemodel has been employed as a tool to solve different design problems, such as de novo backbone modelling, sequence design in a fixed backbone, loop modelling, disulfide design, motif grafting and motif deletion and remodelling of proteins. Huang *et al.* used the RosettaRemodel application to design a four-fold repeat and symmetrical TIM-barrel protein. The capability to design the TIM-barrel catalyst is of great interest because the fold of this protein is one of the most common enzyme topologies and has opened new possibilities for the *de novo* design of functional enzymes [119]. Parmeggiani et al. [31] developed a computational method for repeat protein design, taking sequence and structural information from the repeat protein families. On that paper, sets of sequences were designed for six protein families with different secondary structures: tetratricopeptide repeat (TPR), ankyrin (ank), armadillo (arm), HEAT, WD40 and leucine-rich repeats (LRR). [120,121] A similar design protocol was used later for *de novo* design of repeat proteins with open[122] and closed [123] structural architectures.

(II) A second strategy involves a flexible design approach based on the iteration between a fixed backbone sequence optimization via Monte Carlo search and flexible backbone minimization to adjust the designed sequences. [109,124]

FastDesign is a Rosetta design protocol that integrates the sequence design in the FastRelax method for the backbone minimization [125–128]. The algorithm proceeds in two main steps. In the first step (fixed-backbone sequence design), the backbone is kept fixed, but the side chains' mutation and the rotameric conformations' optimisation are allowed. In the second step (fixed-sequence backbone minimization), a gradient-based minimization of torsional degrees of freedom is applied to relax the entire structure while the sequence is maintained fixed. The main principle of the FastDesign protocol is the iteration of these two steps. A single FastDesign cycle consists of distinct rounds (default is 4) of design and repacking of the side chains follow by backbone and side-chain minimization. At each round, the repulsive part of the van der Walls energy contribution is progressively scaled from 2% to 100% of its total value to avoid clashes due to the amino acid mutations. The protocol runs different cycles (usually 5), and the best scoring pose, among all the cycles performed, is selected representing the output structure. The Fast Design method found many applications for the design of new protein functions [14,129,130].

(III) BackrubEnsemble [131–135] is a method of flexible backbone design that leads to a structural ensemble of the main chain by rotating backbone segments through the application of the Backrub algorithm[136]. The protocol works in two steps. The first step generates random backbone ensembles after applying the Backrub motion. This algorithm rotates as a rigid body, a backbone protein segment around the axis defined by the segment's starting and ending C_a atoms. The moves are accepted or rejected using a Metropolis criterion. The second step carries out a fixed-backbone sequence design. The sampling of the side chains conformational space depends on the probability distributions described by the Dunbrack rotamer library, and the Metropolis criterion selects the proposed solutions.

The Backrub Ensemble was shown to reproduce better the experimental observed sequence conformational fluctuations [134,137,138] and sequence variations in protein-protein [132,133,135] interface compared with the fixed-backbone sequence design applications. The algorithm also found its application for the design of protein with recognition functionality [139].

(IV) CoupleMoves [140] is a Rosetta application that "couples" in a single Monte Carlo step, backbone and sidechains movements. In this way, the backbone can react at once to the conformational and identity changes of the side chains, enabling sampling of backbone and amino acid sequences movements, which may be previously rejected for the noncouple FastDesign and BackrubEnsemble methods due to sidechain clashes.

Mutations of side chains to shorter lengths are more favourable, as they reduce the likelihood of collisions between side chains. However, this can cause the backbone to collapse to accommodate the amino acid replacement. To minimise the possibility that mutations occur with smaller side chains, the CoupleMoves application uses a different strategy for the sampling of side chains: at each side chain move, all the possible rotamers are considered, and the mutation and torsional angle with the highest probability is selected, according to the Boltzmann-weighted Rosetta score. The CoupleMoves method has also been used for designing small ligand binding sites, combining ligand translation and rotations with the switching of ligand conformers. The original CoupleMoves uses the Backrub algorithm to sample the backbone move, but recently the kinematic closure (KIC) algorithm [141] has been introduced to perform the backbone moves.

The ability to design sequences is not only limited to the creation of a protein with a specific function and increased thermodynamic stability but also aim to greater ambition. For example the multi-specificity design[142], generates protein sequences with low energy affinity to multiple binding partners.

RECON [12] is a Rosetta multi-specificity design method that designs proteins with the ability to bind with multiple different partners. The algorithm allows each protein-energy state to explore their local sequence and conformational space to reach its energetic minimum. Then, sequence constraints are iteratively applied such that the corresponding positions in the different states converge to the same amino acid. RECON can be helpful for the antibody design to recognise a new variant of the virus [143].

Interestingly, Rosetta design algorithms produce a solution space that is quite distinct from one of the natural protein sequences [144]. Of course, considering the astronomical size of the protein solution space, it is likely that computer-generated sequences will have a low chance of finding a natural solution. However, it has to be noted that typically Rosetta tends to diverge from natural sequences imposed as initial conditions to the design simulation [144].

Hence, it might be possible that there is space for the development of design algorithms capable of exploring sequences closer to the natural ones.

3. What makes protein designable

The protein design success strengthens the interest in a fundamental question about proteins:" What makes a protein designable?". In other words, what is so exceptional about the proteins compared to the other members of the large class of heteropolymers.

3.1. Fundamental aspect of design

In this section, we summarize the essential aspects that connect the folding of a generalized protein with the design of its sequence. To this end, we will follow the derivation and analysis of the pioneers in the field [7,145–148].

Although the derivation is valid only in a mean-field approximation, the final result will give a clear and simple physical explanation of what it means to design a protein. The Random Energy Model (REM) [145] is a powerful theory that inspired the mean-field description of the freezing transition of heteropolymers [7,146]. The equivalence between REM and random heteropolymers (RHP) , hypothesized by Bryngelson and Wolynes [7], was proven valid in the mean-field limit and for an alphabet size larger than the number of residues by Shakhnovich and Gutin [147]. An RHP protein is represented as a collection of beads connected by a backbone, interacting with others. Each bead is a residue, and the residue-residue interaction depends on the amino acids' particular identity. Hence, a REM protein is defined by a *conformation*, the specific arrangement of the backbone, and a *sequence* that is the ordered list of amino acids along the backbone. Since we are in a mean-field approximation, we can assume we can thread any possible sequence on each conformation. This hypothesis might appear as an oversimplification because of the excluded volume of the amino acid side chains. However, if small backbone fluctuations are allowed, the number of possible threads (or capacity) of know protein structures are astronomical [149].

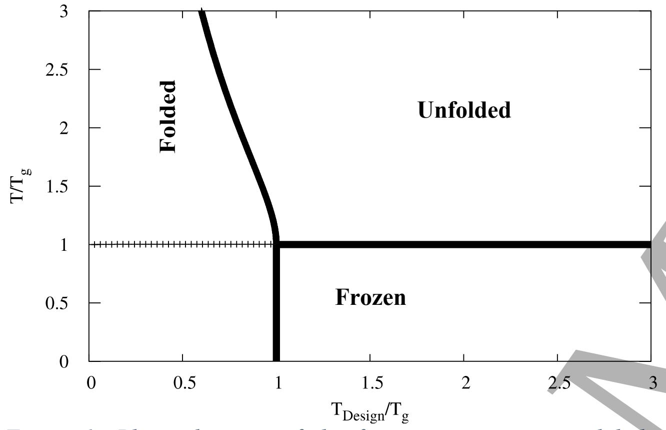


Figure 1: Phase diagram of the freezing transition in globular heteropolymers with a designed sequence at rescaled temperature T_{Design}/T_g versus the rescaled temperature T/T_g at which folding is performed. We can identify three phases: 1) ***Frozen*** *phase in the region $T/T_g<1$ and $T_{Design}/T_g>1$, in which the folding dynamics is glassy. 2)* ***Unfolded*** *phase for $T/T_g>$phase line and $T_{Design}/T_g>1$ where the design and folding explore random sequences and conformations respectively. 3)* ***Folded*** *phase for $T_{Design}/T_g<$ phase liens where the design can successfully optimize sequences for a target structure that is then dynamically accessible. For $T/T_g<1$ the kinetics is slow.*

In other words, the probability $P(E_A,E_B)$ of observing a protein in conformation A with energy E_A and a second one with energy E_B is simply the product of the probabilities $P(E_A,E_B)= P(E_A)\, P(E_B)$.

In REM, the total free energy $\mathcal{F}(T)$ of a random heteropolymer is:

$$\mathcal{F}(T) = \langle \mathcal{F}_{seq}(T) \rangle = -T \langle \ln \mathcal{F}_{seq}(T) \rangle \quad (2)$$

where $\mathcal{F}_{seq}(Z_{seq})$ is the free energy (partition function) for a possible random sequence and T is the temperature. The averages $\langle \ldots \rangle$ are done over all possible sequences. The free energy per monomer is defined as:

$$F(T)/N = \begin{cases} \mathcal{L}\left[\overline{E} - \frac{\sigma_B^2}{2\mathrm{T}}\right] - T\omega & \text{if } T > T_g \\ \mathcal{L}\left[\overline{E} - \frac{\sigma_B^2}{2T_g}\right] - T_g\omega & \text{if } T \le T_g \end{cases} \quad (3)$$

Where $\overline{E}$ and σ_B^2 are the average and variance of the interaction matrix, respectively, $\mathcal{L}$ is the valence of each residue, and ω is the conformational entropy per monomer defined such that $\mathcal{M} = e^{\omega N}$ is the number of states. The meaning of ω is crucial to answering the initial question about the designability of the proteins. Still, its definition is not practical because it depends on the arbitrary definition of the number of states or “compact” states as in the original REM. We propose that a more viable parameter is the folding resolution. Section 3.3 will demonstrate this argument using models beyond the lattice protein approximation. But in the meantime, we keep deriving the theory of heteropolymer freezing.

In REM there is the temperature $T_g = \frac{\sigma_{\mathrm{B}}\mathcal{L}^{\frac{1}{2}}}{(2\omega)^{1/2}}$ below which the distribution of states become discrete and the entropy per monomer vanishes:

$$S(T) = -\frac{dF(T)}{dT}\Big|_{T=T_g} = \omega - \mathcal{L}\frac{\sigma_B^2}{\mathrm{T}_g{}^2} = 0. \quad (4)$$

The temperature T_g is called *glass temperature* because below it the system is trapped in one of the conformations that belong to the discrete region of the density of states. Above the glass temperature T_g, the random-energy heteropolymer explores many states practically independent of the particular sequence of amino acids. However, as the temperature is lowered, the equilibrium is dominated by a few discrete states of low energy highly dependent on the specific sequence. The transition at $T = T_g$ is called the freezing transition [147,150].

Initially, it was suggested that the random-energy model might provide a valuable model for protein folding, as it yields a unique ground state with a probability independent of the system size. However, the energy differences between structurally distinct states in the discrete region of the energy spectrum are only of the order of $\sqrt{N}$, which does not allow for a robust equilibrium state. The question is then if it is possible to design particular sequences that freeze into a stable ground state.

For such an approach to work, the energy of the target state must be well separated from the boundaries of the continuous distribution of states, where the glassy states accumulate (at typical distances of order $\sqrt{N}$). Using mean-field arguments similar to the ones used above, we can derive an expression for the average energy of the designed state E_d as a function of the temperature of the canonical ensemble of sequences T_d. We start by choosing a target conformation C_d as our tentative native state. This conformation is characterized by the energy $E_d = \mathcal{H}(S_d, C_d)$ that depends on the sequence S_d. The partition function obtained by summing over all possible sequences is denoted by W, and it defines a free energy F_W per monomer through:

$$\begin{aligned}\frac{F_W}{N} \equiv -T_d \ln W(T_d) &= -T_d \ln\left[\left\langle\frac{\exp\left[-\mathcal{H}(S_d, C_d)\right]}{T_d}\right\rangle\right] \\ &\simeq \langle\mathcal{H}\rangle - \frac{1}{2T_d\left[\langle\mathcal{H}^2\rangle - \langle\mathcal{H}\rangle^2\right]} \\ &= \mathcal{L}\left[\overline{E} - \frac{\sigma_B^2}{2T_d}\right].\end{aligned} \quad (5)$$

Where T_d represents the design temperature. In terms of F_W we can write an approximate expression for the average energy of the designed sequence $\frac{\langle E_d\rangle}{N} = -\frac{\partial \ln W}{\partial\left(\frac{1}{T}\right)}\Big|_{T\to T_d}$, which does not depend on the target conformation, but instead shows that the energy per monomer is linear in the inverse design temperature

$$\frac{\langle E_d\rangle}{N} = \mathcal{L}\left[\overline{E} - \frac{\sigma_B^2}{T_d}\right] \quad (6)$$

For a target conformation C_d to be the global energy minimum, it must be the equilibrium configuration at a temperature $T_f > T_g$. In the protein folding funnel picture [7], this condition also means that the folding follows a downhill dynamic. Eq. (6) translate into the equality $F(T_f) = \langle E_d\rangle$ or

$$\mathcal{L}\left[\overline{E} - \frac{\sigma_B^2}{2\mathrm{T}_f}\right] - T_f\omega = \mathcal{L}\left[\overline{E} - \frac{\sigma_B^2}{T_d}\right] \quad (7)$$

that rewritten in terms of Tg

$$\mathcal{L}\left[\overline{E}-\frac{\sigma_B^2}{2\mathrm{T}_f}\left(1+\frac{T_f^2}{T_g^2}\right)\right]=\mathcal{L}\left[\overline{E}-\frac{\sigma_B^2}{T_d}\right] \qquad (8)$$

Which leads to a simple expression

$$\frac{1}{T_f^2}+\frac{1}{T_g^2}=\frac{2}{T_f T_d} \qquad (9)$$

which depends on the variance σ_B, but is independent of the mean value of the interaction.

$$\frac{T_g^2}{T_f^2}+1=\frac{2T_g^2}{T_f T_d} \qquad (10)$$

Using such relation is possible to construct a phase diagram that describes the general link between design and folding in heteropolymers (see Figure 1). The phase diagram entirely depends on the glass temperature T_g. The larger T_g the more prominent will be the Folded region or more effortless it will be to find solutions to the design problems.

For example, maximising the alphabet size *q* would undoubtedly do the trick as it reduces frustration. The limit of *q*→∞ guarantees the lowest possible frustration.

An analogous phase diagram to the one plotted in Figure 1 can be done following the pioneering paper of Bryngelson and Wolynes [7]. In Figure 1 of Ref. [7] the freezing phase diagram is plotted as a function of the distribution width of the non-native states $\frac{\Delta \mathrm{L}}{\mathrm{T}}$, a measure of the frustration versus the gap between the native energies *L* and the average non-native ones $\bar{L}$ $(L-\bar{L})/T$. For large gaps, the proteins fold, indicating again that the solutions to the design problems should be located by minimizing the energy of the native state, reducing the frustration to the minimum. A particular solution is to create a set of interactions so that the native state is by construction the lowest energy state. Such models are generally referred to as Gō-Models [3,7,151–173]. According to the "minimum frustration principle" introduced by Wolynes and Onuchic [7], evolution optimized natural sequences, and Gō-proteins share a folding energy landscape with a single global minimum and folding proceeds as a downhill process. Hence, in a Gō-protein, the glass transition is suppressed by construction.

In nature and for most practical applications, it is difficult to reach high values of *q*, so an alternative approach to increase designability is to control the configuration entropy *ω*.

3.2. *Designability and Configurational Entropy ω*

A formidable prediction of REM is identifying the condition for which a solution to the design problem exists [146,148,174].

We can start by taking the entropy in sequence space for a given target conformation *C* of the design process to define such requirements. From the

$$\mathrm{S_C}=\frac{\partial - T_d \ln \sum_{\mathrm{seq}} \exp\left[\mathrm{H(seq,C)/T_d}\right]}{\partial T_d} \qquad (11)$$

where the sum is performed over all possible sequencesN_{seq}that might be generated with the N_C residues of the conformation *C* and an alphabet of *q* amino acid types. N_{seq} and q are connected via the effective number of amino acid types q_{eff} used during the design:

$$N_{seq}=q_{eff}^{N_C};\ \ln q_{eff}=-\sum_{i=1}^{q} p_i \ln p_i \leq \ln q \qquad (12)$$

where p_i is the fraction of each residue used. q_{eff} has its maximum in *q* when the composition is perfectly heterogeneous ($p_i = 1/q$).

Hence,

$$S_C=\ln N_{seq}-\frac{\mathcal{L}\sigma_B^2}{2T_d^2}=\ln q_{eff}-\omega\frac{T_g^2}{T_d^2} \qquad (13)$$

which in terms of the number of solutions to the design problem N_{sol}

$$N_{sol}(T_d)=q_{eff}\, e^{-\omega\frac{T_g^2}{T_d^2}}=e^{\ln q_{eff}-\omega\frac{T_g^2}{T_d^2}} \qquad (14)$$

Designed sequences are obtained when $T_d/T_g \leqq 1$, hence for the design to have a chance of success $N_{sol}(T_g)\geq 1$, which requires the condition $\ln q_{eff} > \omega$ or the simple and powerful prediction of REM $\mathrm{q} > \mathrm{e}^{\omega}$ introduced by Finkelstein et al. [174] in 1993.

The prediction defines the intuitive condition that the alphabet used must be larger than the encoding space of the structure.

In the original formulation of REM, ω was defined as $\omega = \frac{\ln \mathcal{M}}{N}$ and $\mathcal{M}$ is the number of accessible, *compact* conformations per monomer [146,175]. It is important to stress that the *compact* polymer conformations are less than the total possible ones, hence $\omega < s$ where s is the entropy of the backbone. An operative definition of *compact* for off-lattice polymers is not given in the REM, making it difficult to establish a general methodology to estimate ω and, in turn, the designability of a heteropolymer.

3.3. Role of folding resolution and directionality of the interactions

Ultimately, a successful design should produce a protein that folds into the original target structure. The folding success is usually measured as the structural difference between the target and the refolded structures. That difference is the refolding resolution of the model. The resolution has a profound meaning on the understanding of protein design. The reason is the connection between resolution and space of compact structures ω.

ω represents the space of all possible target structures, which is an arbitrary definition depending on how conformations are classified.

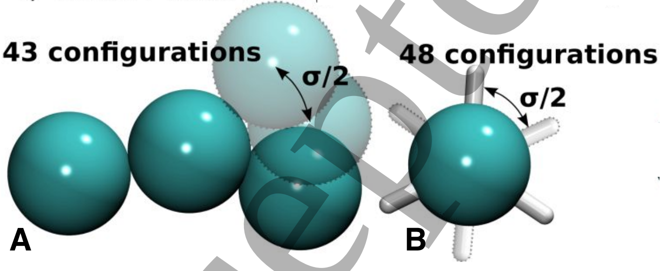


Figure 2: Scheme of the contributions to total conformation entropy $s_{saw}|_{N=3}$ *of a self-avoiding trimer including considering a resolution* $a = \frac{\sigma}{2}$. *There are then 43 backbone configurations (A) and 48 rotational degrees of freedom of each bead (B).*

A solution is to consider the desired folding resolution. Such resolution is defined through the characteristic length a that defines minimum separation to distinguish two atoms in two backbone conformations. Recently Cardelli *et al.* [176] reformulate the definition of ω as the number of accessible configurations partitioned by a, effectively introducing the resolution back into the protein folding theory.

The higher the desired resolution, the larger the conformational space ω, involving a more extensive alphabet q to design successfully. That is why the entire description of protein design must depend on the definition of the resolution a used. In the original formulation of the theory, such parameter was not essential because the reference model systems were proteins on the lattice with a discrete conformational space.

To prove the necessity of the resolution a, Cardelli *et al.* introduced a designable heteropolymer model of which ω is computed as a function of a.

Cardelli's new approach allows testing the predictions of the REM that a system is designable whenever $q = e^{\omega}$. Moreover, the procedure allows assessing the importance of directional interactions to the alphabet size. The latter is done by introducing patches on the surface of the beads, reminiscent of the protein backbone hydrogen bonds

To compute ω, the authors connected the entropy of a protein chain to a system for which the entropy can be computed analytically.

First, we need to compute the absolute entropy of a self-avoiding polymer $s_{saw} = \ln(N_{saw})$ where N_{saw} is the number of conformations of a self-avoiding chain.

To correctly compute s_{saw}, it is necessary to know the number of conformations of a reference state.

The chosen reference state is a trimer of self-avoiding bonded beads, whose conformations can be enumerated analytically as a function of resolution a.

Introducing the resolution $a = \frac{\sigma}{2}$, with σ the hard-core bead radius, the number of conformations $s_{saw}|_{N=3} = 12.9$ can be computed analytically.

Starting from the trimer as the reference system, the total entropy for a self-avoiding polymer of length $N = 50$ is calculated with a potent particle insertion method [177,178] that computes the variation in the partition function upon the particle addition.

$$s_{saw}|_{N=50} = s_{saw}|_{N=3} - \left[s_{saw}^{simul}|_{N=3} + 3\ln\left(12\left(\frac{\sigma}{a}\right)^2\right)\right] + \\ + s_{saw}^{simul}|_{N=50} + 50\ln\left(12\left(\frac{\sigma}{a}\right)^2\right) = 368 \qquad (15)$$

Where the authors have considered the rotational degrees of freedom of the particles. Using the expression in Eq.15, it is possible to compute the entropy variation for different values of a confirming that the number of configurations, and hence ω increase with the resolution.

In fact, for $a = \frac{\sigma}{10}$ (which in protein would correspond to 0.4 Å resolution [179]) $s_{saw}|_{N=50} = 377$, while for $a = 1.5\sigma$, $s_{saw}|_{N=50} = 359$, corresponding to a 2% increase (see Figure 3).

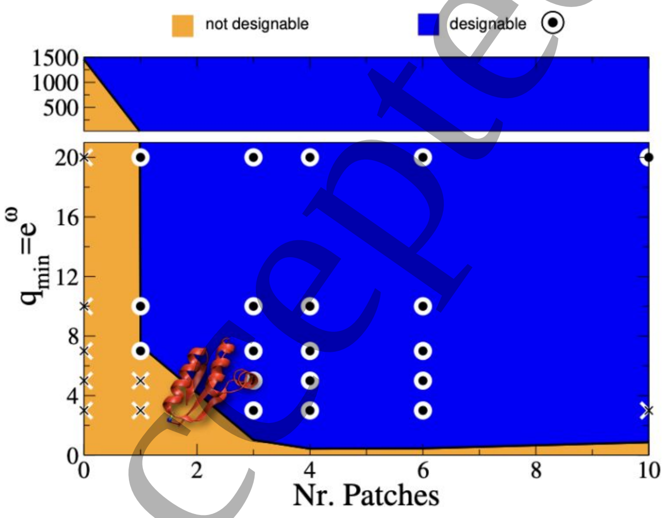


Figure 4: The line represents the alphabet size $q = e^{\omega}$ at which the transition between not designable and designable occurs. Accordingly, two areas are defined: yellow area (not designable) and blue area (designable). The circles are the designable cases, i.e. where the polymer designed with the indicated alphabet has been tested to fold into the target structure, while the crosses the ones where it does not (not designable) [179]. *For 2 directional interactions, like in proteins, the minimum alphabet size for design is predicted to 4 letters, a prediction that has been verified computationally* [180].

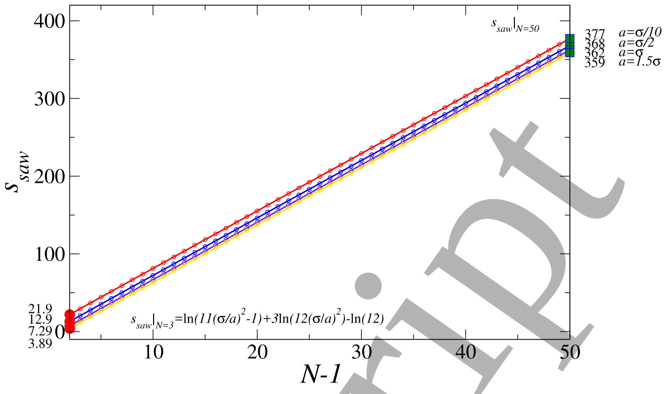


Figure 3: Dependence of the chain entropy s_{saw} as a function of the chain length N. Different curves depend on different resolutions a.

The study offered three major conclusions. First, the relation between alphabet and designability works only once a target resolution is defined. Secondly, directional interactions are imperative for any practical application of polymer design as few patches quickly reduce the minimum alphabet size from q=1500 to just q=7 (see Figure 4). This is a massive reduction with profound implications on the evolution of life that ultimately depends on the possibility of optimizing and storing structures using a code of 20 letters. The third key result predicts that any polymer with 2 to 8 directional interactions should be designable with tiny alphabets of 3,4 letters (see Figure 4). It is again confirming the importance of directional interactions. Proteins are a particular case of the 2-patches scenario, and we confirmed the prediction of the phase diagram in Figure 4 in a recent publication [180]. The study of the origin of the 20-amino acid alphabet is a fascinating problem that has been extensively studied in the past 30 years. We will discuss it in section 5.

4. Coarse graining

The introduction of the REM theory for protein folding and design paved the way for a new protein coarse-graining approach.

As for any computational molecular model, the system is fully characterized by the Hamiltonian that describes the interaction between the different atoms. A coarse-grained model is no different in this respect, but effective interactions between groups of atoms replace the atomic interactions. A carefully constructed coarse-grained model retains the full description of the phenomena under study at a fraction of the computational cost. We will present coarse-grained models that have proven to be designable or have the potential to be, although they have not been tested. Hence, our primary requirement for a coarse-grained model to be a viable protein representation is that it satisfies the REM requirements.

4.1. Lattice proteins

The success of the REM in describing the relation between folding and freezing has been proved by many studies performed using lattice models of proteins[1,5,159,181–187]. In this section, we focus on applying lattice models to understand the fundamental properties of protein folding. However, it is essential to mention that lattice models have been extended to accurately describe protein folding structure prediction [188–193]. They are simple enough to allow for extensive screening of protein sequences and structures aiming at the fundamental mechanism of proteins function. An exhaustive overview of the applications of lattice proteins is beyond the scope of this review.

However, we think it is instructive to list exciting examples. It is important to note that such simple models often cannot provide a quantitative description but instead offer the possibility to test the hypothesis against large protein populations. In particular, the possibility of quickly performing protein design allows studying complex problems related to protein evolution [5,194–197], protein aggregation [198–202], and even intricate protein knotting [187,203,204].

Protein-Protein interaction is a fascinating application of lattice proteins. Lattice proteins models represent a powerful tool to reach problems at large time and size scales. They allow for efficient design of molecule-substrate binding specificity [1,4,184].

One of the critical properties of biological molecules is that they can bind strongly to specific substrates yet interact only weakly with the many other molecules they encounter in the cellular environment.

After the synthesis at the ribosome, polypeptide chains are exposed to a highly crowded cellular environment. Despite many non-specific interactions, the chain can select a subset of amino acid contacts that funnel the free energy landscape toward a unique native/folded state. For instance, it was observed that proteins designed to interact strongly with each other are unlikely to bind non-specifically to other substrates[184,205]. This result has also been verified off-lattice by Nerattini et al. [206]. Therefore, the conflict between specific interactions and weak non-specific interaction among small numbers of biomolecules need not be a severe design constraint.

However, protein aggregation and denaturation are mostly unavoidable when proteins are over-expressed at concentrations higher than the physiological ones. That is why protein expression is highly regulated in cells. The concentration of each protein is kept below a critical value. In 2008 Zhang et al. [202] presented a statistical analysis to rationalize the relative concentrations of monomeric, complex and misbound proteins. The authors concluded that in addition to strong specific

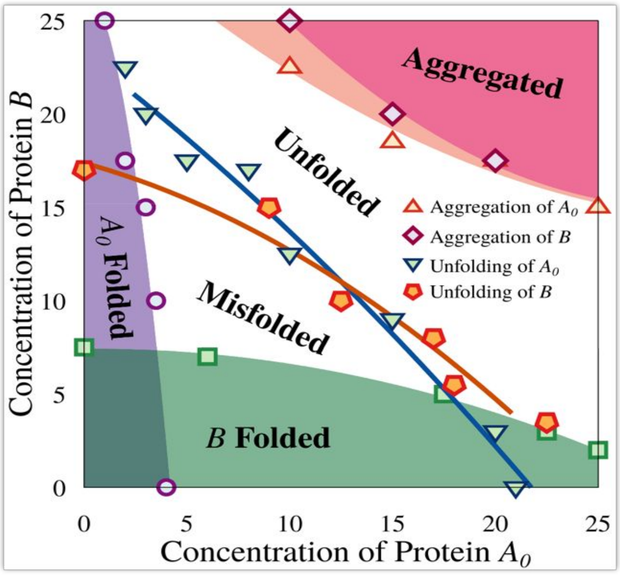


Figure 5: Aggregation phase diagram for two designed proteins. The folded regions are orthogonal to each other proving that cross-aggregation is not a major problem for evolution.

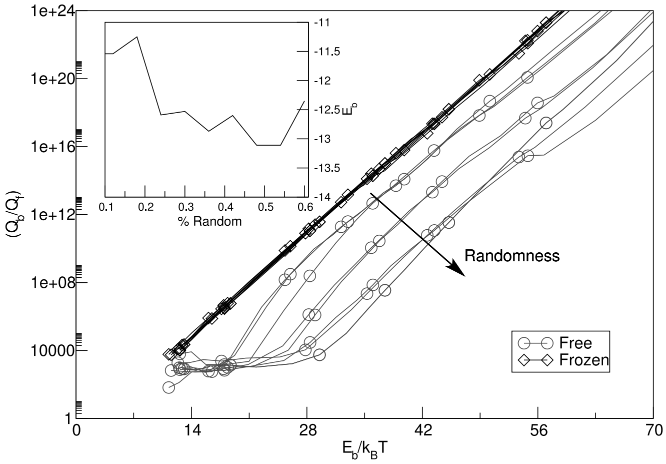


Figure 7: The binding strength of a protein is determined by the ratio Q_b/Q_f (where Q_b as the partition sum of all protein conformations that have at least one contact to the substrate, and Q_f is the partition sum of a ``free" protein in the bulk) [184]. *When the protein is frozen in its native state (diamonds), the conformational entropy does not change upon unbinding. At a fixed (reduced) temperature, proteins that fold upon binding (circles) are less strongly bound than ordered proteins (diamonds) with the same binding strength E_b (plotted in the inset).*

interactions, the presence of compartments and reduced protein-protein interactions (PPIs) could be beneficial in solving the mis-interaction problem.

However, protein expression levels are linearly anti-correlated with their aggregation propensity [207]. This observation suggests that the simple arguments of weaker non-specific interactions are not enough because in a high protein concentration soup, eventually, they should dominate. Still, cells regulate each protein independently of the overall protein concentration. Hence, there is more to the story.

Recently Bianco *et al.* [200] showed that in protein mixtures, each component could maintain its folded state at densities more significant than the one they would precipitate in single-species solutions (see Figure 5). The authors demonstrate the generality of their observation over many different proteins using computer simulations capable of fully characterizing all the mixtures' cross-aggregation phase diagrams. Dynamic light scattering experiments were performed to evaluate the aggregation of two proteins, bovine serum albumin (BSA) and consensus tetratricopeptide repeat (CTPR), in solutions of one or both proteins. The experiments confirm their hypothesis and simulations. These findings demonstrate that below the aggregation concentration, a protein folds unperturbed by the presence of other proteins. Thanks to this property, cells can just regulate the expression of each protein regardless of the concentration of the others, enormously simplifying the entire problem.

Protein-protein interactions can also be tuned to induce folding to a specific configuration upon binding [4,184,185]. Moreover, the disordered state does not affect the protein's binding selectivity but reduces the affinity in a controllable fashion.

In Figure 7, we plot the dependence of the binding affinity of a protein designed to bind to a given substrate as a function of the degree of disorder (``Randomness") induced in the protein. The disorder is added during the design procedure by allowing the identity of a few residues to fluctuate freely hence creating random spots along the protein chain. When the number of random residues becomes too large, the protein cannot fold when unbound, and the binding affinity is significantly reduced (see Figure 7). The behaviour of such randomised proteins is reminiscent of the well-known intrinsically disordered proteins (IDPs) [198], and the design protocol could be used to produce artificial IDPs.

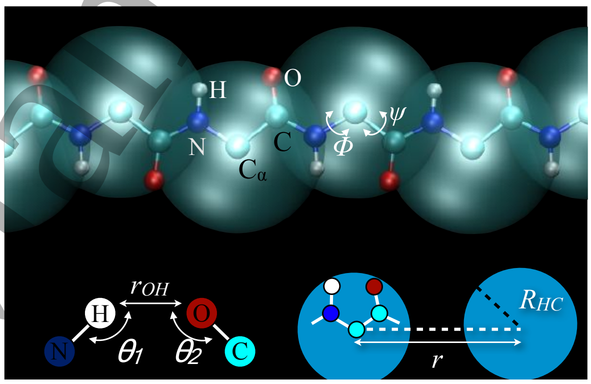


Figure 6: Real-space representation of the backbone of the caterpillar model. The large blue sphere represents the self-avoidance volume $R_{HC} = 2.0$ Å of the C_α atoms. The H and O atoms interact through a 10-12 Lennard-Jones potential tuned with a quadratic orientation term that selects for alignment of the C, H, O, and N atoms involved in a bond. The backbone fluctuates only around the torsional angles ϕ and ψ.

4.2. Caterpillar

In what follows, we will give more details about the Caterpillar protein model.

Recently, inspired by the tube model of Maritan and co-workers [209–211], the Caterpillar protein model approximates a typical protein with the full-atomistic backbone but without the side chains that define each amino acid [22,24]. Instead, the chemical differences are represented by an effective spherically symmetric potential centred on the C_α atoms (see Figure 6). The sphere's zig-zag arrangement that follows the backbone reminds of a Caterpillar worm, hence the name ``Caterpillar''.

The model has two key ingredients, the backbone hydrogen bond interactions and the heterogenous 20 letter amino acid alphabet.
The first element sets in the directional interactions. The presence of the hydrogen bonds was a necessary condition to induce a local protein-like secondary structure and, at the same time, recovered the designability properties [22] with a 20 letter alphabet. The results show that the Caterpillar model describes a system with designable folding behaviour strengthening the importance of directional interactions highlighted in section 3.

The 20 letters instead represent the chemical variability of the amino acids, and their accuracy defines how quantitative the model will be. The interactions were obtained by combining the maximum entropy principle [212–214] with the design algorithm developed for the Caterpillar model. Following the REM protein design described in section 3, two sequences are optimal solutions to the folding protein if they have the same energy. To this end, the Caterpillar algorithm optimizes the energy function by simultaneously designing over 120 test proteins and comparing the designed and the natural sequences. The simulation converges when the design and the natural sequences have matching Caterpillar energies and hydrophilic/phobic profiles.
Given that the native sequence is nature´s solution, the Caterpillar interaction matrix can be viewed as the one by which the natural and designed sequences are equivalent solutions to the inverse folding problem.

The uniqueness of such an approach is that it uses protein design instead of protein folding to predict the structural properties of proteins quantitatively.

It is important to stress that the same methodology can be used to fit a larger spectrum of available experimental data (e.g. iso-electric point, physiological pH) or even other force fields such as ROSETTA described in section 2.

Description of the interaction optimization algorithm

Given a set of N_{Prot}single-domain proteins, for each protein, an ensemble of N_{Seq} sequences are generated. Hence the probability $P(S_i,\Gamma_j)$ of having a sequence $S_i$on a structure Γ_j is given by the Boltzmann weight:

$$P(S_i,\Gamma_j) = \frac{e^{-\beta H(S_i,\Gamma_j)}}{\sum_i^{N_{Seq}} e^{-\beta H(S_i,\Gamma_j)}}, \qquad (16)$$

where H is the Caterpillar force field Hamiltonian.

The objective is to determine the parameters of the force fields by simultaneously designing the N_{Prot} proteins and comparing the N_{Seq} generated sequences with the natural one and select the parameters that give the best match. According to the maximum entropy principle, the optimal values for the parameters are found by maximizing the entropy S

$$S = -\sum_j^{N_{Prot}} \sum_i^{N_{Seq}} P(S_i,\Gamma_j) lnP(S_i,\Gamma_j) \qquad (17)$$

associated with the distribution $P(S_i,\Gamma_j)$. The maximization procedure can the constrained by using the method of Lagrange multipliers, each associated with a given fitness function. The optimal matrix corresponds then to the extremal of the function Λ defined as follows:

$$\Lambda = S + \sum_{j}^{N_{\text{Prot}}} \sum_{k}^{N_j} \lambda_{jk} \left(\sum_{i}^{N_{\text{Seq}}} P(S_i, \Gamma_j) \alpha^i_{jk} - \alpha^{\text{Real}}_{jk} \right) + \sum_{j}^{N_{\text{Prot}}} \sum_{k}^{N_j} \lambda'_{jk} \left(\sum_{i}^{N_{\text{Seq}}} P(S_i, \Gamma_j) E^i_{jk} - E^{\text{Real}}_{jk} \right) + \sum_{j}^{N_{\text{Prot}}} \gamma_j \left(Z_j - 1 \right) \tag{18}$$

Here, λ_{jk}, λ'_{jk} and γ_j are the Lagrange multipliers associated with the HP nature of the amino acids α_{jk} , the total energy of the sequences E_{jk} and the normalization condition $Z_j = \sum_i^{N_{Seq}} P(S_i,\Gamma_j) = 1$.

According to the Euler-Lagrange method, the maximum of the function Λ will correspond to the maximum of the entropy S under the constraints imposed on the system. Hence, we can perform the derivative of Λ with respect to $P(S_i,\Gamma_j)$ keeping the Lagrange multiplier constant and equate the derivative with 0.

$$\frac{d\Lambda}{dP(S_i,\Gamma_j)} = 0 \tag{19}$$

From the maximization, we collect independent relationships for all the Lagrange multipliers. For instance, for the α parameters. We get:

$$\begin{aligned} \frac{\partial \Lambda}{\partial \lambda_{jk}} &= \frac{1}{Z'_j} \frac{\partial Z'_j}{\partial \lambda_{jk}} - \alpha^{\text{Real}}_{jk} \\ &= \frac{1}{Z'_j} \sum_{i}^{N_{\text{Seq}}} \alpha^i_{jk} e^{\sum_k^{N_j} \lambda_{jk} \alpha^i_{jk}} - \alpha^{\text{Real}}_{jk} = 0 \end{aligned} \tag{20}$$

Eq. 20 implies that the distribution generated by the Lagrange multiplier that makes the average hydrophobic/hydrophilic profile equal to the natural one also maximizes the entropy.

Hence, the best model is the one with the parameters that make the natural and artificial sequences have the energy and the hydrophobic/hydrophilic profiles as similar as possible.

4.3. Tube Models

In 2000, Maritan and co-workers [209–211] introduced the “Tube” protein model, where a typical protein is represented as a flexible self-avoiding tube with a radius of ~2.5Å and effective hydrogen bonds interactions along the tube. The configurations of the tube model are controlled by just two parameters, the total hydrophobicity and the bending rigidity. The model then reproduced all secondary and many known protein tertiary structures by local changes in the two model parameters. Hence, the results obtained with the tube model strongly suggest that the typical protein structures are inherent in the geometrical constraints of the backbone, as the latter are the main features of the tube model. To put in the words of the authors, the tube "pre-sculpts`` the free energy landscape. Recently their findings have been further expanded by Kukic et al. [208], who demonstrated how their ``CamTube" model could map the protein structural space. More recently, Škrbić et al. have shown how the symmetry breaking created by the side chain along a polymer backbone can also induce a collapse of the configurational space into sub-space with helices and beta sheets [216,217].

4.4. Martini

The Martini force field has gained popularity for its applications in protein simulations and materials science [218,219]. This force field, developed by the Marrink group [213,214], provides an effective way of simulating the behaviour of a wide range of molecules.[222] Their lipid and protein parameterizations have given the opportunity of simulating membrane proteins in large simulations [223,224]. The scale of these simulations, almost reaching 100 nm, has granted the term of computational microscopy and has offered a unique view of the dynamic behaviour of membranes and the proteins embedded in them [225,226]. As well as lipids and proteins, the force field currently includes parameters for other molecules present in membranes such as sterols, [227] carbohydrates,

[228] glycolipids, [229] and photosynthesis cofactors, [230] in addition to molecules that display interesting behaviours in membranes, with numerous contributions from other groups that have helped to extend the parameter library [231–233]. DNA and RNA complete the list of available biomolecular parameters, allowing for studying complex biological systems.[234,235] The Martini scope has been expanded into materials science with excellent results in peptide self-assembly, [236–238] peptoids mesoscale behaviour, [239] polymers dynamics, [240,241] organic semiconductor layers formation, [242] and ionic liquids phase studies [243,244].

Although the coarse-grained resolution, with a bead representing 2 to 5 heavy atoms, has been vital for the efficiency of Martini to afford such simulation size and times, the development of the polarized version has helped in increasing the accuracy to represent specific interactions, such as cation-π, of great interest for proteins [245–247]. Martini has also been employed in mixed resolution methodologies combined with all-atoms to gain accuracy of the interactions in lipid bilayers.[248] Additionally, this force field has been combined with highly coarse-grained bilayers using dynamically triangulated surfaces to achieve the semi-atomistic resolution of Martini in a whole mitochondria simulation.[249]

However, on its website, Martini's team explicitly states that this force field cannot be used to model protein folding, despite its success with small peptide self-assembly. The mapping of proteins into Martini resolution, or Martinizing of proteins, requires the input of the secondary structure tuning the bonded and non-bonded parameters to preserve it. To maintain the 3D structure of proteins, Martini often needs to be combined with elastic potentials between Cα within a threshold called the ElNeDyn model [250]. Therefore, the input structure is too rigid to reproduce unfolding events. In 2017, Poma et al. overcame this limitation by substituting the harmonic potentials with Lennard-Jones interactions using the contact map of the native state in protein, similarly to Go-models.[251] The Martini team seems to have adopted this idea for its version 3, stating in its open beta version documentation that they improve protein flexibility using Go-models. Although it is still unclear to which extent these new interactions will improve the model towards studying protein unfolding, the latest version has already shown some advances in protein structure and protein-ligand events. The beta version has been employed for high throughput protein-ligan binding, improving the modelling of protein cavities and binding pathways to assess the effects of mutations on the binding of different small drugs.[252] They claim that their coarse-grained approach is similarly effective and more efficient than the corresponding atomistic approaches. In addition to this, Grunewald et al. have recently published the Martini approach for constant pH simulations, with excellent results reproducing experimental pKas.[253]

5. Application of coarse-grained models

This section highlights applications of coarse-grained models trying to answer fundamental questions related to protein evolution. Due to the timescale and size of the protein sequence space, coarse-grained models represent an ideal investigation tool.

5.1. Role of the alphabet

The amino acids are the building blocks of proteins, whose chemical diversity in a sequence is responsible for many three-dimensional structures and biological functions, playing a crucial role in the protein sequence evolution.

The protein sequence is typically noted as a string of letters to represent each amino acid. The protein alphabet contains 20 different characters for the amino acids, unlike DNA and RNA, consisting of 4 letters.

An important issue that attracts the interest of the scientific community is the nature of the amino acid alphabet [1,7,26–29,183–185,254–281] and, in particular, the effects of a reduced alphabet size on protein folding. Previous studies applied different computational methods for the protein design at different alphabet sizes. Using lattice protein models, a large variety of protein-like heteropolymers were designed at different alphabets [1,7,183–185,255–259]. From those studies emerged that a minimum number of residue types is required to get target configurations [260]. It was also possible to investigate the effect of a minimalistic alphabet on protein-protein interactions [261–264]. Also, experimental works were conducted by designing proteins with simplified amino acid sequences [265–269]. Statistical analysis of protein databases also showed that a large part of the information, [254,270–275] encoded in natural proteins, could be enclosed into a small alphabet of only 5 residues types [254,265,267,276,277,281].

Nerattini *et al.* devise a computational protein design strategy that consists of a competition for available amino acids between a protein and an artificial interaction partner. No previous studies have considered the possibility of competition for the availability of amino acids. However, lack of materials may have played an essential role in the evolution of protein alphabets. Hence, it is interesting to estimate the effect of such competition.

Nerattini's scheme spontaneously drives the protein design to the generation of sequences with a reduced number of residue types. Moreover, the reduced alphabets chosen during the design process allows for the folding stability of the protein. The investigation results show that for the folding of a protein, the minimum size of the amino-acid alphabet is just 4 letters. The results have interesting parallelism with the 4-letter alphabet of RNA, which is considered the precursor of proteins during the early stage of life. However, the precision of the folding increases with the alphabet size: 6 letters are the minimum alphabet necessary to maintain the structure of the protein with the same accuracy commonly obtained with 20 letter alphabets. The observation is consistent with the experimental studies confirming that 6 letters are essential for maintaining protein folding and functionality. [254,265,267,276,277,281]

Besides having a binary system, the authors investigate how the alphabet reduction affects the heterogeneity of protein-protein interaction [1,184,262,264,282], observing a strong tendency of the designed protein to absorb and aggregate on a potential binding site. The 4 letters alphabet of the designed sequences has an average intra-protein residue interaction higher than the inter-protein interaction energy. This affinity makes it impossible for the folded state of the protein to be stable in contact with the artificial partner; hence, to avoid the absorption. Conversely, increasing the alphabet size to 6 letters, the intra-protein residue interaction stabilizes the folded structure upon binding due to its lower value with respect to the inter protein one. Living systems are under constant pressure for using the least variety of amino acids to reduce the resources necessary to construct specialised tRNA molecules for the translation process.[283] It is reasonable to assume that it could be advantageous to design proteins with a smaller alphabet during the early stages of life. Thus, it suggests that the optimization of the specificity of protein-protein interactions could have been the driving force for the evolution of the large protein alphabet.

1.1 Protein Design as a tool to test evolution constraints

The rate of Protein sequence evolution varies from protein to protein, and several factors such as the processing of the protein in the cell (*e.g.,* translation time) [284,285], or molecular characteristics specific to each protein [197,286,287], as well as from interactions with other proteins [288]. In contrast, the nature and rate of protein structural evolution are much less well understood. Viksna *et al.*[289] presented an estimate of the rate of structural changes based on the measure of topological distances between proteins structures. Meyerguz et al. [290] grouped all known proteins into basins corresponding to the common native structures. The authors have then built a network of sequences from the collected data and considered the frequency of "transition" sequences (separated by a single point mutation from a different basin). Structural evolution has also been studied in the context of the lattice protein model by Deeds et al.[196], where the structural similarities among all possible 103346 distinct structures of a 3x3x3 lattice polymer have been mapped. Other work has concentrated on structural topologies connected by a relatively small set of structural evolutionary moves (e.g. domain swapping or duplications) [154,197,286].

Coluzza et al. [291] considered the entire evolutionary process without focusing on a detailed description of cell physiology. In that case, the evolutionary process is equivalent to screening a large number of different sequences under the constraint that only a few structures are acceptable. The full evolutionary path can then be represented as a transition sequence between the allowed structures (steppingstones). Such steppingstones represent the possible structures that are still allowed by the selection function and are not identical to the initial and final target structure. The number of intermediate structures reflects the degree of restriction applied to the evolutionary process. Hence the larger the number of steppingstones, the more closely the evolutionary process approximates a free drift in protein space. The entire evolutionary trajectory between two targets is then represented as a path connecting the steppingstones, where each jump is weighted by its probability of occurrence. Accordingly, the main objective of Coluzza's work is to measure the rate of each elementary jump and identify the analytic dependence of such rates from a small set of structural differences.

The first point it is vital to realize is that the number of sequences that can fold into a structure is an astronomically large number [149].

The objective is to sample the rate at which an ensemble of sequences defined by the design procedure with target structure ***A*** will evolve to an equivalent ensemble defined by the design of structure ***B***.

First, the overlap between the most probable sequences of ***A*** and ***B*** is minimal, independently of the structural differences between ***A*** and ***B***. In other words, provided that the structures are not identical, the Hamming distance between the ensemble of the folding sequences is always sizeable. This gap does not necessarily mean that the evolutionary process must proceed with large jumps with many concurrent mutations. Still, it means that the folding sequences in "common" (so with small Hamming distance) between the two distributions are pretty rare. Hence the evolutionary rate is highly dependent on the probability of finding such sequences that are still able to fold but are separated by a small number of mutations. For this reason, the neutral evolution inside each island is assumed to occur at a higher rate than it does between islands.

According to such a hypothesis, the evolution rate is defined as the rate of crossing the point at which a sequence goes from having lower total energy in structure ***A*** to having lower energy in ***B***. This choice can be justified as a measure of the propensity of those sequences to fold into ***B*** instead of ***A*** because of the entropic contribution to the free energy of the native structure is assumed to be the same across all steppingstones, then the only relevant pressure is the energetic contribution. The probability of observing such a sequence can then be measured using the Boltzmann distribution function in the space of all possible proteins (all sequences on all structures);

$$R_{A\to B} = \langle \theta[\Delta \mathrm{E}_{AB}] \rangle_A = \frac{\left\langle e^{\beta E_B} \theta[\Delta \mathrm{E}_{AB}] \right\rangle_{AB}}{\left\langle e^{\beta E_B} \right\rangle_{AB}} \qquad (21)$$

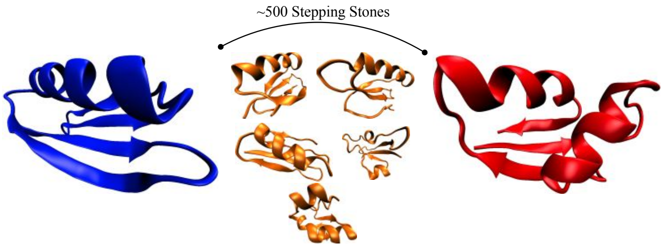
where the ensemble average $\langle \ldots \rangle_{AB}$ is performed over the ***AB*** joined ensemble. Alternatively, the equation can be interpreted as a simulation in the ensemble of sequences that fold into structure ***A*** but in the presence of a bias towards sequences that fold into structure ***B***.

Each rate is then sampled by applying the design procedure described above to the joined ***AB*** ensemble for each ***A***, ***B*** pair with the following acceptance rule

$$P_{\mathrm{acc}} = \min\left\{1, \exp\left[-\frac{\left(\Delta E_{A+B} - E_p \ln \frac{N_P^{\mathrm{new}}}{N_P^{\mathrm{old}}}\right)}{k_B T}\right]\right\}. \qquad (22)$$

Such an acceptance rule also guarantees that homopolymers sequences are not included in the rate calculations that might significantly alter the results towards non-physical solutions with their significant enthalpic weight.

Hence the jumping rate from the island associated with structure ***A*** to ***B*** is going to be equal to the rate of accumulating enough mutations for each sequence of the island of ***A*** to become equal to one of the sequences in the island of ***B***, as the evolutionary process will spontaneously continue towards the optimal sequences of ***B*** at a much faster rate.



Such a rate can be calculated efficiently and allows for a large-scale study of jumps across many structures.

By putting together all $R_{A \to B}$ measured for 490X490 structure pairs, the rate is well described as a function of three structural parameters that measure the difference between structures ***A*** and ***B***: the difference in the number of hydrogen bonds ΔH_{AB}, the difference in the number of residue-residue parameters ΔQ_{AB} and the difference in the number of native contacts Q_N.

$$\begin{aligned} \ln R_{A \to B} = 151 \ln\left(\frac{1}{1+e^{0.005(7.2\Delta H_{AB} - \Delta Q_{AB})}}\right) \\ +222 \ln\left(\frac{1}{1+e^{-20.5(0.5-Q_N)}}\right) \end{aligned} \qquad (23)$$

In particular, this expression demonstrates that it is much easier to jump towards a compact structure with many hydrogen bonds than evolve towards a configuration that is either compact with few hydrogen bonds or non-compact with many hydrogen bonds.

A result that comes naturally from our analysis is the probability of occurrence of a structure, which can also be interpreted as the designability of a protein structure.

$$P_i = \frac{e^{-A_2 A_0 (A_1 H_i - Q_i)}}{\sum e^{-A_2 A_0 (A_1 H_i - Q_i)}} \qquad (24)$$

That is a crucial result of this study. The designability of a protein does not depend just on how compact it is but mainly on the optimization of both the number of hydrogen bonds and the number of contacts between the residues.

This result again highlights the vital role those directional interactions play in the designability of proteins and heteropolymers in general.

1.2 Protein-Protein Interactions

Protein-protein recognition is one of the multiple types of molecular recognition tools that nature employs and, as it is involved in countless physiological processes, is crucial for living beings [292,293]. Synthetic systems, such as polymers, have also copied this mechanism, giving rise to artificial molecular recognition [294–302].

Molecular recognition requires highly specific binding with a high discriminatory resolution. In other words, the molecules must bind strongly to a minimum number of possible partners and weakly, if anything, with the rest. The design of binding sites introduces constraints to ensure a strong and specific interaction. Protein binding sites are in the range of 75 – 150 nm, [303] and often fit the ligand tightly. Therefore, the selectivity of protein-ligand recognition lies in both steric compatibility and chemical patterning of the pocket surface. Coluzza et al. designed patterned surfaces to bind a reduced number of partners selectively using a lattice model. [1,184] They showed that by designing the ligands in the bound state, the selectivity of the binding to the target surface is boosted. This result is based on the probability (P) of non-specific interactions for having a binding energy (E):

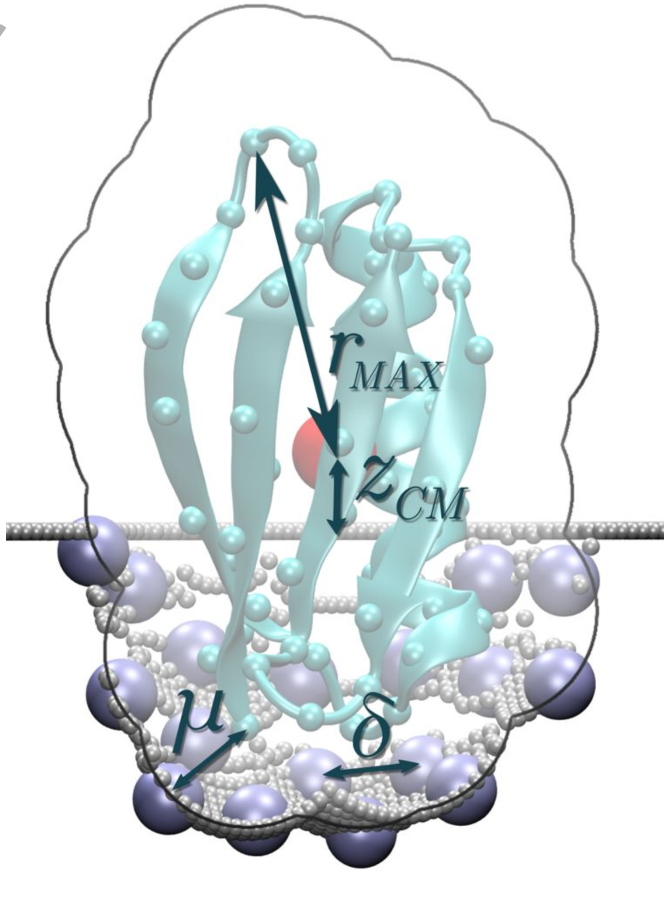


Figure 8: Schematic representation of the parameters used to generate the pocket moulds.

$$P(E) = \left(2\pi N\sigma^2\right)^{-\frac{1}{2}} e^{-\left[\frac{E^2}{2N\sigma^2}\right]} \qquad (25)$$

Where N is the number of interaction sites to account for the size of the binding. The Boltzmann factor $exp\,(-\beta E)$ gives the probability of an interaction energy E in the bound state. Consequently, to be selective, surfaces must have a binding energy lower than the random average Boltzmann factor, $\langle exp(-)\rangle = exp\left(N\sigma^2\beta^2/2\right)$. Additionally, random binding sites are not strictly inert as they will still have a relevant probability to bind if they are sufficiently large (great N).

Nerattini *et al.* [206] employed the Caterpillar protein model [22,24] to explore pockets' precision and binding selectivity with optimal shape and poor steric selectivity. They conducted this study attending to hot spots at the protein-protein interface, which are currently recognized as a critical component for Protein-Protein Interactions (PPI). [304] They did not aim to reproduce PPI quantitatively and could afford to use a coarse-grained model with implicit solvent, which is inappropriate to identify hot spots. Instead, they did examine the steric effect of certain features of the binding sites, such as depth and surface area. They carried out the design of a given protein with a second target protein by modelling the binding region of the latter on a plane. The explicit protein partner was here modelled with the mentioned Caterpillar model, as described in a previous section. The protein-like surface was constructed as a mould by pushing the protein on a dense flat mesh of self-avoiding beads, which mimic the portion of interest of the protein surface. This approach allows controlling the direction of the interaction. Binding site interactions were modelled using only the C_α of the Caterpillar. A certain number of beads scattered within the mesh mimicking the protein surface are conferred Caterpillar C_α character. The model is based on three parameters (Figure 8): ζ, the height of the centre of mass (CM) with respect to the flat mesh plane; μ, minimum C_α protein–Cαsurface distance; and δ, the distance between beads with Cα character in the binding site. Binding sites were generated by setting the last two parameters to typical natural values in globular proteins (both to 5 Å) and varying ζ. Firstly, the maximum CM–C_α distance, corresponding to the entire protein radius (r_{MAX}) was determined to normalize the rest of the CM–Surface distances. Thus, being z the CM–Surface distance, $\zeta = \frac{z}{r_{MAX}}$. For each value of ζ, the flat mesh was tuned to represent each protein orientation to find the orientation that gives a binding site with maximum surface area. It must be noticed that the surface area of the binding site is inversely proportional to ζ.

The distance root mean square displacement (DRMSD) was used as an order parameter for the bias potential, measuring the deviation from the target structure:

$$DRMSD = \sqrt{\frac{1}{C}\sum_{ij}\left(\left|\overrightarrow{\Delta r_{ij}}\right| - \left|\overrightarrow{\Delta r_{ij}}^{T}\right|\right)^2} \qquad (26)$$

Where $DRMSD$ it is calculated as the sum over the *ij* contact pairs in the structure between residues in the same ($DRMSD_{intra}$) or different ($DRMSD_{inter}$) proteins. $\overrightarrow{\Delta r_{ij}}$ is the distance between the pairs, while $\overrightarrow{\Delta r_{ij}}^{T}$ is the corresponding distance in the target structure. This differs from most protein approaches where the RMSD is used instead, using the atom positions rather than distances. The system conformational space was projected over the collective variables $DRMSD_{intra}$ and $DRMSD_{inter}$ generating the free energy landscape $F[DRMSD_{intra}, DRMSD_{inter}]$. $F[DRMSD_{intra}, DRMSD_{inter}]$ can qualitatively show the relative stability between folded and unfolded in bound and unbound states. The profiles show that although the size of the binding site affects the strength of the binding, all the proteins can bind in their folded state to their target binding site, including the small ones.

To quantify the binding affinity and selectivity, the authors measured the free energy difference ΔF between the bound and unbound of the folded. This free energy difference is defined by:

$$\Delta F = -k_B T ln\left(\frac{Q_b}{Q_f}\right) \qquad (27)$$

Qb accounts for the bound protein conformations and Qf for the unbound, free in the bulk. $exp\left(-\frac{\Delta F}{k_B T}\right)$ defines the binding strength, leading to an association constant that follows the expression:

$$K_a = exp\left(-\frac{\Delta F}{k_B T}\right)\frac{V_{bulk}}{n} \qquad (28)$$

Being n the number of binding sites, which was set to 2 in the example and V_{bulk} the volume of the bulk.

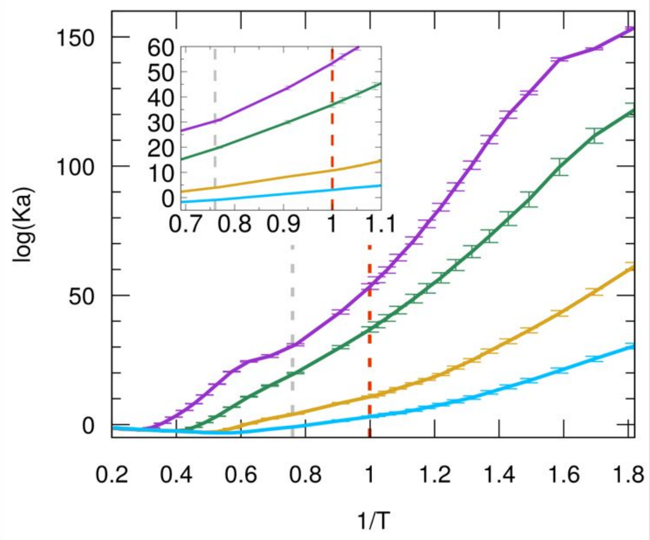


Figure 9: Van't Hoff plot of the binding affinity K_a [l/mol] as a function of the inverse of reduced temperature 1/T for the investigated systems. The grey dashed line shows the folding temperature $T/T_F = 1$. The red dashed line is the reference ambient temperature $T/T_A = 1$ in reduced units. The curves' colour scheme refers to the pockets' size ζ going from large to small: purple, green, yellow and light blue.

Figure 9 shows the van't Hoff plot [305,306] of the binding affinity K_a for the different pocket sizes.

The results showed that binding site surfaces decreased with ζ, the topology matching between the protein and the surface creates an effective pattern of steric repulsion, key for the binding site selectivity.

The specificity of the binding sites towards their target was tested for the artificial binding sites employing different scenarios. Firstly, by crossing proteins and surfaces resulting from different ζ values, we tested the selectivity among proteins with different sequences but identical structures. Secondly, the folding and binding of a protein with different structures but similar sizes were tested. The first scenario showed the differences between small and large binding sites. The former showed negligible binding to large proteins, while the wider binding sites showed stronger binding and a disruptive effect in protein structure, leading to denaturation. The second scenario confirmed the lower specificity of large binding.

Therefore, this work presented an attractive approach for designing protein-protein interactions. Nerattini et al. designed specific sequences for target binding sites. The fact that the folded bound state is favoured in the resulting sequences and their binding energy increases with the size of the pocket is evidence of the approach's success to design protein-protein interactions. Additionally, the results shine a light on the specificity of the pockets, showing that large binding pockets have higher binding affinities, they also show lower specificity. The upper limit determined by the model matches with the size range of binding sites of natural proteins. Therefore, this method is an efficient approach to designing protein-protein interactions and provides fundamental information for understanding natural proteins and how specific parameters may have affected their evolution.

1.3 Compare artificial and natural sequences

Protein sequence maintains a delicate balance between structural stability and biological function, making it difficult to untangle the two contributions. It has been proved that a protein function, such as the catalytic activity of an enzyme, depends on the interaction between specific sequence positions and exhibits a balance between structural stability and flexibility. Also, It is challenging to classify residues as strictly functional or structural due to a correspondence between these two categories; their mutual correlation is essential for the protein activity. [307,308] It is meant by strictly structural residues such as amino acids responsible for protein stability. The loss of the folded structure can affect the functionality of the protein.

On the other hand, strictly functional residues can mutate without altering the structure's stability. An accurate characterization of structural and functional protein residues is fundamental for developing proteome mapping, protein engineering, and new pharmaceutical applications based on the design of target protein.[309–313] The experimental identification of residues is a time-consuming and expensive process: a high-throughput tool requires a large scale mutation assay [308,309], whereas *in-silico* screening has a lower cost. Several Computational methods [316–320] have been developed for studying protein evolution. Most of them are based on the search for sequence conservation and co-evolution.

The residues co-evolution assumes that mutations of interacting amino acids are correlated. Co-evolution allows proteins to change residue identities while maintaining specific residue-residue interactions [19,321]. The residues involved in co-evolution events can be fundamental for the protein activity (*e.g.,* catalytic site residues) and for the structure stability (*e.g.,* hydrophobic core residues), or, in some instances, for both, when there is an interdependence between functional and structural residues. The Direct Coupling Analysis (DCA)[322–330] is one of the most promising computational tools for estimating residues pairs with direct reciprocal constraints in the evolution. The method for protein contact prediction is based purely on sequence information and can analyse a large number of protein domains. However, from DCA alone is not possible to distinguish between structural and functional residues due to the same signal given by the two types of coevolving residues during the analysis. Some information can be deduced from comparing the DCA and the distance between residues in the contact map [319,330–332]. But functional residues do not have always have long-range co-evolution signals.

Searching for amino acid sites of a protein sequence that preserve their identity in the evolutionary residue conservation (or site entropy) analysis is another method for identifying functionally essential protein regions. The evolutionary site conservation can be measured using Casari et al. technique.[333], based on the principal component analysis (PCA) of the sequence alignments.

Nerattini *et al.*[334] introduced a methodology to rank the residues according to their functional (F) or structural (S) nature within the ones that are involved in both events (OFSR, overlapping functional, structural residue [307]).

Their methodology hypothesises is that an artificial evolution process only results in a co-evolutionary structural residue due to the absence of any functional constraints.

Thus, to identify residue and further categorize them into structural, functional or OFSR, it is necessary to generate an artificial protein family that, by construction, contains only structural information. Any protein design method can generate artificial sequences with a specific target conformation [5,6,17,20,21,108,335–339]. The design doesn’t need to generate lab folding proteins. The only requirement is that the artificial sequences fold computationally into the target structure.

After selecting the protein family to analyse, single-site conservation and co-evolution analysis are carried out on artificial and natural alignments. Protein design generates artificial sequences, whereas natural sequences are found in the Pfam database.[316]

The analysis of artificial sequences identifies residues essential for structural stability; on the other hand, signals from natural sequence analysis encode structural and functional information. Residues with high co-evolution signals only in the natural alignments are residues with a functional role in the protein if a similar signal is not present in the analysis of the artificial set. Conversely, structural signals are strongly conserved and co-evolved in the artificial evolution but poorly in natural ones. Residues that display comparable signals between natural and artificial analysis are classified as overlapping-functional-structural residues OFSR, whose mutation would lead to the loss of both functionality and tertiary structure.

DiPA methodology has demonstrated the validity to detect functional residues in protein families without requiring prior knowledge of the biological role of the analysed protein. Hence, in the study of a whole proteome, the DiPA algorithm could give a crucial contribution to the identification of the functional protein regions. By analysing the artificial evolution of protein dimers, the approach can also classify functional residues for the implication of protein-protein interactions, confirming the annotation mentioned above on the direct importance of the structural residues on the protein's function.

Conclusions

Computational Protein design is one of the most promising tools in protein engineering. The long-term objective is to autonomously design new artificial enzymes and drugs with sequences tailored to specific functions and perform better than their natural contour parts. Additionally, protein design offers an ideal benchmark tool to test fundamental hypotheses about the evolution of life's basic building blocks.

In this review, we tried to overview both basic and applied protein designs. The challenges ahead are still many. Although successful in many applications, it is still tough to systematically design proteins with high expression yields that vary a lot from application to application. The reason for such difficulties can be found both at the algorithm level (e.g. sampling), modelling (e.g. accuracy), and fundamental understanding of the central ingredients for successful design.
In terms of algorithms, essential developments are coming from multi-scale approaches mixing coarse-graining and full-atomistic representations and the introduction of deep learning methods like the recent AlphaFold [340]. On the modelling side, it is essential to stress the emerging importance of constant pH simulations that take into account the charge fluctuations that occur on the protonable end of polar amino acids. Constant pH simulations are still growing, and there is not yet a single established method to perform them. However, many studies indicate that they are strategic in understanding protein-protein interaction phenomena [341] and hence for design [342].

Furthermore, protein design has the potential to push the development of parallel fields such as supramolecular peptide polymers. These materials exploit the tendency of small peptides to self-assemble into protein-like structures driven by similar rules to proteins themselves. Some efforts have been carried out in modelling the behaviour of these materials using molecular dynamics simulations. Tuttle *et al.* screened short peptides using the MARTINI force field to find new self-assembling sequences [236,343]. However, these had computing limitations that drove them to combine this with machine learning to screen peptide sequences consisting of up to 8 amino acids [344]. Ferguson et al. also employed this approach on a hybrid system [345]. Although machine learning has significantly reduced the computational effort of these procedures, these methods are far from the level of validations and efficiency of protein design. We believe that using a modular approach like the one employed for repeat proteins [122], protein design methodologies could be applied to self-assembling peptides, which would boost the development of these synthetic materials.

Finally, on the fundamental understanding of the relation between protein folding and protein design, we have stressed the physical role of directional interaction in sculping the conformational landscape. A landscape that can only be defined if a proper length scale is introduced to discriminate between conformations. Such length scale is nothing else than the target folding resolution. With such knowledge, it is possible to extend protein design beyond the biological kingdom to venture into the unknown, mimicking life, fully synthetic materials. That will be the era of bionic proteins.

Acknowledgements

I.C. gratefully acknowledges support from the Ministerio de Economià y Competitividad (MINECO) (FIS2017-89471-R). I. R. S. acknowledges financial support from Gipuzkoa Foru Aldundia (Gipuzkoa Fellows Program, Diputacion Foral de Gipuzkoa: 2019-FELL-000017-01). This work was performed under the Maria de Maeztu Units of Excellence Program from the Spanish State Research Agency – Grant No. MDM-2017-0720. This research was supported by Programa Red Guipuzcoana de Ciencia, Tecnología e Información 2019-CIEN-000051-01. We acknowledge support from the BIKAINTEK program (grant No. 008-B1/2020).

References

[1] Coluzza I and Frenkel D 2004 Designing specificity of protein-substrate interactions

Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. **70** 51917

[2] Coluzza I, van Oostrum P D J, Capone B, Reimhult E and Dellago C 2013 Sequence controlled self-knotting colloidal patchy polymers *Phys. Rev. Lett.* **110** 075501

[3] Coluzza I 2015 Constrained versus unconstrained folding free-energy landscapes *Mol. Phys.* **113** 2905–12

[4] Rubenstein B M B M B M B M, Coluzza I and Miller M M A M A M A M A 2012 Controlling the folding and substrate-binding of proteins using polymer brushes *Phys. Rev. Lett.* **108** 1–5

[5] Shakhnovich E I 1994 Proteins with selected sequences fold into unique native conformation *Phys. Rev. Lett.* **72** 3907–10

[6] Gutin A M and Shakhnovich E I 1993 Ground state of random copolymers and the discrete random energy model *J. Chem. Phys.* **98** 8174–7

[7] Bryngelson J D D and Wolynes P G G 1987 Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **84** 7524–8

[8] Frauenfelder H, Sligar S and Wolynes P 1991 The energy landscapes and motions of proteins *Science (80-.).* **254** 1598–603

[9] Bryngelson J D, Onuchic J N, Socci N D and Wolynes P G 1995 Funnels, pathways, and the energy landscape of protein folding: A synthesis *Proteins Struct. Funct. Bioinforma.* **21** 167–95

[10] Onuchic J N, Luthey-Schulten Z and Wolynes P G 1997 Theory of Protein Folding: The Energy Landscape Perspective *Annu. Rev. Phys. Chem.* **48** 545–600

[11] Dahiyat B I and Mayo S L S L S L L 1997 De Novo Protein Design: Fully Automated Sequence Selection *Science (80-.).* **278** 82–7

[12] Sevy A M, Jacobs T M, Crowe Jr. J E, Meiler J, Crowe J E and Meiler J 2015 Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences ed B Peters *PLOS Comput. Biol.* **11** e1004300

[13] Pelay-Gimeno M, Glas A, Koch O and Grossmann T N 2015 Structure-Based Design of Inhibitors of Protein-Protein Interactions: Mimicking Peptide Binding Epitopes *Angew. Chemie - Int. Ed.* **54** 8896–927

[14] Chevalier A, Silva D-A, Rocklin G J, Hicks D R, Vergara R, Murapa P, Bernard S M, Zhang L, Lam K-H, Yao G, Bahl C D, Miyashita S-I, Goreshnik I, Fuller J T, Koday M T, Jenkins C M, Colvin T, Carter L, Bohn A, Bryan C M, Fernández-Velasco D A, Stewart L, Dong M, Huang X, Jin R, Wilson I A, Fuller D H and Baker D 2017 Massively parallel de novo protein design for targeted therapeutics *Nature* **550** 74–9

[15] Marcos E, Basanta B, Chidyausiku T M, Tang Y, Oberdorfer G, Liu G, Swapna G V T, Guan R, Silva D-A, Dou J, Pereira J H, Xiao R, Sankaran B, Zwart P H, Montelione G T and Baker D 2017 Principles for designing proteins with cavities formed by curved β sheets *Science (80-.).* **355** 201–6

[16] Bianco V, Franzese G, Dellago C and Coluzza I 2017 Role of Water in the Selection of Stable Proteins at Ambient and Extreme Thermodynamic Conditions *Phys. Rev. X* **7** 21047

[17] Coluzza I 2017 Computational protein design: a review *J. Phys. Condens. Matter* **29** 143001

[18] Koehl P and Levitt M 1999 De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.* **293** 1161–81

[19] Kortemme T and Baker D 2004 Computational design of protein-protein interactions *Curr. Opin. Chem. Biol.* **8** 91–7

[20] Fung H K, Welsh W J and Floudas C A 2008 Computational de novo peptide and protein design: Rigid templates versus flexible templates *Ind. Eng. Chem. Res.* **47** 993–1001

[21] Samish I, Macdermaid C, Perez-Aguilar J and Saven J 2011 Theoretical and computational protein design *Annu. Rev. Phys. Chem.* **62** 129–49

[22] Coluzza I 2011 A Coarse-Grained approach to protein design: Learning from design to understand folding *PLoS One* **6** e20853

[23] Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton T B B, Montelione G T T and Baker D 2012 Principles for designing ideal protein structures *Nature* **491** 222–7

[24] Coluzza I 2014 Transferable coarse-grained potential for de novo protein folding and design. ed Y Zhang *PLoS One* **9** e112852

[25] Thomson a. R, Wood C W, Burton a. J, Bartlett G J, Sessions R B, Brady R L and Woolfson D N 2014 Computational design of water-soluble -helical barrels *Science (80-.).* **346** 485–8

[26] Davidson A R and Sauer R T 1994 Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci.* **91** 2146–50

[27] Riddle D S, Santiago V J, Bray-Hall S T, Doshi N, Grantcharova V P, Yi Q, Baker D, Santiago J V., Bray-Hall S T, Doshi N, Grantcharova V P, Yi Q and Baker D 1997 Functional rapidly folding proteins from simplified amino acid sequences *Nat. Struct. Biol.* **4** 805–9

[28] Cordes M H J, Davidson A R and Sauer R T 1996 Sequence space, folding and protein design *Curr. Opin. Struct. Biol.* **6** 3–10

[29] Davidson A R, Lumb K J and Sauer R T 1995 Cooperatively folded proteins in random sequence libraries *Nat. Struct. Biol.* **2** 856

[30] Huang P-S, Boyken S E and Baker D 2016 The coming of age of de novo protein design *Nature* **537** 320–7

[31] Parmeggiani F and Huang P-S 2017 Designing repeat proteins: a modular approach to protein design *Curr. Opin. Struct. Biol.* **45** 116–23

[32] Baran D, Pszolla M G, Lapidoth G D, Norn C, Dym O, Unger T, Albeck S, Tyka M D and Fleishman S J 2017 Principles for computational design of binding antibodies *Proc. Natl. Acad. Sci.* **114** 10900–5

[33] Mejias S H, Lopez-Andarias J, Sakurai T, Yoneda S, Erazo K P, Seki S, Atienza C, Martin N and Cortajarena A L 2016 Repeat protein scaffolds: ordering photo- and electroactive molecules in solution and solid state *Chem. Sci.* **7** 4842–7

[34] Cortajarena A L, Liu T Y, Hochstrasser M and Regan L 2010 Designed proteins to modulate cellular networks *ACS Chem. Biol.* **5** 545–52

[35] Mejias S H, Aires A, Couleaud P and Cortajarena A L 2016 Designed Repeat Proteins as Building Blocks for Nanofabrication *Advances in Experimental Medicine and Biology* vol 940, ed A L Cortajarena and T Z Grove (Cham: Springer International Publishing) pp 61–81

[36] Bianchi E, Capone B, Coluzza I, Rovigatti L and van Oostrum P D J D J 2017 Limiting the valence: advancements and new perspectives on patchy colloids, soft functionalized nanoparticles and biomolecules *Phys. Chem. Chem. Phys.* **19** 19847–68

[37] Sorenson J M and Head-Gordon T 2000 Matching simulation and experiment: A new simplified model for simulating protein folding *Journal of Computational Biology* vol 7 (Mary Ann Liebert, Inc., publishers) pp 469–81

[38] Song Y, Dimaio F, Wang R Y R, Kim D, Miles C, Brunette T, Thompson J and Baker D 2013 High-resolution comparative modeling with RosettaCM *Structure* **21** 1735–42

[39] Ovchinnikov S, Park H, Varghese N, Huang P S, Pavlopoulos G A, Kim D E, Kamisetty H, Kyrpides N C and Baker D 2017 Protein structure determination using metagenome sequence data *Science (80-.).* **355** 294–8

[40] Park H, Ovchinnikov S, Kim D E, DiMaio F and Baker D 2018 Protein homology model refinement by large-scale energy optimization *Proc. Natl. Acad. Sci. U. S. A.* **115** 3054–9

[41] Gront D, Kulp D W, Vernon R M, Strauss C E M and Baker D 2011 Generalized fragment picking in rosetta: Design, protocols and applications *PLoS One* **6** e23294

[42] Wernisch L, Hery S and Wodak S J 2000 Automatic protein design with all atom force-fields by exact and heuristic

optimization11Edited by J. Thorton *J. Mol. Biol.* **301** 713–36

[43] Opuu V, Sun Y J, Hou T, Panel N, Fuentes E J and Simonson T 2020 A physics-based energy function allows the computational redesign of a PDZ domain *Sci. Rep.* **10** 1–9

[44] Damborsky J and Brezovsky J 2014 Computational tools for designing and engineering enzymes *Curr. Opin. Chem. Biol.* **19** 8–16

[45] Marze N A, Roy Burman S S, Sheffler W and Gray J J 2018 Efficient flexible backbone protein-protein docking for challenging targets *Bioinformatics* **34** 3461–9

[46] Roy Burman S S, Yovanno R A and Gray J J 2019 Flexible Backbone Assembly and Refinement of Symmetrical Homomeric Complexes *Structure* **27** 1041-1051.e8

[47] Meiler J and Baker D 2006 ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility *Proteins Struct. Funct. Genet.* **65** 538–48

[48] DeLuca S, Khar K and Meiler J 2015 Fully flexible docking of medium sized ligand libraries with rosettaligand *PLoS One* **10** 1–19

[49] Mills J H, Khare S D, Bolduc J M, Forouhar F, Mulligan V K, Lew S, Seetharaman J, Tong L, Stoddard B L and Baker D 2013 Computational design of an unnatural amino acid dependent metalloprotein with atomic level accuracy *J. Am. Chem. Soc.* **135** 13393–9

[50] Davis I W and Baker D 2009 RosettaLigand docking with full ligand and receptor flexibility *J. Mol. Biol.* **385** 381–92

[51] Gowthaman R, Miller S A, Rogers S, Khowsathit J, Lan L, Bai N, Johnson D K, Liu C, Xu L, Anbanandam A, Aubé J, Roy A and Karanicolas J 2016 DARC: Mapping Surface Topography by Ray-Casting for Effective Virtual Screening at Protein Interaction Sites *J. Med. Chem.* **59** 4152–70

[52] Johnson D K and Karanicolas J 2013 Druggable Protein Interaction Sites Are More Predisposed to Surface Pocket Formation than the Rest of the Protein Surface *PLoS Comput. Biol.* **9** e1002951

[53] Johnson D K and Karanicolas J 2015 Selectivity by small-molecule inhibitors of protein interactions can be driven by protein surface fluctuations *PLoS Comput Biol* **11** e1004081

[54] Fu D Y and Meiler J 2018 RosettaLigandEnsemble: A Small-Molecule Ensemble-Driven Docking Approach | ACS Omega *ACS omega* **3** 3655–64

[55] Moretti R, Bender B J, Allison B and Meiler J 2016 Rosetta and the Design of Ligand Binding Sites. *Methods Mol. Biol.* **1414** 47–62

[56] Stein A and Kortemme T 2013 Improvements to Robotics-Inspired Conformational Sampling in Rosetta *PLoS One* **8** e63090

[57] Canutescu A A and Dunbrack R L 2003 Cyclic coordinate descent: A robotics algorithm for protein loop closure *Protein Sci.* **12** 963–72

[58] Bhardwaj G, Mulligan V K, Bahl C D, Gilmore J M, Harvey P J, Cheneval O, Buchko G W, Pulavarti S V, Kaas Q and Eletsky A 2016 Accurate de novo design of hyperstable constrained peptides *Nature* **538** 329–35

[59] Marcos E, Chidyausiku T M, McShan A C, Evangelidis T, Nerli S, Carter L, Nivón L G, Davis A, Oberdorfer G, Tripsianes K, Sgourakis N G and Baker D 2018 De novo design of a non-local β-sheet protein with high stability and accuracy *Nat. Struct. Mol. Biol.* **25** 1028–34

[60] Nerli S and Sgourakis N G 2019 Cs-rosetta *Methods Enzymol.* **614** 321–62

[61] Rohl C A and Baker D 2002 De Novo Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta *J. Am. Chem. Soc.* **124** 2723–9

[62] Yagi H, Pilla K B, Maleckis A, Graham B, Huber T and Otting G 2013 Three-dimensional protein fold determination from backbone amide pseudocontact shifts generated by lanthanide tags at multiple

sites *Structure* **21** 883–90

[63] Schmitz C, Vernon R, Otting G, Baker D and Huber T 2012 Protein structure determination from pseudocontact shifts using ROSETTA *J. Mol. Biol.* **416** 668–77

[64] Pilla K B, Otting G and Huber T 2016 Pseudocontact Shift-Driven Iterative Resampling for 3D Structure Determinations of Large Proteins *J. Mol. Biol.* **428** 522–32

[65] Evangelidis T, Nerli S, Nováček J, Brereton A E, Karplus P A, Dotas R R, Venditti V, Sgourakis N G and Tripsianes K 2018 Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra *Nat. Commun.* **9** 1–13

[66] Lange O F 2014 Automatic NOESY assignment in CS-RASREC-Rosetta *J. Biomol. NMR* **59** 147–59

[67] Kuenze G, Bonneau R, Leman J K and Meiler J 2019 Integrative Protein Modeling in RosettaNMR from Sparse Paramagnetic Restraints *Structure* **27** 1721-1734.e5

[68] Raveh B, London N and Schueler-Furman O 2010 Sub-angstrom modeling of complexes between flexible peptides and globular proteins *Proteins Struct. Funct. Bioinforma.* **78** 2029–40

[69] Pacella M S, Koo D C E, Thottungal R A and Gray J J 2013 Using the Rosetta surface algorithm to predict protein structure at mineral surfaces *Methods in Enzymology* vol 532, ed J J B T-M in E De Yoreo (Academic Press) pp 343–66

[70] Raveh B, London N, Zimmerman L and Schueler-Furman O 2011 Rosetta FlexPepDockab-initio: Simultaneous folding, docking and refinement of peptides onto their receptors *PLoS One* **6** e18934

[71] Alam N, Goldstein O, Xia B, Porter K A, Kozakov D and Schueler-Furman O 2017 High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock *PLoS Comput. Biol.* **13** e1005905

[72] Sedan Y, Marcu O, Lyskov S and Schueler-Furman O 2016 Peptiderive server: derive peptide inhibitors from protein-protein interactions *Nucleic Acids Res.* **44** W536–41

[73] Hosseinzadeh P, Bhardwaj G, Mulligan V K, Shortridge M D, Craven T W, Pardo-Avila F, Rettie S A, Kim D E, Silva D A, Ibrahim Y M, Webb I K, Cort J R, Adkins J N, Varani G and Baker D 2017 Comprehensive computational design of ordered peptide macrocycles *Science (80-.).* **358** 1461–6

[74] Dang B, Wu H, Mulligan V K, Mravic M, Wu Y, Lemmin T, Ford A, Silva D A, Baker D and DeGrado W F 2017 De novo design of covalently constrained mesosize protein scaffolds with unique tertiary structures *Proc. Natl. Acad. Sci. U. S. A.* **114** 10852–7

[75] Rubenstein A B, Pethe M A and Khare S D 2017 MFPred: Rapid and accurate prediction of protein-peptide recognition multispecificity using self-consistent mean field theory *PLoS Comput. Biol.* **13** e1005614

[76] Lubin J H, Pacella M S and Gray J J 2018 A Parametric Rosetta Energy Function Analysis with LK Peptides on SAM Surfaces *Langmuir* **34** 5279–89

[77] Pacella M S and Gray J J 2018 A Benchmarking Study of Peptide–Biomineral Interactions *Cryst. Growth Des.* **18** 607–16

[78] Das R 2013 Atomic-Accuracy Prediction of Protein Loop Structures through an RNA-Inspired Ansatz *PLoS One* **8** e74830

[79] Sripakdeevong P, Kladwang W and Das R 2011 An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling *Proc. Natl. Acad. Sci. U. S. A.* **108** 20573–8

[80] Watkins A M, Geniesse C, Kladwang W, Zakrevsky P, Jaeger L and Das R 2018 Blind prediction of noncanonical RNA structure at atomic accuracy *Sci. Adv.* **4** eaar5316

[81] Kappel K and Das R 2019 Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking *Structure* **27** 140-151.e5

[82] Das R, Karanicolas J and Baker D 2010 Atomic accuracy in predicting and

designing noncanonical RNA structure *Nat. Methods* **7** 291–4

[83] Cheng C Y, Chou F C and Das R 2015 Modeling complex RNA tertiary folds with Rosetta *Methods in Enzymology* vol 553, ed S-J Chen and D H B T-M in E Burke-Aguero (Academic Press) pp 35–64

[84] Chou F C, Sripakdeevong P, Dibrov S M, Hermann T and Das R 2013 Correcting pervasive errors in RNA crystallography through enumerative structure prediction *Nat. Methods* **10** 74–6

[85] Chou F C, Kladwang W, Kappel K and Das R 2016 Blind tests of RNA nearest-neighbor energy prediction *Proc. Natl. Acad. Sci. U. S. A.* **113** 8430–5

[86] Kappel K, Liu S, Larsen K P, Skiniotis G, Puglisi E V, Puglisi J D, Zhou Z H, Zhao R and Das R 2018 De novo computational RNA modeling into cryo-EM maps of large ribonucleoprotein complexes *Nat. Methods* **15** 947–54

[87] Sircar A, Kim E T and Gray J J 2009 RosettaAntibody: antibody variable region homology modeling server *Nucleic Acids Res.* **37** W474–9

[88] Weitzner B D, Jeliazkov J R, Lyskov S, Marze N, Kuroda D, Frick R, Adolf-Bryfogle J, Biswas N, Dunbrack R L and Gray J J 2017 Modeling and docking of antibody structures with Rosetta *Nat. Protoc.* **12** 401–16

[89] Sivasubramanian A, Sircar A, Chaudhury S and Gray J J 2009 Toward high-resolution homology modeling of antibody F v regions and application to antibody-antigen docking *Proteins Struct. Funct. Bioinforma.* **74** 497–514

[90] Norn C H, Lapidoth G and Fleishman S J 2017 High-accuracy modeling of antibody structures by a search for minimum-energy recombination of backbone fragments *Proteins Struct. Funct. Bioinforma.* **85** 30–8

[91] Lapidoth G, Parker J, Prilusky J and Fleishman S J 2019 AbPredict 2: A server for accurate and unstrained structure prediction of antibody variable domains *Bioinformatics* **35** 1591–3

[92] Toor J S, Rao A A, McShan A C, Yarmarkovich M, Nerli S, Yamaguchi K, Madejska A A, Nguyen S, Tripathi S, Maris J M, Salama S R, Haussler D and Sgourakis N G 2018 A recurrent mutation in anaplastic lymphoma kinase with distinct neoepitope conformations *Front. Immunol.* **9** 99

[93] Gowthaman R and Pierce B G 2018 TCRmodel: High resolution modeling of T cell receptors from sequence *Nucleic Acids Res.* **46** W396–401

[94] Sircar A and Gray J J 2010 SnugDock: Paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models *PLoS Comput. Biol.* **6** e1000644

[95] Adolf-Bryfogle J, Kalyuzhniy O, Kubitz M, Weitzner B D, Hu X, Adachi Y, Schief W R and Dunbrack R L 2018 RosettaAntibodyDesign (RAbD): A general framework for computational antibody design *PLoS Comput. Biol.* **14** e1006112

[96] King C, Garza E N, Mazor R, Linehan J L, Pastan I, Pepper M and Baker D 2014 Removing T-cell epitopes with computational protein design *Proc. Natl. Acad. Sci.* **111** 8577 LP – 8582

[97] Nivón L G, Bjelic S, King C and Baker D 2014 Automating human intuition for protein design *Proteins Struct. Funct. Bioinforma.* **82** 858–66

[98] Lapidoth G D, Baran D, Pszolla G M, Norn C, Alon A, Tyka M D and Fleishman S J 2015 AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins* **83** 1385–406

[99] Leman J K, Mueller B K and Gray J J 2017 Expanding the toolkit for membrane protein modeling in Rosetta *Bioinformatics* **33** 754–6

[100] Koehler Leman J, Lyskov S and Bonneau R 2017 Computing structure-based lipid accessibility of membrane proteins with mp_lipid_acc in RosettaMP *BMC Bioinformatics* **18** 115

[101] Koehler Leman J and Bonneau R 2018 A Novel Domain Assembly Routine for

Creating Full-Length Models of Membrane Proteins from Known Domain Structures *Biochemistry* **57** 1939–44

[102] Bender B J, Cisneros A, Duran A M, Finn J A, Fu D, Lokits A D, Mueller B K, Sangha A K, Sauer M F, Sevy A M, Sliwoski G, Sheehan J H, DiMaio F, Meiler J and Moretti R 2016 Protocols for Molecular Modeling with Rosetta3 and RosettaScripts *Biochemistry* **55** 4748–63

[103] Labonte J W, Adolf-Bryfogle J, Schief W R and Gray J J 2017 Residue-centric modeling and design of saccharide and glycoconjugate structures *J. Comput. Chem.* **38** 276–87

[104] Frenz B, Rämisch S, Borst A J, Walls A C, Adolf-Bryfogle J, Schief W R, Veesler D and DiMaio F 2019 Automatically Fixing Errors in Glycoprotein Structures with Rosetta *Structure* **27** 134-139.e3

[105] Gordon D B, Marshall S A and Mayo S L 1999 Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9** 509–13

[106] Hoffman J G, Metropolis N and Gardiner V 1956 Digital computer studies of cell multiplication by monte carlo methods *J. Natl. Cancer Inst.* **17** 175–88

[107] Kuhlman B and Baker D 2000 Native protein sequences are close to optimal for their structures *Proc. Natl. Acad. Sci. U. S. A.* **97** 10383–8

[108] Alford R F, Leaver-Fay A, Jeliazkov J R, O'Meara M J, DiMaio F P, Park H, Shapovalov M V., Renfrew P D, Mulligan V K, Kappel K, Labonte J W, Pacella M S, Bonneau R, Bradley P, Dunbrack R L, Das R, Baker D, Kuhlman B, Kortemme T, Gray J J, O'Meara M J, DiMaio F P, Park H, Shapovalov M V., Renfrew P D, Mulligan V K, Kappel K, Labonte J W, Pacella M S, Bonneau R, Bradley P, Dunbrack R L, Das R, Baker D, Kuhlman B, Kortemme T and Gray J J 2017 The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design *J. Chem. Theory Comput.* **13** 3031–48

[109] Kuhlman B, Dantas G, Ireton G C, Varani G, Stoddard B L and Baker D 2003 Design of a novel globular protein fold with atomic-level accuracy. *Science* **302** 1364–8

[110] Ponder J W and Richards F M 1987 Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes *J. Mol. Biol.* **193** 775–91

[111] Dunbrack R L and Karplus M 1993 Backbone-dependent rotamer library for proteins: Application to side-chain prediction *J. Mol. Biol.* **230** 543–74

[112] Tuffery P, Etchebest C, Hazout S and Lavery R 1991 A new approach to the rapid determination of protein side chain conformations *J. Biomol. Struct. Dyn.* **8** 1267–89

[113] Dunbrack R L 2002 Rotamer libraries in the 21st century *Curr. Opin. Struct. Biol.* **12** 431–40

[114] Baldwin E P, Hajiseyedjavadi O, Baase W A and Matthews B W 1993 The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme *Science (80-.).* **262** 1715–8

[115] Keedy D A, Georgiev I, Triplett E B, Donald B R, Richardson D C and Richardson J S 2012 The Role of Local Backrub Motions in Evolved and Designed Mutations *PLOS Comput. Biol.* **8** e1002629

[116] Jacobs T M, Williams B, Williams T, Xu X, Eletsky A, Federizon J F, Szyperski T and Kuhlman B 2016 Design of structurally distinct proteins using strategies inspired by evolution *Science (80-.).* **352** 687–90

[117] Guffy S L, Teets F D, Langlois M I and Kuhlman B 2018 Protocols for Requirement-Driven Protein Design in the Rosetta Modeling Program. *J. Chem. Inf. Model.* **58** 895–901

[118] Huang P-S, Ban Y-E A, Richter F, Andre I, Vernon R, Schief W R and Baker D 2011 RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design *PLoS One* **6** e24109

[119] Huang P S, Feldmeier K, Parmeggiani F, Velasco D F, Hocker B and Baker D 2016 De novo design of a four-fold symmetric TIM-barrel protein with atomic-level

accuracy *Nat. Chem. Biol.* **12** 29–34

[120] Parmeggiani F, Huang P-S, Vorobiev S, Xiao R, Park K, Caprari S, Su M, Seetharaman J, Mao L, Janjua H, Montelione G T, Hunt J and Baker D 2015 A general computational approach for repeat protein design. *J. Mol. Biol.* **427** 563–75

[121] Park K, Shen B W, Parmeggiani F, Huang P-S, Stoddard B L and Baker D 2015 Control of repeat-protein curvature by computational protein design *Nat. Struct. & Mol. Biol.* **22** 167—174

[122] Brunette T J, Parmeggiani F, Huang P-S, Bhabha G, Ekiert D C, Tsutakawa S E, Hura G L, Tainer J A and Baker D 2015 Exploring the repeat protein universe through computational protein design *Nature* **528** 580–4

[123] Doyle L, Hallinan J, Bolduc J, Parmeggiani F, Baker D, Stoddard B L and Bradley P 2015 Rational design of α-helical tandem repeat proteins with closed architectures *Nature* **528** 585–8

[124] Saunders C T and Baker D 2005 Recapitulation of protein family divergence using flexible backbone protein design. *J. Mol. Biol.* **346** 631–44

[125] Khatib F, Cooper S, Tyka M D, Xu K, Makedon I, Popović Z, Baker D and Players F 2011 Algorithm discovery by protein folding game players *Proc. Natl. Acad. Sci. U. S. A.* **108** 18949–53

[126] Tyka M D, Keedy D A, André I, Dimaio F, Song Y, Richardson D C, Richardson J S and Baker D 2011 Alternate states of proteins revealed by detailed energy landscape mapping *J. Mol. Biol.* **405** 607–18

[127] Nivón L G, Moretti R and Baker D 2013 A Pareto-Optimal Refinement Method for Protein Design Scaffolds *PLoS One* **8** e59004

[128] Conway P, Tyka M D, DiMaio F, Konerding D E and Baker D 2014 Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23** 47–55

[129] Dou J, Vorobieva A A, Sheffler W, Doyle L A, Park H, Bick M J, Lee M Y, Gagnon L A, Carter L, Sankaran B, Mao B and Glenna W 2018 De novo design of a fluorescence-activating β -barrel *Nature*

[130] Silva D-A, Yu S, Ulge U Y, Spangler J B, Jude K M, Labão-Almeida C, Ali L R, Quijano-Rubio A, Ruterbusch M, Leung I, Biary T, Crowley S J, Marcos E, Walkey C D, Weitzner B D, Pardo-Avila F, Castellanos J, Carter L, Stewart L, Riddell S R, Pepper M, Bernardes G J L, Dougan M, Garcia K C and Baker D 2019 De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565** 186–91

[131] Ollikainen N and Kortemme T 2013 Computational Protein Design Quantifies Structural Constraints on Amino Acid Covariation *PLoS Comput. Biol.* **9** e1003313

[132] Smith C A and Kortemme T 2010 Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *J. Mol. Biol.* **402** 460–74

[133] Smith C A and Kortemme T 2011 Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design *PLoS One* **6** e20451

[134] Friedland G D, Lakomek N-A, Griesinger C, Meiler J and Kortemme T 2009 A Correspondence Between Solution-State Dynamics of an Individual Protein and the Sequence and Conformational Diversity of its Family *PLOS Comput. Biol.* **5** e1000393

[135] Humphris E L and Kortemme T 2008 Prediction of Protein-Protein Interface Sequence Diversity Using Flexible Backbone Computational Protein Design *Structure* **16** 1777–88

[136] Smith C A and Kortemme T 2008 Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction *J. Mol. Biol.* **380** 742–56

[137] Ollikainen N, Smith C A, Fraser J S and Kortemme T 2013 Flexible backbone

sampling methods to model and design protein alternative conformations *Methods in Enzymology* vol 523 pp 61–85

[138] Friedland G D, Linares A J, Smith C A and Kortemme T 2008 A simple model of backbone flexibility improves modeling of side-chain conformational variability *J. Mol. Biol.* **380** 757—774

[139] Kapp G T, Liu S, Stein A, Wong D T, Remeńyi A, Yeh B J, Fraser J S, Taunton J, Lim W A and Kortemme T 2012 Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair *Proc. Natl. Acad. Sci. U. S. A.* **109** 5277–82

[140] Ollikainen N, de Jong R M and Kortemme T 2015 Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-design of Protein-Ligand Specificity *PLOS Comput. Biol.* **11** e1004335

[141] Mandell D J, Coutsias E A and Kortemme T 2009 Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling *Nat. Methods* **6** 551–2

[142] Leaver-Fay A, Jacak R, Stranges P B and Kuhlman B 2011 A Generic Program for Multistate Protein Design *PLoS One* **6** e20937

[143] Sevy A M, Wu N C, Gilchuk I M, Parrish E H, Burger S, Yousif D, Nagel M B M, Schey K L, Wilson I A, Crowe J E and Meiler J 2019 Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses *Proc. Natl. Acad. Sci. U. S. A.* **116** 1597–602

[144] Sormani G, Harteveld Z, Rosset S, Correia B and Laio A 2021 A Rosetta-based protein design protocol converging to natural sequences *J. Chem. Phys.* **154** 74114

[145] Derrida B 1981 Random-energy model: An exactly solvable model of disordered systems *Phys. Rev. B* **24** 2613–26

[146] Pande V S V S, Grosberg A Y Y A Y A Y and Tanaka T 1997 Statistical mechanics of simple models of protein folding and design. *Biophys. J.* **73** 3192–210

[147] Shakhnovich E I and Gutin A M 1989 Formation of unique structure in polypeptide chains: Theoretical investigation with the aid of a replica approach *Biophys. Chem.* **34** 187–99

[148] Shakhnovich E I 1998 Protein design: A perspective from simple tractable models *Fold. Des.* **3** 45–58

[149] Tian P and Best R B 2017 How Many Protein Sequences Fold to a Given Structure? A Coevolutionary Analysis *Biophys. J.* **113** 1719–30

[150] Shoemaker B A, Portman J J and WOLYNES P G 2000 Speeding molecular recognition by using the folding funnel: The fly-casting mechanism *Proc. Natl. Acad. Sci. U. S. A.* **97** 8868-+

[151] Shehu A, Kavraki L E and Clementi C 2009 Multiscale characterization of protein conformational ensembles *Proteins Struct. Funct. Bioinforma.* **76** 837–51

[152] Larriva M M, Prieto L, Bruscolini P and Rey A 2010 A simple simulation model can reproduce the thermodynamic folding intermediate of apoflavodoxin *Proteins Struct. Funct. Bioinforma.* **78** 73–82

[153] Hills R D, Lu L and Voth G A 2010 Multiscale coarse-graining of the protein energy landscape *PLoS Comput. Biol.* **6** 1–15

[154] Kinch L N, Shi S, Cheng H, Cong Q, Pei J, Mariani V, Schwede T and Grishin N V. 2011 CASP9 target classification *Proteins* **79** 21–36

[155] Bowman G R, Voelz V A and Pande V S 2011 Taming the complexity of protein folding. *Curr. Opin. Struct. Biol.* **21** 4–11

[156] Wolynes P G, Eaton W A and Fersht A R 2012 Chemical physics of protein folding *Proc. Natl. Acad. Sci. U. S. A.* **109** 17770–1

[157] Distasio R A, von Lilienfeld O A and Tkatchenko A 2012 Collective many-body van der Waals interactions in molecular systems. *Proc. Natl. Acad. Sci. U. S. A.* **109** 14791–5

[158] Kellogg E H, Lange O F and Baker D 2012 Evaluation and optimization of discrete state models of protein folding *J. Phys. Chem. B* **116** 11405–13

[159] Krobath H, Estácio S G, Faísca P F N N

and Shakhnovich E I 2012 Identification of a conserved aggregation-prone intermediate state in the folding pathways of Spc-SH3 amyloidogenic variants *J. Mol. Biol.* **422** 705–22

[160] Lin M M and Zewail A H 2012 Protein folding - Simplicity in complexity *Ann. Phys.* **524** 379–91

[161] Go N 1983 Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12** 183–210

[162] Estácio S G, Fernandes C S, Krobath H, Faísca P F N N, Shakhnovich E I, Fa\'\isca P F N, Shakhnovich E I, Faísca P F N N, Shakhnovich E I, Fa$\$’$\$isca P F N, Shakhnovich E I, Faísca P F N N and Shakhnovich E I 2012 Robustness of atomistic Gō models in predicting native-like folding intermediates *J. Chem. Phys.* **137** 085102

[163] Mochalin V N, Shenderova O, Ho D and Gogotsi Y 2012 The properties and applications of nanodiamonds *Nat. Nanotechnol.* **7** 11–23

[164] Noid W G 2013 Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **139** 90901

[165] Clementi C, Nymeyer H and Onuchic J N 2000 Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298** 937–53

[166] Whitford P C, Noel J K, Gosavi S, Schug A, Sanbonmatsu K Y and Onuchic J N J N 2009 An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields *Proteins Struct. Funct. Bioinforma.* **75** 430–41

[167] Garnier, J. and Robson B, Richardson J S, Richardson D C, Garnier, J. and Robson B, Richardson J S and Richardson D C 1989 *Prediction of Protein Structure and the Principles of Protein Conformation* ed G D Fasman (New York: Springer US)

[168] Tirion M M M M 1996 Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **77** 1905–8

[169] Atilgan A R, Durell S R, Jernigan R L, Demirel M C, Keskin O and Bahar I 2001 Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80** 505–15

[170] Ollerenshaw J E, Kaya H, Chan H S and Kay L E 2004 Sparsely populated folding intermediates of the Fyn SH3 domain: matching native-centric essential dynamics and experiment. *Proc. Natl. Acad. Sci. U. S. A.* **101** 14748–53

[171] Tozzini V 2005 Coarse-grained models for proteins *Curr. Opin. Struct. Biol.* **15** 144–50

[172] Clementi C 2008 Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.* **18** 10–5

[173] Sułkowska J I, Cieplak M, Su$\$lkowska J I, Cieplak M, Sułkowska J I and Cieplak M 2008 Selection of optimal variants of Gō-like models of proteins through studies of stretching. *Biophys. J.* **95** 3174–91

[174] Finkelstein A V., Gutun A M, Badretdinov A Y, Finkelstein V A, Gutun A M and Badretdinov A Y 1993 Why are the same protein folds used to perform different functions? *FEBS Lett.* **325** 23–8

[175] Pande V S, Grosberg A Y and Tanaka T 2000 Heteropolymer freezing and design: Towards physical models of protein folding *Rev. Mod. Phys.* **72** 259–314

[176] Cardelli C, Nerattini F, Tubiana L, Bianco V, Dellago C, Sciortino F and Coluzza I 2019 General Methodology to Identify the Minimum Alphabet Size for Heteropolymer Design *Adv. Theory Simulations* **2** 1900031

[177] Vissers T, Smallenburg F, Munaò G, Preisler Z and Sciortino F 2014 Cooperative polymerization of one-patch colloids *J. Chem. Phys.* **140** 144902

[178] Ronti M, Rovigatti L, Tavares J M, Ivanov A O, Kantorovich S S and Sciortino F 2017 Free energy calculations for rings and chains formed by dipolar hard spheres *Soft Matter*

[179] Cardelli C, Bianco V, Rovigatti L, Nerattini F, Tubiana L, Dellago C and Coluzza I 2017

The role of directional interactions in the designability of generalized heteropolymers *Sci. Rep.* **7** 4986

[180] Nerattini F, Tubiana L, Cardelli C, Bianco V, Dellago C and Coluzza I 2020 Protein design under competing conditions for the availability of amino acids *Sci. Rep.* **10** 2684

[181] Shakhnovich E I and Gutin A M 1993 A new approach to the design of stable proteins *Protein Eng. Des. Sel.* **6** 793–800

[182] Shakhnovich E I and Gutin A M 1990 Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* **346** 773–5

[183] Coluzza I, Muller H G G and Frenkel D 2003 Designing refoldable model molecules. *Phys. Rev. E* **68** 46703

[184] Coluzza I and Frenkel D 2007 Monte Carlo study of substrate-induced folding and refolding of lattice proteins. *Biophys. J.* **92** 1150–6

[185] Abeln S and Frenkel D 2008 Disordered flanks prevent peptide aggregation ed B Rost *PLoS Comput. Biol.* **4** e1000241

[186] Abeln S and Frenkel D 2011 Accounting for protein-solvent contacts facilitates design of nonaggregating lattice proteins *Biophys. J.* **100** 693–700

[187] Faísca P F N 2015 Knotted proteins: A tangled tale of Structural Biology *Comput. Struct. Biotechnol. J.* **13** 459–68

[188] Kolinski A and Skolnick J 1994 Monte carlo simulations of protein folding. I. Lattice model and interaction scheme *Proteins Struct. Funct. Bioinforma.* **18** 338–52

[189] Allouche A 2012 Software News and Updates Gabedit — A Graphical User Interface for Computational Chemistry Softwares *J. Comput. Chem.* **32** 174–82

[190] Kolinski A and Skolnick J 2004 Reduced models of proteins and their applications *Polymer (Guildf).* **45** 511–24

[191] Kolinski A and Skolnick J 1992 Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides *J. Chem. Phys.* **97** 9412–26

[192] Godzik A, Kolinski A and Skolnick J 1993 Lattice representations of globular proteins: How good are they? *J. Comput. Chem.* **14** 1194–202

[193] Skolnick J, Kolinski A, Brooks C L, Godzik A and Rey A 1993 A method for predicting protein structure from sequence *Curr. Biol.* **3** 414–23

[194] Zeldovich K B and Shakhnovich E I 2008 Understanding Protein Evolution: From Protein Physics to Darwinian Selection *Annu. Rev. Phys. Chem.* **59** 105–27

[195] Hubner I A, Oliveberg M and Shakhnovich E I 2004 Simulation, experiment, and evolution: Understanding nucleation in protein S6 folding *Proc. Natl. Acad. Sci. U. S. A.* **101** 8354–9

[196] Deeds E J, Dokholyan N V and Shakhnovich E I 2003 Protein evolution within a structural space *Biophys. J.* **85** 2962–72

[197] Dokholyan N V and Shakhnovich E I 2001 Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* **312** 289–307

[198] Ni R, Abeln S, Schor M, Cohen Stuart M A and Bolhuis P G 2013 Interplay between folding and assembly of fibril-forming polypeptides *Phys. Rev. Lett.* **111** 1–5

[199] Abeln S, Vendruscolo M, Dobson C M, Frenkel D and Riekel C 2014 A simple lattice model that captures protein folding, aggregation and amyloid formation ed I V Baskakov *PLoS One* **9** e85185

[200] Bianco V, Alonso-Navarro M, Di Silvio D, Moya S, Cortajarena A L and Coluzza I 2019 Proteins are Solitary! Pathways of Protein Folding and Aggregation in Protein Mixtures *J. Phys. Chem. Lett.* **10** 4800–4

[201] Bianco V, Franzese G and Coluzza I 2020 In Silico Evidence That Protein Unfolding is a Precursor of Protein Aggregation *ChemPhysChem* **21** 377–84

[202] Zhang J, Maslov S and Shakhnovich E I 2008 Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size *Mol. Syst. Biol.* **4** 1–11

[203] Sułkowska J I, Sułkowski P and Onuchic J 2009 Dodging the crisis of folding proteins with knots *Proc. Natl. Acad. Sci. U. S. A.* **106** 3119–24

[204] Soler M A, Nunes A, Faísca P F N and Fa\'\isca P F N 2014 Effects of knot type in the folding of topologically complex lattice proteins *J. Chem. Phys.* **141** 07B607_1

[205] Deeds E J, Ashenberg O, Gerardin J and Shakhnovich E I 2007 Robust protein-protein interactions in crowded cellular environments *Proc. Natl. Acad. Sci. U. S. A.* **104** 14952–7

[206] Nerattini F, Tubiana L, Cardelli C, Bianco V, Dellago C and Coluzza I 2019 Design of Protein–Protein Binding Sites Suggests a Rationale for Naturally Occurring Contact Areas *J. Chem. Theory Comput.* **15** 1383–92

[207] Tartaglia G G, Pechmann S, Dobson C M and Vendruscolo M 2007 Life on the edge: a link between gene expression levels and aggregation rates of human proteins *Trends Biochem. Sci.* **32** 204–6

[208] Sharma R, De Sancho D and Muñ V 2017 Interplay between the folding mechanism and binding modes in folding coupled to binding processes *Phys. Chem. Chem. Phys* **19** 28512–6

[209] Maritan A, Micheletti C, Trovato A and Banavar J R 2000 Optimal shapes of compact strings. *Nature* **406** 287–90

[210] Hoang T X, Trovato A, Seno F, Banavar J R and Maritan A 2004 Geometry and symmetry presculpt the free-energy landscape of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **101** 7960–4

[211] Magee J E, Vasquez V R and Lue L 2006 Helical structures from an isotropic homopolymer model *Phys. Rev. Lett.* **96** 2078028

[212] Banavar J and Maritan A 2007 The maximum relative entropy principle *arXiv* **cond**-mat.s

[213] Hoang T X, Seno F, Trovato A, Banavar J R and Maritan A 2008 Inference of the solvation energy parameters of amino acids using maximum entropy approach *J. Chem. Phys.* **129** 35102

[214] Seno F, Trovato A, Banavar J R and Maritan A 2008 Maximum entropy approach for deducing amino acid interactions in proteins *Phys. Rev. Lett.* **100** 1–4

[215] Kukic P, Kannan A, Dijkstra M J J, Abeln S, Camilloni C and Vendruscolo M 2015 Mapping the Protein Fold Universe Using the CamTube Force Field in Molecular Dynamics Simulations ed A Shehu *PLOS Comput. Biol.* **11** e1004435

[216] Škrbić T, Badasyan A, Hoang T X, Podgornik R, Giacometti A, Skrbic T, Badasyan A, Hoang T X, Podgornik R, Giacometti A, Škrbić T, Badasyan A, Hoang T X, Podgornik R, Giacometti A, Skrbic T, Badasyan A, Hoang T X, Podgornik R, Giacometti A, Škrbić T, Badasyan A, Hoang T X, Podgornik R and Giacometti A 2016 From polymers to proteins: the effect of side chains and broken symmetry on the formation of secondary structures within a Wang–Landau approach *Soft Matter* **12** 4783–93

[217] Škrbić T, Hoang T X and Giacometti A 2016 Effective stiffness and formation of secondary structures in a protein-like model *J. Chem. Phys.* **145** 84904

[218] Marrink S J and Tieleman D P 2013 Perspective on the martini model *Chem. Soc. Rev.* **42** 6801–22

[219] Alessandri R, Grünewald F and Marrink S J 2021 The Martini Model in Materials Science *Adv. Mater.* **33**

[220] Marrink S J, De Vries A H and Mark A E 2004 Coarse Grained Model for Semiquantitative Lipid Simulations *J. Phys. Chem. B* **108** 750–60

[221] Marrink S J, Risselada H J, Yefimov S, D. Peter Tieleman and and de Vries A H 2007 The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations

[222] Bruininks B M H, Souza P C T and Marrink S J 2019 A Practical View of the Martini Force Field *Methods in Molecular Biology* vol 2022 (Springer) pp 105–27

[223] Monticelli L, Kandasamy S K, Periole X, Larson R G, Tieleman D P and Marrink S-J

2008 The MARTINI Coarse-Grained Force Field: Extension to Proteins *J. Chem. Theory Comput.* **4** 819–34

[224] Herzog F A, Braun L, Schoen I and Vogel V 2016 Improved Side Chain Dynamics in MARTINI Simulations of Protein-Lipid Interfaces *J. Chem. Theory Comput.* **12** 2446–58

[225] Ingólfsson H I, Arnarez C, Periole X and Marrink S J 2016 Computational 'microscopy'of cellular membranes *J. Cell Sci.* **129** 257–68

[226] Arnarez C, Marrink S J and Periole X 2016 Molecular mechanism of cardiolipin-mediated assembly of respiratory chain supercomplexes *Chem. Sci.* **7** 4435–43

[227] Melo M N, Ingólfsson H I and Marrink S J 2015 Parameters for Martini sterols and hopanoids based on a virtual-site description *J. Chem. Phys.* **143** 243152

[228] López C A, Rzepiela A J, de Vries A H, Dijkhuizen L, Hünenberger P H and Marrink S J 2009 Martini Coarse-Grained Force Field: Extension to Carbohydrates *J. Chem. Theory Comput.* **5** 3195–210

[229] López C A, Sovova Z, Van Eerden F J, De Vries A H and Marrink S J 2013 Martini force field parameters for glycolipids *J. Chem. Theory Comput.* **9** 1694–708

[230] de Jong D H, Liguori N, van den Berg T, Arnarez C, Periole X and Marrink S J 2015 Atomistic and coarse grain topologies for the cofactors associated with the photosystem II core complex *J. Phys. Chem. B* **119** 7791–803

[231] Hinner M J, Marrink S-J and de Vries A H 2009 Location, tilt, and binding: a molecular dynamics study of voltage-sensitive dyes in biomembranes *J. Phys. Chem. B* **113** 15807–19

[232] Ingólfsson H I, Thakur P, Herold K F, Hobart E A, Ramsey N B, Periole X, De Jong D H, Zwama M, Yilmaz D and Hall K 2014 Phytochemicals perturb membranes and promiscuously alter protein function *ACS Chem. Biol.* **9** 1788–98

[233] Salassi S, Simonelli F, Bartocci A and Rossi G 2018 A Martini coarse-grained model of the calcein fluorescent dye *J. Phys. D. Appl. Phys.* **51** 384002

[234] Uusitalo J J, Ingólfsson H I, Akhshi P, Tieleman D P and Marrink S J 2015 Martini Coarse-Grained Force Field: Extension to DNA *J. Chem. Theory Comput.* **11** 3932–45

[235] Uusitalo J J, Ingólfsson H I, Marrink S J and Faustino I 2017 Martini Coarse-Grained Force Field: Extension to RNA *Biophys. J.* **113** 246–56

[236] Frederix P W J M, Scott G G, Abul-Haija Y M, Kalafatovic D, Pappas C G, Javid N, Hunt N T, Ulijn R V and Tuttle T 2015 Exploring the sequence space for (tri-)peptide self-assembly to design and discover new hydrogels *Nat. Chem.* **7** 30–7

[237] Sather N A, Sai H, Sasselli I R, Sato K, Ji W, Synatschke C V, Zambrotta R T, Edelbrock J F, Kohlmeyer R R, Hardin J O, Berrigan J D, Durstock M F, Mirau P and Stupp S I 2021 3D Printing of Supramolecular Polymer Hydrogels with Hierarchical Structure *Small* **17** 2005743

[238] Sasselli I R, Moreira I P, Ulijn R V and Tuttle T 2017 Molecular dynamics simulations reveal disruptive self-assembly in dynamic peptide libraries *Org. Biomol. Chem.* **15** 6541–7

[239] Zhao M, Sampath J, Alamdari S, Shen G, Chen C L, Mundy C J, Pfaendtner J and Ferguson A L 2020 MARTINI-Compatible Coarse-Grained Model for the Mesoscale Simulation of Peptoids *J. Phys. Chem. B* **124** 7745–64

[240] Panizon E, Bochicchio D, Monticelli L and Rossi G 2015 MARTINI coarse-grained models of polyethylene and polypropylene *J. Phys. Chem. B* **119** 8209–16

[241] Rossi G, Giannakopoulos I, Monticelli L, Rostedt N K J, Puisto S R, Lowe C, Taylor A C, Vattulainen I and Ala-Nissila T 2011 A MARTINI coarse-grained model of a thermoset polyester coating *Macromolecules* **44** 6198–208

[242] Alessandri R, Uusitalo J J, De Vries A H, Havenith R W A and Marrink S J 2017 Bulk Heterojunction Morphologies with Atomistic Resolution from Coarse-Grain

Solvent Evaporation Simulations *J. Am. Chem. Soc.* **139** 3697–705

[243] Crespo E A, Schaeffer N, Coutinho J A P and Perez-Sanchez G 2020 Improved coarse-grain model to unravel the phase behavior of 1-alkyl-3-methylimidazolium-based ionic liquids through molecular dynamics simulations *J. Colloid Interface Sci.* **574** 324–36

[244] Vazquez-Salazar L I, Selle M, De Vries A H, Marrink S J and Souza P C T 2020 Martini coarse-grained models of imidazolium-based ionic liquids: From nanostructural organization to liquid-liquid extraction *Green Chem.* **22** 7376–86

[245] de Jong D H, Singh G, Bennett W F D, Arnarez C, Wassenaar T A, Schäfer L V, Periole X, Tieleman D P and Marrink S J 2013 Improved Parameters for the Martini Coarse-Grained Protein Force Field *J. Chem. Theory Comput.* **9** 687–97

[246] Khan H M, Souza P C T, Thallmair S, Barnoud J, De Vries A H, Marrink S J and Reuter N 2020 Capturing Choline-Aromatics Cation-πInteractions in the MARTINI Force Field *J. Chem. Theory Comput.* **16** 2550–60

[247] Yesylevskyy S O, Schäfer L V, Sengupta D and Marrink S J 2010 Polarizable Water Model for the Coarse-Grained MARTINI Force Field ed M Levitt *PLoS Comput. Biol.* **6** e1000810

[248] Liu Y, De Vries A H, Barnoud J, Pezeshkian W, Melcr J and Marrink S J 2020 Dual Resolution Membrane Simulations Using Virtual Sites *J. Phys. Chem. B* **124** 3944–53

[249] Pezeshkian W, König M, Wassenaar T A and Marrink S J 2020 Backmapping triangulated surfaces to coarse-grained membrane models *Nat. Commun.* **11** 2296

[250] Periole X, Cavalli M, Marrink S J and Ceruso M A 2009 Combining an elastic network with a coarse-grained molecular force field: Structure, dynamics, and intermolecular recognition *J. Chem. Theory Comput.* **5** 2531–43

[251] Poma A B, Cieplak M and Theodorakis P E 2017 Combining the MARTINI and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins *J. Chem. Theory Comput.* **13** 1366–74

[252] Souza P C T, Thallmair S, Conflitti P, Ramírez-Palacios C, Alessandri R, Raniolo S, Limongelli V and Marrink S J 2020 Protein–ligand binding with the coarse-grained Martini model *Nat. Commun.* **11** 3714

[253] Grünewald F, Souza P C T, Abdizadeh H, Barnoud J, de Vries A H and Marrink S J 2020 Titratable Martini model for constant pH simulations *J. Chem. Phys.* **153** 24118

[254] Murphy L R, Wallqvist A and Levy R M 2000 Simplified amino acid alphabets for protein fold recognition and implications for folding *Protein Eng. Des. Sel.* **13** 149–52

[255] Salvi G, Mölbert S and De Los Rios P 2002 Design of lattice proteins with explicit solvent *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **66** 061911

[256] Wang T R, Miller J, Wingreen N S, Tang C and Dill K A 2000 Symmetry and designability for lattice protein models *J. Chem. Phys.* **113** 8329–36

[257] Deutsch J M and Kurosky T 1995 A New Algorithm for Protein Design *Phys. Rev. Lett.* **76** 10

[258] Shakhnovich E I and Gutin A M 1993 Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. U. S. A.* **90** 7195–9

[259] Yue K and Dill K A 1992 Inverse protein folding problem: designing polymer sequences. *Proc. Natl. Acad. Sci. U. S. A.* **89** 4163–7

[260] Chan H S and Dill K A 1996 Comparing folding codes for proteins and polymers *Proteins Struct. Funct. Genet.* **24** 335–44

[261] Sear R P and Cuesta J a 2003 Instabilities in Complex Mixtures with a Large Number of Components *Phys. Rev. Lett.* **91** 245701

[262] Sear R P 2004 Specific protein–protein binding in many-component mixtures of proteins *Phys. Biol.* **1** 53–60

[263] Sear R P 2004 Highly specific protein-

protein interactions, evolution and negative design *Phys. Biol.* **1** 166–72

[264] Madge J and Miller M A 2015 Design strategies for self-assembly of discrete targets *J. Chem. Phys.* **143** 44905

[265] Plaxco K W, Riddle D S, Grantcharova V and Baker D 1998 Simplified proteins: Minimalist solutions to the “protein folding problem” *Curr. Opin. Struct. Biol.* **8** 80–5

[266] Walter K U, Vamvaca K and Hilvert D 2005 An active enzyme constructed from a 9-amino acid alphabet *J. Biol. Chem.* **280** 37742–6

[267] Reetz M T and Wu S 2008 Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions *Chem. Commun.* 5499

[268] Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X and Chou K C 2014 IDNA-Prot|dis: Identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition *PLoS One* **9**

[269] Sun Z, Lonsdale R, Kong X D, Xu J H, Zhou J and Reetz M T 2015 Reshaping an Enzyme Binding Pocket for Enhanced and Inverted Stereoselectivity: Use of Smallest Amino Acid Alphabets in Directed Evolution *Angew. Chemie - Int. Ed.* **54** 12410–5

[270] Wang J and Wang W 2016 Simplification of complexity in protein molecular systems by grouping amino acids: a view from physics *Adv. Phys. X* **1** 444–66

[271] Buchfink B, Xie C and Huson D H 2014 Fast and sensitive protein alignment using DIAMOND *Nat. Methods* **12** 59–60

[272] Ferreiro D U, Komives E A and Wolynes P G 2014 Frustration in biomolecules *Q. Rev. Biophys.* **47** 285–363

[273] Uversky V N 2013 A decade and a half of protein intrinsic disorder: Biology still waits for physics *Protein Sci.* **22** 693–724

[274] Longo L M and Blaber M 2012 Protein design at the interface of the pre-biotic and biotic worlds *Arch. Biochem. Biophys.* **526** 16–21

[275] Li T, Fan K K, Wang J and Wang W 2003 Reduction of protein sequence complexity by residue grouping *Protein Eng. Des. Sel.* **16** 323–30

[276] Chan H S 1999 Folding alphabets *Nat. Struct. Biol.* **6** 994–6

[277] Solis A D 2015 Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins *Proteins Struct. Funct. Bioinforma.* **83** 2198–216

[278] Wolynes P G 1997 As simple as can be? *Nat. Struct. Biol.* **4** 871–4

[279] Dokholyan V N and Dokholyan N V. 2004 What Is the Protein Design Alphabet? *Proteins Struct. Funct. Genet.* **54** 622–8

[280] Betancourt M R and Onuchic J N 1995 Kinetics of proteinlike models: The energy landscape factors that determine folding *J. Chem. Phys.* **103** 773–87

[281] Wang J and Wang W 1999 A computational approach to simplifying the protein folding alphabet *Nat. Struct. Biol.* **6** 1033–8

[282] Sear R P 2004 Highly specific protein–protein interactions, evolution and negative design *Phys. Biol.* **1** 166–72

[283] Alberts et al B 2007 *Molecular Biology of the Cell* (Garland Science)

[284] Heizer Jr. E M, Raiford D W, Raymer M L, Doom T E, Miller R V and Krane D E 2006 Amino acid cost and codon usage biases in six prokaryotic genomes: a whole genome analysis *Mol. Biol. Evol.* **23** 1670–80

[285] Drummond D A, Bloom J D, Adami C, Wilke C O and Arnold F H 2005 Why highly expressed proteins evolve slowly *Proc Natl Acad Sci U S A* **102** 14338–43

[286] Koehl P and Levitt M 2002 Protein topology and stability define the space of allowed sequences *Proc. Natl. Acad. Sci. U. S. A.* **99** 1280–5

[287] Lobkovsky A E, Wolf Y I and Koonin E V. 2010 Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc. Natl. Acad. Sci. U. S. A.* **107** 2983–8

[288] Pál C, Papp B, Lercher M J, Pál C, Papp B,

Papp B, Lercher M J, Lercher M J, Pál C, Papp B, Papp B, Lercher M J and Lercher M J 2006 An integrated view of protein evolution. *Nat. Rev. Genet.* **7** 337–48

[289] Gilbert D, Viksna J, Gilbert D, Viksna J, Gilbert D, Viksna J and Gilbert D 2007 Structural bioinformatics Assessment of the probabilities for evolutionary structural changes in protein folds *Bioinformatics* **23** 832–41

[290] Meyerguz L, Kleinberg J and Elber R 2007 The network of sequence flow between protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **104** 11627–32

[291] Coluzza I, MacDonald J T, Sadowski M I, Taylor W R and Goldstein R a. 2012 Analytic Markovian rates for generalized protein structure evolution ed A Tramontano *PLoS One* **7** e34228

[292] Sotriffer C A, Flader W, Winger R H, Rode B M, Liedl K R and Varga J M 2000 Automated docking of ligands to antibodies: methods and applications *Methods* **20** 280–91

[293] Fahmy A and Wagner G 2002 TreeDock: a tool for protein docking based on minimizing van der Waals energies *J. Am. Chem. Soc.* **124** 1241–50

[294] Poma A, Turner A P F and Piletsky S A 2010 Advances in the manufacture of MIP nanoparticles *Trends Biotechnol.* **28** 629–37

[295] Piletska E V, Guerreiro A R, Whitcombe M J and Piletsky S A 2009 Influence of the polymerization conditions on the performance of molecularly imprinted polymers *Macromolecules* **42** 4921–8

[296] Ye L and Mosbach K 2008 Molecular imprinting: synthetic materials as substitutes for biological antibodies and receptors *Chem. Mater.* **20** 859–68

[297] Alexander C, Andersson H S, Andersson L I, Ansell R J, Kirsch N, Nicholls I A, O'Mahony J and Whitcombe M J 2006 Molecular imprinting science and technology: a survey of the literature for the years up to and including 2003 *J. Mol. Recognit.* **19** 106–80

[298] Yan S, Fang Y and Gao Z 2007 Quartz crystal microbalance for the determination of daminozide using molecularly imprinted polymers as recognition element *Biosens. Bioelectron.* **22** 1087–91

[299] Whitcombe M J, Alexander C and Vulfson E N 1997 Smart polymers for the food industry *Trends Food Sci. Technol.* **8** 140–5

[300] Mosbach K and Ramström O 1996 The emerging technique of molecular imprinting and its future impact on biotechnology *Nat. Biotechnol.* **14** 163–70

[301] Wulff G and Sarhan A 1972 Use of polymers with enzyme-analogous structures for resolution of racemates *Angewandte Chemie-International Edition* vol 11 (WILEY-V CH VERLAG GMBH MUHLENSTRASSE 33-34, D-13187 BERLIN, GERMANY) p 341

[302] Takagishi T and Klotz I M 1972 Macromolecule-small molecule interactions; introduction of additional binding sites in polyethyleneimine by disulfide cross–linkages *Biopolymers* **11** 483–91

[303] Arkin M R and Wells J a 2004 Small-molecule inhibitors of protein–protein interactions: progressing towards the dream *Nat. Rev. Drug Discov.* **3** 301–17

[304] Clackson T and Wells J A 1995 A hot spot of binding energy in a hormone-receptor interface *Science (80-.).* **267** 383–6

[305] Lim C W and Kim T W 2012 Dynamic [2]Catenation of Pd(II) Self-assembled Macrocycles in Water *Chem. Lett.* **41** 70–2

[306] Hino S, Ichikawa T and Kojima Y 2010 Thermodynamic properties of metal amides determined by ammonia pressure-composition isotherms *J. Chem. Thermodyn.* **42** 140–3

[307] Magyar C, Tüdős É and Simon I 2004 Functionally and structurally relevant residues of enzymes: are they segregated or overlapping? *FEBS Lett.* **567** 239–42

[308] Zikmanis P and Kampenusa I 2014 Relationship between Metabolic Fluxes and Sequence-Derived Properties of Enzymes *Int. Sch. Res. Not.* **2014** 1–9

[309] Wells J A and McClendon C L 2007 Reaching for high-hanging fruit in drug

discovery at protein-protein interfaces *Nature* **450** 1001–9

[310] Vanhee P, van der Sloot A M, Verschueren E, Serrano L, Rousseau F and Schymkowitz J 2011 Computational design of peptide ligands *Trends Biotechnol.* **29** 231–9

[311] Song C M, Lim S J and Tong J C 2009 Recent advances in computer-aided drug design *Brief. Bioinform.* **10** 579–91

[312] Lavecchia A and Giovanni C 2013 Virtual Screening Strategies in Drug Discovery: A Critical Review *Curr. Med. Chem.* **20** 2839–60

[313] Coluzza I, Creamean J, Rossi M J, Wex H, Alpert P A, Bianco V, Boose Y, Dellago C, Felgitsch L, Fröhlich-Nowoisky J, Herrmann H, Jungblut S, Kanji Z A, Menzl G, Moffett B, Moritz C, Mutzel A, Pöschl U, Schauperl M, Scheel J, Stopelli E, Stratmann F, Grothe H and Schmale D G 2017 Perspectives on the future of ice nucleation research: Research needs and Unanswered questions identified from two international workshops *Atmosphere (Basel).* **8** 138

[314] Cusick M E, Klitgord N, Vidal M and Hill D E 2005 Interactome: Gateway into systems biology *Hum. Mol. Genet.* **14** 171–81

[315] Emili A Q and Cagney G 2000 Large-scale functional analysis using peptide or protein arrays *Nat. Biotechnol.* **18** 393–7

[316] Finn R D, Coggill P, Eberhardt R Y, Eddy S R, Mistry J, Mitchell A L, Potter S C, Punta M, Qureshi M, Sangrador-Vegas A, Salazar G A, Tate J and Bateman A 2016 The Pfam protein families database: Towards a more sustainable future *Nucleic Acids Res.* **44** D279–85

[317] McGinnis S and Madden T L 2004 BLAST: At the core of a powerful and diverse set of sequence analysis tools *Nucleic Acids Res.* **32** 20–5

[318] Lever E and Sheer D 2010 The role of nuclear organization in cancer. *J. Pathol.* **220** 114–25

[319] Cheng R R, Morcos F, Levine H and Onuchic J N 2014 Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information *Proc. Natl. Acad. Sci.* **111** E563–71

[320] De Juan D, Pazos F and Valencia A 2013 Emerging methods in protein co-evolution *Nat. Rev. Genet.* **14** 249

[321] Kortemme T, Joachimiak L A, Bullock A N, Schuler A D, Stoddard B L and Baker D 2004 Computational redesign of protein-protein interaction specificity *Nat. Struct. Mol. Biol.* **11** 371–9

[322] Cocco S, Monasson R and Weigt M 2013 From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction *PLoS Comput. Biol.* **9** e1003176

[323] Dago A E, Schug A, Procaccini A, Hoch J A, Weigt M and Szurmant H 2012 Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis *Proc. Natl. Acad. Sci.* **109** E1733–42

[324] Ekeberg M, Lövkvist C, Lan Y, Weigt M and Aurell E 2013 Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models *Phys. Rev. E* **87** 012707

[325] Ho B K, Perahia D and Buckle A M 2012 Hybrid approaches to molecular simulation *Curr. Opin. Struct. Biol.* **22** 386–93

[326] Lunt B, Szurmant H, Procaccini A, Hoch J A, Hwa T and Weigt M 2010 Inference of direct residue contacts in two-component signaling *Methods in enzymology* vol 471 (Elsevier) pp 17–41

[327] Marks D S, Hopf T A and Sander C 2012 Protein structure prediction from sequence variation *Nat. Biotechnol.* **30** 1072

[328] Morcos F, Hwa T, Onuchic J N and Weigt M 2014 Direct coupling analysis for protein contact prediction *Protein Structure Prediction* (Springer) pp 55–70

[329] Morcos F, Jana B, Hwa T and Onuchic J N 2013 Coevolutionary signals across protein lineages help capture multiple protein

conformations *Proc. Natl. Acad. Sci.* **110** 20533–8

[330] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D S, Sander C, Zecchina R, Onuchic J N, Hwa T and Weigt M 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families *Proc. Natl. Acad. Sci.* **108** E1293–301

[331] Schug A, Weigt M, Onuchic J N, Hwa T and Szurmant H 2009 High-resolution protein complexes from integrating genomic information with molecular simulation *Proc. Natl. Acad. Sci.* **106** 22124–9

[332] Weigt M, White R A, Szurmant H, Hoch J A and Hwa T 2009 Identification of direct residue contacts in protein-protein interaction by message passing *Proc. Natl. Acad. Sci. U. S. A.* **106** 67–72

[333] Casari G, Sander C and Valencia A 1995 A method to predict functional residues in proteins *Nat. Struct. Mol. Biol.* **2** 171–8

[334] Nerattini F, Figliuzzi M, Cardelli C, Tubiana L, Bianco V, Dellago C and Coluzza I 2020 Identification of Protein Functional Regions *ChemPhysChem* **21** 335–47

[335] Mignon D, Panel N, Chen X, Fuentes E J and Simonson T 2017 Computational Design of the Tiam1 PDZ Domain and Its Ligand Binding *J. Chem. Theory Comput.* **13** 2271–89

[336] Thomson A R, Wood C W, Burton A J, Bartlett G J, Sessions R B, Brady R L and Woolfson D N 2014 Computational design of water-soluble α-helical barrels *Science (80-.).* **346** 485–8

[337] Huang P-S S, Oberdorfer G, Xu C, Pei X Y, Nannenga B L, Rogers J M, DiMaio F, Gonen T, Luisi B and Baker D 2014 High thermodynamic stability of parametrically designed helical bundles *Science (80-.).* **346** 481–5

[338] Chino M, Maglio O, Nastri F, Pavone V, DeGrado W F and Lombardi A 2015 Artificial Diiron Enzymes with a De Novo Designed Four-Helix Bundle Structure *Eur. J. Inorg. Chem.* **2015** 3371–90

[339] Gaillard T and Simonson T 2017 Full Protein Sequence Redesign with an MMGBSA Energy Function *J. Chem. Theory Comput.* **13** 4932–43

[340] Senior A W, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson A W R R, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones D T, Silver D, Kavukcuoglu K and Hassabis D 2020 Improved protein structure prediction using potentials from deep learning *Nature* **577** 706–10

[341] Li W, Persson B A, Morin M, Behrens M a., Lund M and Zackrisson Oskolkova M 2015 Charge-induced patchy attractions between proteins *J. Phys. Chem. B* **119** 503–8

[342] Boyken S E, Benhaim M A, Busch F, Jia M, Bick M J, Choi H, Klima J C, Chen Z, Walkey C, Mileant A, Sahasrabuddhe A, Wei K Y, Hodge E A, Byron S, Quijano-Rubio A, Sankaran B, King N P, Lippincott-Schwartz J, Wysocki V H, Lee K K and Baker D 2019 De novo design of tunable, pH-driven conformational changes *Science (80-.).* **364** 658 LP – 664

[343] Frederix P W J M W J M J M, Ulijn R V V., Hunt N T T, Tuttle T, Ulijn V R, Hunt N T T and Tuttle T 2011 Virtual screening for dipeptide aggregation: Toward predictive tools for peptide self-Assembly *J. Phys. Chem. Lett.* **2** 2380–4

[344] Van Teijlingen A and Tuttle T 2021 Beyond Tripeptides Two-Step Active Machine Learning for Very Large Data sets *J. Chem. Theory Comput.* **17** 3221–32

[345] Shmilovich K, Mansbach R A, Sidky H, Dunne O E, Panda S S, Tovar J D and Ferguson A L 2020 Discovery of Self-Assembling π-Conjugated Peptides by Active Learning-Directed Coarse-Grained Molecular Simulation *J. Phys. Chem. B* **124** 3873–91