Directed Enzyme Evolution: Advances and Applications

Book Chapter

Molecular modeling in enzyme design, towards *in silico* guided directed evolution

Emanuele Monza[1], Sandra Acebes[1], M. Fátima Lucas[1,2], Victor Guallar[1,3*]

[1]Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Jordi Girona 29, E-08034 Barcelona, Spain

[2]Anaxomics Biotech, Balmes 89, E-08008 Barcelona, Spain

[3]ICREA, Passeig Lluís Companys 23, E-08010 Barcelona, Spain

* Corresponding author

Victor Guallar, PhD
Life Science Department
Barcelona Supercomputer Center
Jordi Girona 29, E-08034 Barcelona, Spain
Phone: +34 93-413-7727
Fax +34 93-413-7721
Email: victor.guallar@bsc.es

Abstract

Directed evolution creates diversity in subsequent rounds of mutagenesis in the quest of increased protein stability, substrate binding and catalysis. Although this technique does not require any structural/mechanistic knowledge of the system, the frequency of improved mutations is usually low. For this reason, computational tools are increasingly used to focus the search in sequence space, enhancing the efficiency of laboratory evolution. In particular, molecular modeling methods provide a unique tool to grasp the sequence/structure/function relationship

of the protein to evolve, with the only condition that a structural model is provided. With this book chapter, we tried to guide the reader through the state of art of molecular modeling, discussing their strengths, limitations and directions. In addition, we suggest a possible future template for *in silico* directed evolution where we underline two main points: a hierarchical computational protocol combining several different techniques, and a synergic effort between simulations and experimental validation.

1. Introduction

Biotechnology needs catalysts that can work under harsh conditions, catalyze a broad range of substrates, generate maximum amount of product, and tolerate changes in the environment. Enzymes, which are biodegradable and reusable catalysts [1], in addition to remarkable reaction rates, can work in environmentally friendly pH and temperature ranges, and display control over stereochemistry and regioselectivity which makes them ideal for many applications [2,3]. When thinking about enzymes, people normally associate them to expressions such as "perfect catalysts" or "outstanding reaction rate". In fact, there are examples of enzymes that catalyze reactions at extremely high rates such as triose phosphate isomerase, superoxide dismutase or carbonic anhydrase [4]. These are often limited only by the rate of ligand diffusion into the active site (diffusion-controlled rate). Nevertheless, an extensive analysis by Bar-Even et al., of nearly 2000 enzymes, showed that the median maximal turnover rate value over all measured enzymes is about 10 s^{-1} nowhere close to the values of 10^5 or 10^6 normally associated with catalysts [5,6]. So, it would appear that natural enzymes are "just good enough" for the function they must perform in a given organism [7]. One might conclude that if they had evolved to their optimum performance then trying to improve them (from a kinetic point of view) would be attempting the impossible. On the contrary, as seen by the distribution of reaction rates, k_{cat}, most enzymes function at a lower rate than the diffusion-limit and thus, there is space to further increase their kinetic properties to meet industrial needs. Additionally, we need enzymes capable of catalyzing reactions for which no known enzymes exist, to work with different substrates and for particular conditions that are industrially convenient and economically advantageous. For all these reasons, in most cases, we cannot just use enzymes as they are found but instead we need to change their physical-chemical and functional properties. This is one of the reasons why engineering enzymes for biocatalysis is an incessantly growing field [8-11].

In Nature, enzymes have evolved over millions of years to meet specific demands and operate under tight *in vivo* regulation. Their degree of adeptness includes diverse criteria such as: which substrates they accept, the effective reaction rate, the environment in which they function and how well they tolerate changes in it, inactivation by their own products, etc. These characteristics are precisely the ones that scientist wish to control to their own advantage. Some of the earliest attempts to modify enzymes required a deep knowledge of complex structure/function relationships and (to the authors' contentment) computer simulations have played an important part in it [12,13]. Since the pioneering work [10,14,15] in computationally designed protein sequences (with experimental validation) many remarkable achievements have been obtained. Interesting work includes predicting sequence changes that alter atomic packing arrangements in buried protein regions or the creation of new metal

binding sites which may have many applications along with potential improvement in protein stability [16-18]. In addition to being able to correctly predict changes in protein structure, there has been, of course, a large interest in altering proteins, through computational techniques, to create new function or adapt them to particular conditions. Rational protein design, which involves modification of specific amino acids in the protein's three-dimensional (3D) structure with previous structural/mechanistic knowledge, can be used to alter specificity, stability, selectivity and activity. Literature contains a vastness of examples of rationally designed proteins (which we do not presume to cover here) including creating new recognition [19-26], improve protein stability [27-29], and protein-protein [30-34] or protein-DNA interactions [35,36]. We can find procedures to engineer a protein that binds a specific cofactor [37] or a calcium-binding site [38,39], redesign an enzyme by stabilizing the transition state [40] or create new activity from scratch [41].

A special mention involves the design of new proteins from scratch, commonly known as *de novo* design, and literature displays many truly interesting examples of new proteins [42-45]. Currently one of the most common strategies to design new enzymes is based on encountering complementary active sites for the transition states of interest [46,47]. Despite the success of *de novo* design in providing novel structures and activity, its difficulties in achieving fast kinetics make it still preferable to modify templates available in Nature for the desired chemistry. Indeed, a recent computational study pointed out how target reactivity can be one mutation away from a non-enzymatic protein (if well picked) [48]. Due to the scope of this book, we refer the reader interested in *de novo* design to recent studies on this topic [49].

Despite many promising studies, rational computational protein redesign has its limitations: it requires a reliable three-dimensional structure of the system of interest and an in-depth comprehension of the catalytic mechanism; understanding the relationships between a protein's primary sequence, its three-dimensional structure and its function is therefore a fundamental goal. Regrettably, our knowledge of enzyme activity is still incomplete which makes our attempts to modifying them often limited. Detailed understanding of the enzymatic structure/function relation is, however, not necessary in directed evolution, an alternative engineering technique based on massive mutations and selective evolution.

Directed evolution (DE), has proven to be a powerful tool for adapting enzymes to wider applications [50-53]. Briefly, in DE diversity is first created through mutagenesis or recombination, followed by screening for improvements in desired properties. One of the main advantages of DE is most certainly that it does not require a thorough understanding of structure/function relationships, unlike rational or *de novo* design. The introduction of random mutations throughout the gene allows the discovery of mutations that could be difficult to predict with studies based on structure-function knowledge (mostly focused at the active site region). However, the low frequency of improved mutations, some experimental bias, and the combinatorial explosion of possibilities limits this technique. Furthermore, DE requires the development of high-throughput screening and not all processes can be

adapted. The methodologies and achievements of directed evolution were already discussed in other sections of this book and will not be included here. Also we refer the reader to interesting reviews [54-59].

A remarkable observation of many DE experiments is that the location of the beneficial mutations varies considerably. For example, most modifications in enantioselectivity or substrate specificity are located in the vicinity of the active site or in the access/exit of reactants/products [58,60,61]. Stability and activity however can be affected by mutations in any part of the protein, close or far from the active site [62], increasing significantly the number of possible mutations. To avoid screening massive number of mutations, one can reduce the region to explore by using functional information (from point mutations, random mutagenesis or deduction from sequence alignments) or when structural information exists (by visual inspection, analysis, etc.), it would be advantageous to exploit this by concentrating mutations where they might be the most effective [62]. Methods such as saturation mutagenesis (where all other 19 amino acids are tested) on specific positions, generally near the active site, can increase the probability of finding beneficial mutations [63-65]. This approach is particularly advantageous when a high-throughput screening method is not available. Generally known as **semi-rational approaches**, these are based on "smart" libraries that, in principle, should have a higher success rate and try to overcome the limitations of the directed evolution and rational design [66-69].

Although it is true that many computational approaches exist to complement DE experiments [70,69,71], the scope of this chapter is to center on how physics based molecular modeling can aid in laboratory DE. For this reason, sequence-based strategies that use evolutionary information or statistical data from previous DE rounds will not be explored. These often use phylogenetic analyses and multiple sequence alignments for exploring the amino acid conservation and relationships between homologs protein sequences [67,72-78]. Instead we will center our attention on how computations and structural information may aid and focus mainly in physics-based methods to assist in the improvement of three major aspects in enzyme design: catalytic rate constants, protein stability and protein-ligand binding processes.

The atomic/molecular detailed computational exploration of a protein's amino acid sequence space is a complex problem. As in most simulation fields, a compromise between sampling quality and quantity is necessary. Sampling quality involves construction of the models together with the energy and scoring functions necessary to rank them and evaluate molecular interactions, topics extensively reviewed previously [63,79-92]. An energy function describes the internal energy of the protein and its interactions with the environment such as other proteins, substrates and solvent, aiming at reproducing the features of the folded protein [34,84,93]. The level of theory used in these and their parameters vary considerably but most implementations include bonded (bonds, angles and torsions), non-bonded terms (van der Waals and electrostatics) and solvent components. Associated to the energy functions is the ability to efficiently score a large number of protein structures and protein-ligand interactions. Scoring functions are used to assess the quality of the designed protein, help select the preferred sequence and the lowest energy protein-substrate complex. Just as the energy functions, also these vary considerably and can be

statistics or empirically-based methods such as DMutant or PopMuSiC or physics-based and rely on the derivation of energy terms from basic principles to calculate free energy changes [94-97].

The second key aspect involves the system (model) sampling. Given the large number of degrees of freedom, including possible mutations and structural changes associated with them, sampling near-native protein conformations is difficult. Moreover, in situations where protein-ligand interactions exist, sampling must also extend to all (relevant) protein-ligand conformations. And if we don't consider these issues to be enough of a headache, scoring and sampling are not independent! So, to overcome these limitations it is essential to introduce many approximations. Strategies to limit the sampling include restricting the backbone and side-chain degrees of freedom [34,82]. In most protein design strategies, sampling is simplified by using a fixed backbone which is normally obtained from an experimentally determined protein structure [98] or a high quality homology model. Although controversy exists on the importance of dynamics in catalysis [99-101], currently we see more and more cases where backbone flexibility is being taken into account [18,102-106]. As shown below, development in molecular dynamics (such as high-throughput molecular dynamics (HTMD), steered-MD, etc.) and Monte Carlo techniques, are gaining importance in enzyme engineering.

As mentioned, this book chapter will center on physics-based methods to assist in the improvement of catalytic rate constants, protein stability and protein-ligand binding processes. Before entering in these three topics, we refer to Figure 1 and Table 1 for a quick guide describing the main computational methods and models being used for these purposes (a guide for non-experts in theoretical modeling). Finally, we conclude this book chapter by introducing our perspective on how we believe these techniques will aid in future enzymatic directed evolution.

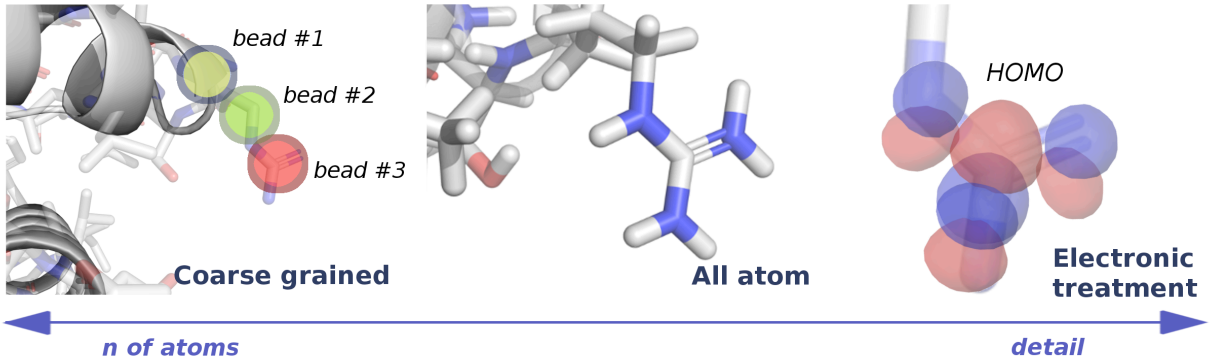

Fig. 1 Scheme showing three different levels of granularity used in molecular modeling: coarse grained, all-atom and electronic. In the coarse grained model the smallest particle is a bead that includes condensed information on a set of atoms. All-atom, as indicated by the name, uses the atom as the smallest unit while in an electronic treatment electrons and nuclei are explicitly included. Here we show the highest energy molecular orbital (HOMO) only possible in an electronic treatment of the system.

Methodology Guide

Length-size based models (see also Figure 1)

Coarse grained (CG) model. A group of atoms is described by a bead enclosing the properties of the aggregation. For example, according to the MARTINI model [107], aminoacids are represented with one to four beads, classified as charged, polar, nonpolar or apolar, and also subdivided depending on their hydrogen bonding capacity. Reduction in the number of beads decreases the number of pairwise interactions thus increasing the speed of the simulation

All-atom model. All the atoms of the system are included in the model, where the energy function used (see below) must describe their interaction. Electrons and the nuclei information is condensed to a single particle that must contain an averaged description of those properties.

Electronic treatment model. Each atom is described as a nucleus with its electrons, requiring, for its description, approximate solution of the Schrödinger equation.

Physical theoretical methods

Molecular mechanics (MM) [108]. These methods perform a classical description where atoms (or beads in a coarse grained model) are represented as spheres (or spheroids) connected by bonds, behaving as springs. Based on MM arise several computing simulations such as molecular dynamics, Monte Carlo and docking methods.

Force field. Set of parameters that define the property of atoms or beads (predefined with a partial charge and radius) and the energy function describing their interactions in MM methods. Typically they include bonding: bond, angle and torsion, and non-bonding: electrostatic and van de Waals terms.

Elastic network model [109] (ENM). Describes the collective dynamics of proteins by an elastic network, typically using a reduced set of nodes, such as alpha carbons.

Molecular Dynamics [110] (MD). Simulate the motion of a model accordingly to the classical Newton's equation. Most MD software uses force fields to describe the properties of atoms and its interactions. With current high performance computing (HPC) simulations can be expanded up to the millisecond time-scale [111] and few millions of atoms; typical values however, involve thousands of atoms and hundreds of nanoseconds.

Monte Carlo simulations (MC). The dynamics of the system are obtained by random (stochastic) motion of the system to assemble a non-time dependent trajectory [112]. As in MD, it is mostly based in a force field description of the model.

PELE [113]. The protein energy landscape exploration (PELE) software is a Monte Carlo based technique including protein structure prediction techniques (such as ENM) capable of quickly modeling protein dynamics and protein/DNA-ligand interactions [114,115].

Docking simulations. These propose the preferred relative bound orientation between molecules, mostly used in protein-ligand (substrate) or protein-protein interactions. Usually, docking methods first provide several conformations which are then classified by scoring functions.

Scoring functions. Are mathematical functions that predict the strength of intermolecular interactions [116]. Scoring functions are mostly parameterized from MM force fields, empirical data or knowledge based functions.

Rotamer library. Contains a restricted number of the most probable conformations (torsion angle

values) for a molecule, mostly applied to amino acid side chains, protein backbone and ligands. They are built from experimental structural data or from accurate quantum simulations (for example in ligands). When used with sampling methods they accelerate the exploration by adopting discrete states instead of continuous values.

Quantum mechanics (QM). These methods are based on solving the Schrödinger equation (normally using approximations) under an electronic model description of the system. The solution provides the wave function which fully describes the system: the electronic distribution, energy and the gradients to describe the motion of the system. The main limitation of QM methods is their high computational cost, limiting the system's size and simulation speed.

***Ab initio* methods.** These are quantum mechanics methods which parameters are obtained exclusively from first principles solution of the equations (still under approximations) but without any usage of parameterized data (see semiempirical methods).

Semiempirical methods. Referred to quantum mechanics methods that use parameters derived from experimental data (or *ab initio* calculation), typically for the parameterization of the electron-electron interaction terms (the most expensive to compute). Thus, they are less computationally expensive and faster than *ab initio* ones, capable of dealing with large systems. Their lack of accuracy, especially when fragments are not in the parameterized data set, is their main limitation.

QM/MM. This methodology is a combination of QM and MM methods to handle (large) biological all-atom systems [117]. One part of the system where we require an electronic description, such as the active site in an enzyme, is treated at the QM level and the rest of the model (remainder of the protein, solvent, etc.) is treated at the MM level.

2. State of the art of molecular modeling in protein design

We provide here a general view of recent computational work on protein design. We do not aim to review all studies produced in the field but to underline several ones which we believe to be important for future developments of *in silico* DE approaches.

2.1 Protein stability improvement

Understanding and quantifying the effect of mutations on the thermodynamical stability of a protein is of paramount importance for industrial applications. Two of the most popular tools to prepare and score mutated proteins are Rosetta [118,119] and FoldX [27]. After introducing a mutation, the protein's torsional degrees of freedom (usually sidechain rotamers) are optimized using an energy function that estimates the folding free energy for the created variant. Such energy functions depends on: i) physics-based terms, which account for van der Waals, hydrogen bond, solvation and electrostatic energies; ii) knowledge-based contributions, which determine the probability of a given rotamer according to the protein data bank (PDB) statistics [120]. Apart from these common energy terms, these functions have unique features. For example, Rosetta approximates the free energy change in the unfolded state due to a mutation with context-independent reference energies for each residue [121], On the other hand, FoldX explicitly estimates the entropy cost to restrict a rotamer in the native state [27]. The relatively low computational cost of these protocols permits to generate and score a large number of mutations in a short time. As shown by Potapov et al., the accuracy/cost trade-off is such that these tools can reproduce overall trends, and

therefore suggest stabilizing mutations with acceptable probabilities, but they are not good enough to provide detailed results [122].

Following Potapov accuracy assessment, Kellogg et al. tested the ability of Rosetta to score mutations combining several energy functions and sampling methods with variable resolution [119]. As a main result, the authors concluded that the choice of the sampling algorithm should be tuned with the resolution of the energy function adopted. In other words, an accurate energy function performs better on a finer sampling; likewise, roughly sampled structures should be scored by smoother functions which can tolerate steric clashes better. Still, flexible backbone protocols improved small to large residue mutations, where significant structural changes can occur. In addition, the authors found that conformational sampling was still insufficient to recover the crystal conformation when a large to small hydrophobic residue mutation was introduced, due to poor packing. Larger errors were found when the polarity of the residue drastically changed upon mutation, which suggests poor trade-off between polar desolvation and buried polar interactions. The lack of explicit water molecules and ligand contacts was another factor in some failed predictions. Finally, the lack of a context-dependent unfolded state modeling (a given mutation was considered to have the same effect on the unfolded state independently of the environment [121] was considered as a source of error, although not a major one. In fact, a free energy variation of the unfolded state upon mutation might change protein stability as well as a variation in the folded state. However, a recent paper shows that an accurate conformational and energetic characterization of the unfolded protein is not trivial and its inclusion in protein stability scoring significantly worsened the prediction [123].

The entropic scoring in FoldX [124,27] only takes into account the change in conformational entropy, which depends on the number of accessible conformers in the unfolded state and their probabilities [124]. Although this entropic variation dominates folding [125], large discrepancies in vibrational entropy (the intrinsic entropy of a given protein conformer [124] have been calculated between thermophilic and mesophilic proteins [126]. Therefore, the thoughtful inclusion of a vibrational entropy contribution in protein design free energy functions might pay-off. Najmanovich and coworkers implemented this strategy in the ENcoM server [126], where they combine FoldX [27] with their ENcoM protocol to rapidly estimate vibrational entropy. ENcoM combines ENM techniques with a pairwise atom-type non-bonded interaction term to include the specific nature of amino acids [127].

In an attempt to quantify free energies more rigorously, de Groot and coworkers employed alchemical free energy MD simulations to score 109 mutants of ribonuclease barnase [128]. In this technique, sampling a convenient number of unphysical ("alchemical") intermediates renders a rigorous evaluation of the free energy difference (ΔG) between two states (e.g. wild type and mutant protein). Unfolded state's free energy differences were calculated using a generic Gly-XXX-Gly peptide with capped termini. This choice provides a universal, albeit less accurate and context-independent, reference state whose values need to be calculated only once and then are stored as a database. The overall Pearson's correlation coefficient with experimental values was 0.86, providing ~72 % of the predicted values within 1 kcal/mol of the experimental one when using 30 ns of simulation time. Notably, most of this accuracy (65 %) is retained with only 5 ns. The generality of this accuracy/cost ratio will need to be tested against a wider benchmark of mutations. Larger errors were detected for mutations that introduced changes in the

electrostatics of buried residues or large structure fluctuation: mutations to glycine, involving bulky and/or well packed residues, etc..

Due to the impossibility of scoring the entire sequence (mutation) space, several strategies have been developed to focus the search in smaller regions. These include: i) the identification of flexible backbone sites which can be rigidified [129,130] introducing salt bridges [131] and/or disulphide bonds [132]; ii) the optimization of surface charge-charge interactions [133-135]; iii) the optimization of core packing [136]; iv) the removal of unsatisfied buried polar groups [137]; v) the localization of critical residues in the active site entry tunnels, especially for co-solute tolerance, with MD [138] or other algorithms like our in-house software PELE [113].

Recently, Wijma and coworkers developed, and applied with success, a mixed approach which aims to obtain highly thermostable protein variants in a short time with minimum experimental screening [139]. In their computational workflow, potentially stabilizing mutations were firstly produced and scored with Rosetta [119] and FoldX [27]. To minimize the risk of affecting catalysis, only residues beyond 10 Å of the substrate were mutated. Mutations were considered potentially stabilizing if $\Delta\Delta G_{fold} \leq -5$ kJ/mol or if $|\Delta\Delta G_{fold}| < 5$ kJ/mol and the mutation type was contained in the set XXX→Arg, XXX→Pro, Gly→XXX. These were then filtered to avoid undesired, typically destabilizing features such as increased unsatisfied hydrogen bond donors and acceptors or hydrophobic surface exposure to water. Then, multiple short MD simulations were used to discard variants with increased backbone flexibility. Finally, variants with experimentally confirmed higher thermostability and preserved activity were combined in the lab. This computational hierarchical workflow helps to unmask false positives (~50 % of the potentially stabilizing mutations), aiding to focus on reliable mutations; it is, therefore, a plausible strategy for future computer-aided directed evolution of thermostable proteins. The main drawback is the exclusion of mildly damaging mutations that could be coupled synergically to other to improve thermostability.

As reported in a recent review [140], there is still substantial room for improvement of structure- and physics-based (thermo)stability design. This will likely pass through a strong synergy of computational and experimental efforts to improve our understanding of protein stability. In addition, significant work will have to center on developing more accurate energy functions, including polarization, solvation and vibrational entropy terms. These methodological developments will necessarily have to couple with improvement of sampling algorithms, including a more effective modeling of unfolded state changes.

2.2 Protein-ligand binding redesign

Whether we are talking about enzymes or receptors, they all share a common feature: at some stage a protein-ligand recognition process must occur. These are however, notoriously slow and complex processes that require extensive sampling of the protein-ligand dynamics which in many cases includes induced-fit protein conformational changes. The accurate *in silico* design of protein-ligand interactions is thus a challenging step [141] toward the engineering of proteins for therapeutic [142] and enzymatic purposes [140]. Its difficulty roots in the low tolerance to error due to the reduced number of protein-ligand interactions. In addition, these are largely dominated by polar interactions, which are very sensitive to small changes in geometry [143]. It is worth noting that, despite the small size of the

protein-ligand interface, we still face a huge number of possible combinations in sequence space (for 10 positions there are ~10^{13} sequences).

In a recent attempt to benchmark the state of art of computational protein-ligand interactions design, Allison and coworkers tested Rosetta's [12] sequence recovery (with respect to the wild type) capability over a set of 43 protein-ligand complexes [143]. The Rosetta protocol involved simultaneous ligand motion and sidechain rotamer discrete optimization. Overall, sequence recovery was more successful when: i) a near-optimal pose was inputted and subjected to limited sampling instead of blindly searched; ii) the ligand was small, non-polar and rigid; iii) the binding pocket packing was neither overcrowded, nor poor. Another interesting result was the significantly higher recovery for non-polar residues. The authors suggested that new terms should be added to the energy function to correct this bias toward non-polar interactions [143]. However, this bias could be an artifact of poor sampling, which might limit the accuracy of polar interactions estimation (see above). In fact, other suggested areas of improvement were the use of continuous, instead of discrete, sampling of backbone and sidechain rotamers [144], concerted rotation of the backbone of two adjacent residues allowing larger sidechain motion (the so-called backrub motion) [145,146] and the calculations of partition functions providing a link between molecular behavior and bulk thermodynamic quantities over structural ensembles [147,148,102]. All these features are grasped by OSPREY [144], a recent open source solution to protein design which includes graphic processing units (GPU) acceleration [149], dead-end elimination algorithm [150,151] and the K* method [151]. K* aims to approximate the partition function of the bound and unbound states over an ensemble of structures; their ratio provides an estimation of the binding constant. The conceptual advantage of this methodology is a mathematically rigorous, albeit approximate, free energy difference calculation that explicitly simulates the free ligand and protein. Consequently, ligand and binding site pre-organization are, in principle, included in the calculation. On the other hand, this absolute free energy calculation is neither accurate nor efficient for systems with a large number of energy minima [152], requiring extensive sampling to reduce errors. However, the error of the method most likely compensates between complex and free monomers calculations, making this strategy a valuable tool for fast free binding energy simulations. Regardless of the methodology chosen, an effort to produce new experimental data will be fundamental to benchmark these high-throughput computational protocols and improve their predictive power [86].

An inaccurate description of the binding site is yet other possible sources of error. Indeed, some mutations could shift the pK_a of ligand's and protein's titratable sites or introduce a new titratable residue. Therefore, the system should be prepared thoroughly before computational mutagenesis and quick pK_a predictors [153] should be used to treat critical mutations. On the other hand, for situations where pK_a is close to the experimental pH conditions, simulation of all the possible combinations for the ambiguous titratable sites is required. For instance, a recent laccase design effort required the simulation of sinapic acid in both protonation states[154]: if one of the two protonation would have been picked, activity changes would have been missed since they mostly involved one of the two accessible protonation states. Finally, waters in the binding site might have an important role in binding and their neglect could affect the quality of the calculation [155].

A way to filter and correct designs is based on MD simulations [156]. Many features of designed protein-ligand complexes can be inspected with this technique, including hydrogen bond geometries, binding site structural integrity, solvent exposure and binding site pre-organization. In particular HTMD [157] was used by Baker and coworkers to filter computationally designed candidates according to the fraction of near attack conformations (NAC), structures that resemble the transition state (TS) [158]. Moreover, MD can help in the future to discern long range effects. In fact, it has been recently used to highlight the impact of distant mutations on active site pre-organization in evolved enzymes [159,160]. Furthermore, proteins are dynamical entities organized in a network of correlated fluctuations whose changes can significantly affect binding at large distances [161]. Importantly, such network can be identified through a correlation matrix (which quantifies the correlation degree of a pair of amino acids) and partitioned in communities of highly correlated residues, giving insights on allosteric interactions [162]. These analyses, with contact and hydrogen bond maps, might be used in the future to identify regions whose motion influences the binding site's dynamics (for example making the sidechain of a catalytic residue too flexible), which can then be subjected to mutagenesis in the lab.

A possible error when studying protein-ligand association arises when focusing mostly on the binding site, as some mutations along a possible entrance channel could hinder the ligand entrance/exit process. Sampling algorithms like PELE [163,113] can help to recognize such mutations. Its combination of ENM, sidechain prediction, ligand perturbation (translations and rotations), all-atom minimization and implicit solvation make it a suitable tool to quickly map the whole ligand migration process with good accuracy, taking protein flexibility into account [164-166]. Analogous MD based techniques, such as HTMD [157], RAMD [167] and steered MD [168] ,have addressed this problem. These, provide more accurate simulations, as it explicitly models water molecules, but also significantly more expensive ones, difficult to apply to massive mutation studies. Additional tools such as Fpocket [169] or CAVER [170] are widely used to quickly identify tunnels and cavities. These techniques, however, do not explicitly simulate ligand or protein dynamics.

Finally, quantum chemical calculations can be used to validate promising mutations, especially when charge transfer and polarization have an important role in the binding process. Mixed QM/MM schemes [171], widely used to model large systems, can significantly improve protein-ligand binding prediction directly, through explicit energy calculations [172], or indirectly [173] by re-calculating ligand's atomic charges in an attempt to model ligand polarization effects. An alternative, more accurate but slower approach to large systems, is the fragment molecular orbital (FMO) method [174]. FMO divides a system in N non-overlapping fragments (e.g. one for each protein residue and ligand) and calculates the total energy as the sum of one-body fragment energies and two-body interaction energy corrections, providing a $\sim N^2$ scalable fully parallelizable QM calculation. Jensen and coworkers used this methodology to energetically score the cleavability of peptides by HIV1-protease [175] by looking at the protein-peptide interaction energy.

2.3 Catalytic rate constant enhancement

The improvement of catalytic activity of bond breaking/formation, passes through the modeling of the TS of the slowest chemical step of the targeted reaction; see below for electron transfer (ET) processes. The problems with the design of optimal activation energies are multiple: i) the energy function should be sensitive enough to effectively discriminate between the reactant (substrate) and the TS, whose charges and geometries might be similar; ii) the nature of the TS can change upon mutation; iii) activation energies are very sensitive to molecular geometry changes.

In OptZyme the TS is approximated by a transition state analogue (TSA), a stable molecule which electronically and sterically resembles the TS [176]. Once the TSA and the substrate are docked in the active site, two parameters drive mutant selection: the enzyme-substrate (ES) and the enzyme-TSA interaction energies. These last two quantities are obtained using classical force fields. Through a number of conceptual and mathematical simplifications, the authors show that the former energy correlates with the Michaelis constant (K_M) while the latter with the specificity constant (defined as the ration between k_{cat} and K_M). Although it yielded satisfactory correlations for their specific case, it is worth noting that these have no general validity. If the rate constant is comparable or much higher than the ES dissociation constant the pre-equilibrium approximation is no longer valid. Then, the Michaelis constant cannot be approximated by the ES dissociation equilibrium constant (K_D) [177]. Notwithstanding, ES and enzyme-TSA interaction energies are still valuable tools for a fast semi-quantitative evaluation of enzyme variants.

Khersonsky et al. combined computational design and directed evolution to optimize a previously designed Kemp eliminase [178]. As in the previous example, the classical (force field) interaction energy between the enzyme and an explicit TS model was the parameter to be optimized during the sequence exploration. The TS model, however, was derived from QM calculations in solution including key catalytic residues. The authors individuated three main factors for the improved activity: a more favorable electrostatic environment, a better packed active site and a higher degree of active site pre-organization.

Although the last two methods, based on classical interaction energies, allow to test a big number of mutants, they both present a conceptual limitation: the use of the enzyme-TS (model) interaction energy, which is size-dependent (extensive property), to score the activation energy. To correct for this approximation, the ES interaction energy should be taken into account, providing a relative value. However, poor sampling and inaccurate energy functions might introduce uncertainties that could make its introduction useless (as it often happens in molecular mechanics/generalized born surface area (MM/GBSA) free energy calculations [179]). Still, they are currently the best approach to test a large number of mutations and find promising protein variants which can then be filtered with MD and quantum chemical methods [156].

The only way to properly calculate the activation energy barriers is the introduction of a QM methodology, capable of describing the electronic effects associated with TS formation. QM/MM schemes, which have been widely applied to elucidate enzymatic reaction mechanisms [180], have been employed to rescore promising candidates [156,180]. A remarkable example is the hierarchical approach of Zheng et al. used to design a cocaine

hydrolase [181]. Firstly, the reaction coordinate and the TS of the rate-determining limiting step are determined in the wild type. Then, many mutations are scored according to their protein-TS interaction energy; if this is lower than the wild type, a QM/MM calculation along the reaction coordinate is used to estimate the energy barrier. To allow fast computation, the authors use a reaction coordinate approach, freezing at each step the reactive coordinate and minimizing all the other degrees of freedom. A main drawback is still the need of extensive sampling, which makes the presented methodology too expensive for a general use.

A cheaper alternative to QM/MM calculations is empirical valence bond (EVB) [182]. EVB is based on a semi-empirical Hamiltonian which describes reactants and products with their resonance structure (explicitly defining atom connectivity). Although EVB energies are less accurate than *ab initio* and DFT QM/MM methods, free energy calculations are orders of magnitude quicker and still can achieve accurate results [183,184], making EVB a suitable tool to score a bigger number of mutants.

In the attempt to describe the entire enzyme or a large part of it with QM calculations, Jensen and coworkers approximated the reaction coordinate with the linear interpolation between reactant and product optimized geometries and calculated each point with semi-empirical methods [185-187]. These fast electronic calculations, united with algorithms for large scale systems such as FMO [174,188] or the much faster Effective Fragment Molecular Orbital (EFMO) [189], make "*ab initio* biochemistry" [190] closer, albeit still far away for design purposes.

In the particular case of oxidoreductases, where charge transfer processes dominate, additional complexity is added to the protein design problem. Electrons must move from a donor to an acceptor, sometimes through a long range electron transfer. According to Marcus' theory, the ET rate constant [191] depends on three parameters: i) electronic coupling, the probability to jump from the reactant to the product's diabatic state, which exponentially depends on the donor-acceptor distance; ii) reorganization energy, which is the energy penalty that accompanies electron transfer; and iii) the free energy difference between product and reactant (driving force). The ET rate constant has a maximum when the sum of reorganization energy and driving force equals zero. Although accurate QM/MM methodologies have been developed to study electron transfer rate in proteins [192,193], their use in enzyme design is limited by their computational cost. To overcome this barrier, we have recently developed a new methodology to approximately evaluate ET rates, which combines fast conformational sampling [163] and quick QM/MM spin density calculations and has been used to evaluate the activity of laccases variants [194,154]. While PELE provides a thorough and quick mapping of enzyme's and substrate's dynamics, substrate's spin density permits to promptly score the relative changes in driving force upon mutation (the higher the spin density, in principle, the higher the driving force). In the same spirit, we rationally improved the oxidation rate of 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonic acid) (ABTS) by a highly stable manganese peroxidase (Fig. 1) relying on electron coupling calculations, after the entire protein-ligand migration studies were performed [195]. In this case, it was assumed that the driving force and the reorganization energies did not change upon mutation, which can be a reasonable approximation in surface ET.

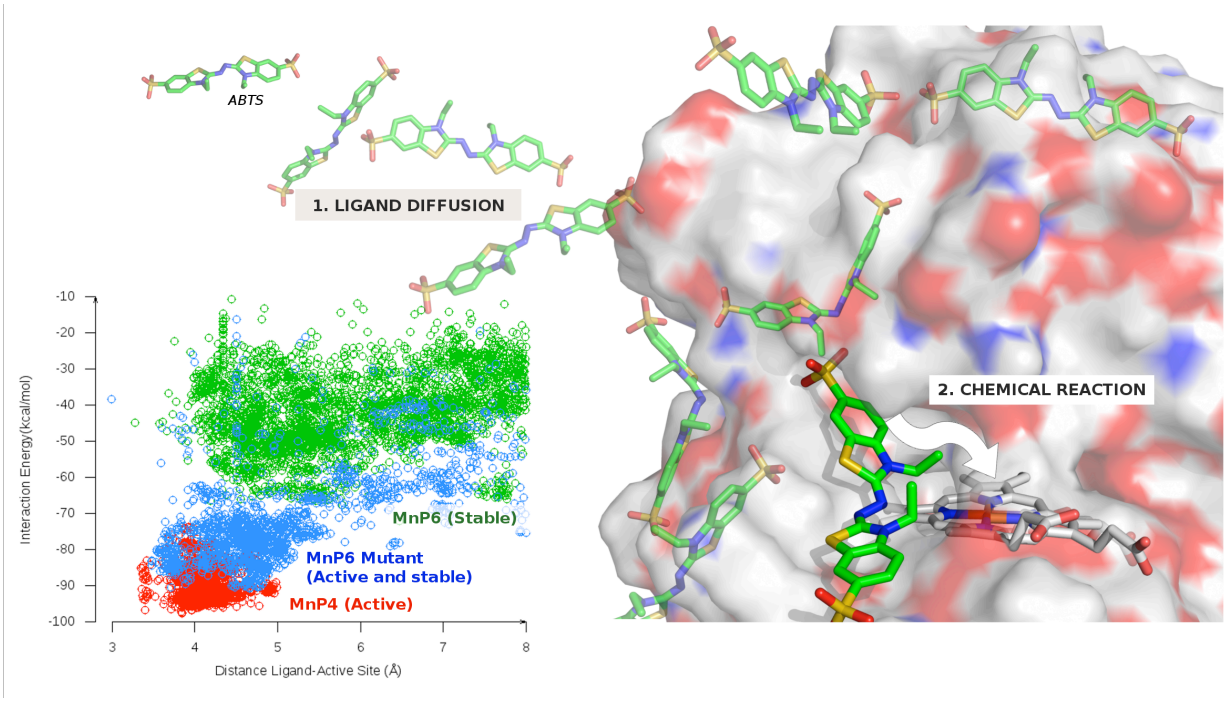

Fig. 2 General scheme for the rational MnP6 engineering (Acebes et al.2016). The study was divided into two steps: 1) ligand diffusion and 2) electronic transfer process. First, we compare the ligand diffusion in the active and inactive enzyme by computing the interaction energy (red and green dots, respectively) and the distance between ligand and receptor. From the active enzyme we extract information about the active site environment that help us to redesign the inactive enzyme by introducing two specific surface mutations (blue dots). Importantly, these mutations involved solvent exposed residues with low conservation and mutability score provided by bioinformatics techniques. The activation was confirmed *in silico* by electronic coupling calculations in the second step. Site-directed experimental mutagenesis validated the success of the new mutant, which combines both stability and activity.

Since long range ET is often the rate limiting step in catalysis, engineering efforts have also centered on mutating residues along the ET pathway. By using the QM/MM e-pathway method [196], Vidal-Limon et al. studied P450BM3's suicide inactivation [197], a common process in heme peroxidases. From the QM-MM calculations they identified key residues in the second heme coordination sphere, aiming at reducing electron delocalization and obtaining a more stable enzyme against H_2O_2. After mass spectrometry assays confirmed the oxidized sites predicted by QM/MM, they generated a variant 260 times more stable against H_2O_2 inactivation.

3. Computer-aided directed evolution, a perspective.

In the previous sections we have seen multiple examples of computationally driven enzyme engineering. While they use, to more or less degree, structure/function knowledge for the modeling, we observe a tendency towards more random massive sequence sampling; we expect to see in the near future full *in silico* directed evolution studies. By full we obviously do not refer to a complete study of the sequence space (all residues and all possible mutations), but to an exhaustive random mutagenesis combination on a large subset of selected mutants. For 100 residues, for example, we have a sequence space of $\sim 10^{130}$, which would take several lifetimes to be evaluated even if using current supercomputers. If we restrict the exploration to single, double and triple mutants, we have now $\sim 10^{9}$ possible variants to model. While this is still a huge number, one can think in a hierarchical scheme where this sequence space can be explored in days. This is particularly true with the current (and future) developments in lower cost multicore servers based on mobile technology (see, for example, the MontBlanc project at https://www.montblanc-project.eu/).

We find a promising example in the work by Wijma et al., where a hierarchical protocol is used to increase thermostabilty [139]. In this line, we expect the development of additional techniques combining quick bioinformatics (or knowledge-based) screening of a large sequence space, with a molecular modeling refinement of selected mutants. This last step could be further hierarchically broken down into a first classical molecular modeling screening followed by selected quantum chemical reevaluations, in a similar manner to the previously described study by Zheng et al. [181]. Even though computational techniques are becoming more precise and easy to implement, a synergic effort between *in silico* predictions and experimental validation will be, in our opinion, the preferred solution. Figure 4 shows a possible workflow combining these ideas. In order to apply such a combined effort, we should keep in mind that molecular modeling will require an accurate 3D structural model, a possible limiting factor.

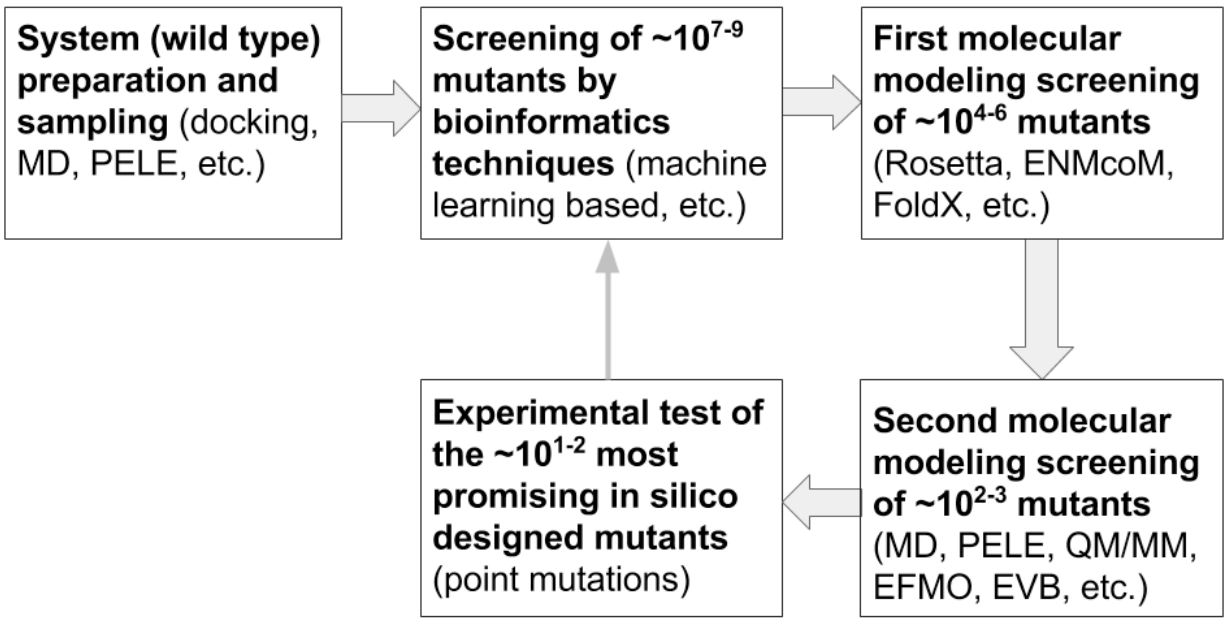

Fig. 3 Proposed computer-aided directed evolution workflow

The workflow must start with a careful preparation of the wild type structure, a fundamental step as it determines the outcome of the computational design. This preparation should include some sampling, aimed at generating conformational diversity and providing useful information for design (such as the protein regions where to look for improved variants). We should emphasize that most molecular modeling predictions are based on relative values (rather than absolute ones), in which case a wild type reference value is needed. In the next step, high-throughput screening of mutations is carried out with quick methods, such as bioinformatics and knowledge-based methods. This step will have to rank the initial sequence space, similar to a high throughput screening in drug design. Taking into account that each bioinformatics score can be accomplished in less than a second, we can aim for several millions mutants in a "doable" time; we are still looking at several days of hundreds/thousands of cores dedication, a feasible effort, however, in near future multicore and accelerated computers (or cloud computing). Bioinformatics screening of millions (billions) mutants will benefit from new sequencing and storage of mutational data in the years to come and, in particular, from its processing using machine learning techniques [198]. At the present stage, such techniques are mostly used to restrict mutagenesis to relevant protein regions [72-75,77,78] or to guide directed evolution "on-the-fly"[199]. A second filtering, for example, could then be performed with fast molecular modeling techniques such as FoldX or Rosetta. These techniques could be applied to (the best ranked) several tens of thousands of compounds. The final goal is to provide a reduced set of candidates, few hundreds/thousands, where we can apply a more accurate molecular modeling refinement. The simulation time required for this last step will highly depend on two factors: i) the exhaustiveness of conformational sampling and, ii) the nature of the property to improve. Conformational sampling aims to determine possible changes in the structure produced by the mutation. Quick assessments, in the order of minutes to hours, can be currently performed by Monte Carlo techniques [154], using MD will require significantly more computational time, limiting the study to only few hundred mutants. Another important aspect is how to re-evaluate the desired property to engineer. Substrate binding energies and thermal stability could be quickly evaluated, for example, with alchemical MM free energy calculations (with respect to the wild type) [128]. Catalytic design, on the other hand, will require expensive QM calculations. Due to their very high cost (hours to days), future work will have to center in designing cheaper methods [200-202] and/or property evaluations. For example, in our current efforts in oxidoreductases's engineering the driving force is approximated with the amount of spin density, calculated after five steps of QM/MM geometry optimization, localized on the substrate (with respect to the wild type) [194].

The proposed workflow includes an iterative computational-experimental approach, where several schemes could be imagined. Currently we find very few studies following such an approach, where we can underline, for example, Privett et al. [203]. When thinking of future implementation, an initial less accurate *in silico* evolution could be tested in the lab and a more accurate second one performed only on those regions that show more promising experimental results. Similarly, more accurate simulations could be performed only in single mutants, followed by experimental site directed mutagesesis and expanded then to a second *in silico* round involving double

mutants, and so on. In this way, synergic mutations can be partially recovered. An alternative strategy to retrieve cooperative mutations, while computing single mutants only, is to rank sequence positions instead of point mutations. Positions are ordered according to their frequency of beneficial mutations, following a fast computational saturated mutagenesis protocol, and the most promising are communicated to the experimental laboratory for (iterative) combinatorial saturated mutagenesis. Contrary to the previous strategy, false positives are not filtered out since a position, instead of a given mutation, is chosen. On the other hand, true positives can be recovered. We are currently employing this strategy to improve oxidoreductases' activity. In our initial test on a high redox potential fungal laccase, initial experimental and theoretical evolutions were run in parallel. In the first DE experimental generation one improved variant was identified. In the *in silico* round, over 40 positions were screened with PELE, using the protein-substrate interaction energy after an induced fit procedure, and the best five identified (Fig. 4 central panel). Interestingly the improved variant found experimentally was within these 5 top positions. Then through combinatorial saturation mutagenesis using NDT degenerated codons three new variant were identified recovering synergetic effect of two of the suggested *in silico* positions. This approach allows quickly guiding experimental mutagenesis: using 100 CPUs ~200 positions can be scored in one day. Although this protocol requires testing more mutants in the lab, it permits to go from the computer to the lab in 24 hours with a focused library of mutants, an appealing feature for industrial purposes where large number of mutants can be assayed.

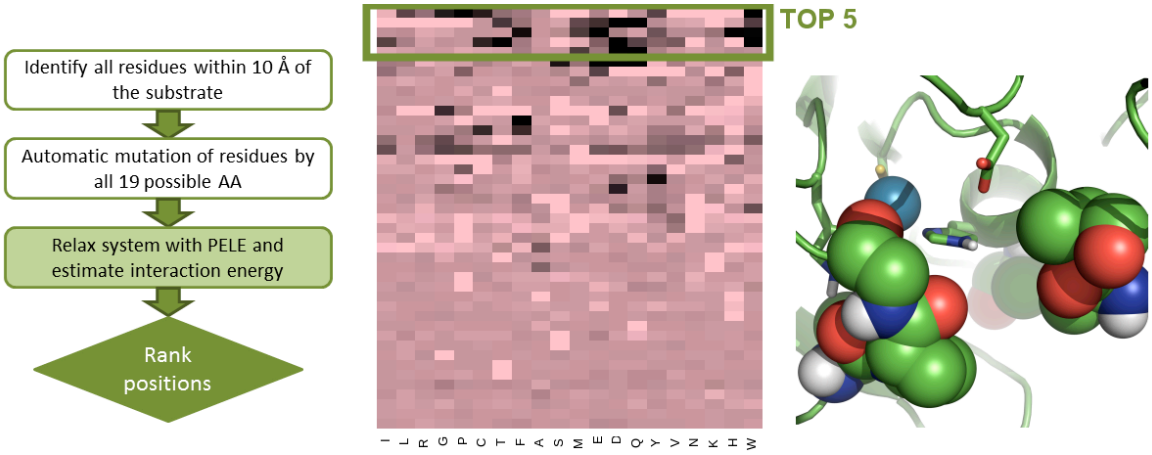

Fig. 4 Left panel: scheme of the used computational protocol. It includes identification of all residues close to the bound substrate, mutation of these amino acids (AA), system relaxation and energy scoring. The image in the center is a heat-map of the tested 43 mutations. The pink color corresponds to an equal or worse value than the wild type, while increasing darker color corresponds to improved classical scoring (protein-substrate interaction energy). From this heat-map the best positions (with the largest number of improved variants relative to the wild type) are identifies, in this example the top 5. On the left panel we find the 5 amino acid positions (in van der Waals representation) suggested to be tested experimentally.

4. Conclusion

Biotechnology needs accurate enzymes evolution techniques, capable of designing new catalysts able to work in environmentally friendly conditions and, importantly, under industrial requirements. In this line, we find great efforts in developing (and improving) site directed mutagenesis and directed evolutions techniques, with computer simulations increasingly being used for this purpose. Different methodologies, from a quick bioinformatics sequence analysis to a robust solution of the Schrödinger equation, seek to aid the experimental efforts. In this book chapter we overviewed recent developments in molecular modeling for three different engineering tasks: protein stability, protein-substrate binding and catalytic rate, with the goal of illustrating how these techniques can influence directed evolution in the near future. We underline two key factors in future implementations: i) hierarchical combination of different computational solutions (with increasing accuracy but also computational cost), and ii) close iterative efforts between *in silico* and *in vitro* approaches.

4. References

1. Schmid A, Dordick JS, Hauer B, Kiener A, Wubbolts M, Witholt B (2001) Industrial biocatalysis today and tomorrow. Nature 409 (6817):258-268.
2. Patel RN (2008) Synthesis of chiral pharmaceutical intermediates by biocatalysis. Coordination Chemistry Reviews 252 (5–7):659-701.
3. Sukumaran J, Hanefeld U (2005) Enantioselective C-C bond synthesis catalysed by enzymes. Chemical Society Reviews 34 (6):530-542.
4. Koenig SH, Brown RD (1972) H(2)CO(3) as Substrate for Carbonic Anhydrase in the Dehydration of HCO(3)(-). Proceedings of the National Academy of Sciences of the United States of America 69 (9):2422-2425.
5. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, Milo R (2011) The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. Biochemistry 50 (21):4402-4410.
6. Milo R, Last RL (2012) Achieving Diversity in the Face of Constraints: Lessons from Metabolism. Science 336 (6089):1663-1667.
7. Currin A, Swainston N, Day PJ, Kell DB (2015) Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. Chemical Society Reviews 44 (5):1172-1239.
8. Gutte B, Däumigen M, Wittschieber E (1979) Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids. Nature 281 (5733):650-655.
9. Russell AJ, Fersht AR (1987) Rational modification of enzyme catalysis by engineering surface charge. Nature 328 (6130):496-500.
10. Hellinga HW, Caradonna JP, Richards FM (1991) Construction of new ligand binding sites in proteins of known structure: II. Grafting of a buried transition metal binding site into Escherichia coli thioredoxin. Journal of Molecular Biology 222 (3):787-803.
11. Jemli S, Ayadi-Zouari D, Hlima HB, Bejar S (2016) Biocatalysts: application and engineering for industrial purposes. Critical Reviews in Biotechnology 36 (2):246-258.
12. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D (2005) Progress in Modeling of Protein Structures and Interactions. Science 310 (5748):638-642.

13. Steiner K, Schwab H (2012) Recent advances in rational approaches for enzyme engineering. Computational and Structural Biotechnology Journal 2:e201209010.

14. Richardson JS, Richardson DC (1989) The de novo design of protein structures. Trends in Biochemical Sciences 14 (7):304-309.

15. Ponder JW, Richards FM (1987) Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of Molecular Biology 193 (4):775-791.

16. Bolon DN, Marcus JS, Ross SA, Mayo SL (2003) Prudent Modeling of Core Polar Residues in Computational Protein Design. Journal of Molecular Biology 329 (3):611-622.

17. Dahiyat BI, Mayo SL (1997) Probing the role of packing specificity in protein design. Proceedings of the National Academy of Sciences of the United States of America 94 (19):10172-10177.

18. Desjarlais JR, Handel TM (1999) Side-chain and backbone flexibility in protein core design1. Journal of Molecular Biology 290 (1):305-318.

19. Scrutton NS, Berry A, Perham RN (1990) Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. Nature 343 (6253):38-43.

20. Carter P, Nilsson B, Burnier JP, Burdick D, Wells JA (1989) Engineering subtilisin BPN′ for site-specific proteolysis. Proteins: Structure, Function, and Bioinformatics 6 (3):240-248.

21. Wells JA, Powers DB, Bott RR, Graycar TP, Estell DA (1987) Designing substrate specificity by protein engineering of electrostatic interactions. Proceedings of the National Academy of Sciences of the United States of America 84 (5):1219-1223.

22. Cedrone F, Ménez A, Quéméneur E (2000) Tailoring new enzyme functions by rational redesign. Current Opinion in Structural Biology 10 (4):405-410.

23. Looger LL, Dwyer MA, Smith JJ, Hellinga HW (2003) Computational design of receptor and sensor proteins with novel functions. Nature 423 (6936):185-190.

24. Craik C, Largman C, Fletcher T, Roczniak S, Barr P, Fletterick R, Rutter W (1985) Redesigning trypsin: alteration of substrate specificity. Science 228 (4697):291-297.

25. Bastianelli G, Bouillon A, Nguyen C, Crublet E, Pêtres S, Gorgette O, Le-Nguyen D, Barale J-C, Nilges M (2011) Computational Reverse-Engineering of a Spider-Venom Derived Peptide Active Against <italic>Plasmodium falciparum</italic> SUB1. PLoS ONE 6 (7):e21812.

26. Oelschlaeger P, Mayo SL (2005) Hydroxyl Groups in the ββ Sandwich of Metallo-β-lactamases Favor Enzyme Activity: A Computational Protein Design Study. Journal of Molecular Biology 350 (3):395-401.

27. Guerois R, Nielsen JE, Serrano L (2002) Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. Journal of Molecular Biology 320 (2):369-387.

28. Yu H, Huang H (2014) Engineering proteins for thermostability through rigidifying flexible sites. Biotechnology Advances 32 (2):308-315.

29. Kuhlman B, Baker D (2004) Exploring folding free energy landscapes using computational protein design. Current Opinion in Structural Biology 14 (1):89-95.

30. Kortemme T, Baker D (2004) Computational design of protein–protein interactions. Current Opinion in Chemical Biology 8 (1):91-97.

31. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D (2004) Computational redesign of protein-protein interaction specificity. Nat Struct Mol Biol 11 (4):371-379.

32. Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, Schwab MS, Serrano L, Gonzalez C (2002) Computer-aided design of a PDZ domain to recognize new target sequences. Nat Struct Mol Biol 9 (8):621-627.

33. Shifman JM, Mayo SL (2002) Modulating Calmodulin Binding Specificity through Computational Protein Design. Journal of Molecular Biology 323 (3):417-423.

34. Lippow SM, Tidor B (2007) Progress in Computational Protein Design. Current opinion in biotechnology 18

(4):305-311.

35. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, Stoddard BL, Baker D (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 441 (7093):656-659.

36. Chevalier BS, Kortemme T, Chadsey MS, Baker D, Monnat RJ, Jr., Stoddard BL (2002) Design, Activity, and Structure of a Highly Specific Artificial Endonuclease. Molecular Cell 10 (4):895-905.

37. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, Therien MJ, DeGrado WF (2005) Computational De Novo Design and Characterization of a Four-Helix Bundle Protein that Selectively Binds a Nonbiological Cofactor. Journal of the American Chemical Society 127 (5):1346-1347.

38. Yang W, Wilkins AL, Ye Y, Liu Z-r, Li S-y, Urbauer JL, Hellinga HW, Kearney A, van der Merwe PA, Yang JJ (2005) Design of a Calcium-Binding Protein with Desired Structure in a Cell Adhesion Molecule. Journal of the American Chemical Society 127 (7):2085-2093.

39. Palmer AE, Giacomello M, Kortemme T, Hires SA, Lev-Ram V, Baker D, Tsien RY (2006) Ca2+ Indicators Based on Computationally Redesigned Calmodulin-Peptide Pairs. Chemistry & Biology 13 (5):521-530.

40. Lassila JK, Keeffe JR, Oelschlaeger P, Mayo SL (2005) Computationally designed variants of Escherichia coli chorismate mutase show altered catalytic activity. Protein Engineering Design and Selection 18 (4):161-163.

41. Bornscheuer UT, Pohl M (2001) Improved biocatalysts by directed evolution and rational protein design. Current Opinion in Chemical Biology 5 (2):137-143.

42. Faiella M, Andreozzi C, de Rosales RTM, Pavone V, Maglio O, Nastri F, DeGrado WF, Lombardi A (2009) An artificial di-iron oxo-protein with phenol oxidase activity. Nature chemical biology 5 (12):882-884.

43. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, Baker D (2013) Computational Design of Ligand Binding Proteins with High Affinity and Selectivity. Nature 501 (7466):212-216.

44. Kaplan J, DeGrado WF (2004) De novo design of catalytic proteins. Proceedings of the National Academy of Sciences of the United States of America 101 (32):11566-11570.

45. Dahiyat BI, Mayo SL (1997) De novo protein design: fully automated sequence selection. Science 278 (5335):82-87.

46. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De Novo Computational Design of Retro-Aldol Enzymes. Science (New York, NY) 319 (5868):1387-1391.

47. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D (2008) Kemp elimination catalysts by computational enzyme design. Nature 453 (7192):190-195.

48. Moroz YS, Dunston TT, Makhlynets OV, Moroz OV, Wu Y, Yoon JH, Olsen AB, McLaughlin JM, Mack KL, Gosavi PM, van Nuland NAJ, Korendovych IV (2015) New Tricks for Old Proteins: Single Mutations in a Nonenzymatic Protein Give Rise to Various Enzymatic Activities. Journal of the American Chemical Society 137 (47):14905-14911.

49. Zanghellini A (2014) de novo computational enzyme design. Current opinion in biotechnology 29:132-138.

50. Petrounia IP, Arnold FH (2000) Designed evolution of enzymatic properties. Current opinion in biotechnology 11 (4):325-330.

51. Arnold FH (2001) Combinatorial and computational challenges for biocatalyst design. Nature 409 (6817):253-257.

52. Minshull J, Willem Stemmer PC (1999) Protein evolution by molecular breeding. Current Opinion in Chemical Biology 3 (3):284-290.

53. Packer MS, Liu DR (2015) Methods for the directed evolution of proteins. Nat Rev Genet 16 (7):379-394.

54. Jaeger K-E, Eggert T (2004) Enantioselective biocatalysis optimized by directed evolution. Current opinion in

biotechnology 15 (4):305-313.

55. Jestin J-L, Kaminski PA (2004) Directed enzyme evolution and selections for catalysis based on product formation. Journal of Biotechnology 113 (1–3):85-103.

56. Tao H, Cornish VW (2002) Milestones in directed enzyme evolution. Current Opinion in Chemical Biology 6 (6):858-864.

57. Williams GJ, Nelson AS, Berry A (2004) Directed evolution of enzymes for biocatalysis and the life sciences. Cellular and Molecular Life Sciences CMLS 61 (24):3034-3046.

58. Dalby PA (2003) Optimising enzyme function by directed evolution. Current Opinion in Structural Biology 13 (4):500-505.

59. Bershtein S, Tawfik DS (2008) Advances in laboratory evolution of enzymes. Current Opinion in Chemical Biology 12 (2):151-158.

60. Park S, Morley KL, Horsman GP, Holmquist M, Hult K, Kazlauskas RJ (2005) Focusing Mutations into the P. fluorescens Esterase Binding Site Increases Enantioselectivity More Effectively than Distant Mutations. Chemistry & Biology 12 (1):45-54.

61. Strausberg SL, Ruan B, Fisher KE, Alexander PA, Bryan PN (2005) Directed Coevolution of Stability and Catalytic Activity in Calcium-free Subtilisin†. Biochemistry 44 (9):3272-3279.

62. Chockalingam K, Chen Z, Katzenellenbogen JA, Zhao H (2005) Directed evolution of specific receptor–ligand pairs for use in the creation of gene switches. Proceedings of the National Academy of Sciences of the United States of America 102 (16):5691-5696.

63. Chica RA, Doucet N, Pelletier JN (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Current opinion in biotechnology 16 (4):378-384.

64. Hill CM, Li W-S, Thoden JB, Holden HM, Raushel FM (2003) Enhanced Degradation of Chemical Warfare Agents through Molecular Engineering of the Phosphotriesterase Active Site. Journal of the American Chemical Society 125 (30):8990-8991.

65. Reetz MT, Carballeira JD (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. Nat Protocols 2 (4):891-903.

66. Lutz S, Patrick WM (2004) Novel methods for directed evolution of enzymes: quality, not quantity. Current opinion in biotechnology 15 (4):291-297.

67. Lutz S (2010) Beyond directed evolution - semi-rational protein engineering and design. Current opinion in biotechnology 21 (6):734-743.

68. Lippow SM, Moon TS, Basu S, Yoon S-H, Li X, Chapman BA, Robison K, Lipovšek D, Prather KLJ Engineering Enzyme Specificity Using Computational Design of a Defined-Sequence Library. Chemistry & Biology 17 (12):1306-1315.

69. Sebestova E, Bendl J, Brezovsky J, Damborský J (2014) Computational Tools for Designing Smart Libraries. Methods Molecular Biology 1179:291-314.

70. Voigt CA, Mayo SL, Arnold FH, Wang Z-G (2001) Computationally focusing the directed evolution of proteins. Journal of Cellular Biochemistry 84 (S37):58-63.

71. Zaugg J, Gumulya Y, Gillam EM, Boden M (2014) Computational tools for directed evolution: a comparison of prospective and retrospective strategies. Methods Mol Biol 1179:315-333.

72. Damborsky J, Brezovsky J (2009) Computational tools for designing and engineering biocatalysts. Current Opinion in Chemical Biology 13 (1):26-34.

73. Pei J (2008) Multiple protein sequence alignment. Current Opinion in Structural Biology 18 (3):382-386.

74. Pavelka A, Chovancova E, Damborsky J (2009) HotSpot Wizard: a web server for identification of hot spots in protein engineering. Nucleic Acids Research 37 (suppl 2):W376-W383.

75. Kuipers RK, Joosten H-J, van Berkel WJH, Leferink NGH, Rooijen E, Ittmann E, van Zimmeren F, Jochens H,

Bornscheuer U, Vriend G, Martins dos Santos VAP, Schaap PJ (2010) 3DM: Systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins: Structure, Function, and Bioinformatics 78 (9):2101-2113.

76. Jochens H, Bornscheuer UT (2010) Natural Diversity to Guide Focused Directed Evolution. ChemBioChem 11 (13):1861-1866.

77. Goldsmith M, Tawfik DS (2012) Directed enzyme evolution: beyond the low-hanging fruit. Current Opinion in Structural Biology 22 (4):406-412.

78. Barak Y, Nov Y, Ackerley DF, Matin A (2007) Enzyme improvement in the absence of structural knowledge: a novel statistical approach. ISME J 2 (2):171-179.

79. Rosenberg M, Goldblum A (2006) Computational Protein Design: A Novel Path to Future Protein Drugs. Current Pharmaceutical Design 12 (31):3973-3997.

80. Poole AM, Ranganathan R (2006) Knowledge-based potentials in protein design. Current Opinion in Structural Biology 16 (4):508-513.

81. Koder RL, Dutton PL (2006) Intelligent design: the de novo engineering of proteins with specified functions. Dalton Transactions (25):3045-3051.

82. Butterfoss GL, Kuhlman B (2006) COMPUTER-BASED DESIGN OF NOVEL PROTEIN STRUCTURES. Annual Review of Biophysics and Biomolecular Structure 35 (1):49-65.

83. Ambroggio XI, Kuhlman B (2006) Design of protein conformational switches. Current Opinion in Structural Biology 16 (4):525-530.

84. Vizcarra CL, Mayo SL (2005) Electrostatics in computational protein design. Current Opinion in Chemical Biology 9 (6):622-626.

85. Morin A, Meiler J, Mizoue LS Computational design of protein–ligand interfaces: potential in therapeutic development. Trends in Biotechnology 29 (4):159-166.

86. Malisi C, Schumann M, Toussaint NC, Kageyama J, Kohlbacher O, Höcker B (2012) Binding Pocket Optimization by Computational Protein Design. PLoS ONE 7 (12):e52505.

87. Saven JG (2011) Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins. Current Opinion in Chemical Biology 15 (3):452-457.

88. Ollikainen N, Smith CA, Fraser JS, Kortemme T (2013) Methods in Enzymology: "Flexible backbone sampling methods to model and design protein alternative conformations". Methods in enzymology 523:61-85.

89. Park S, Yang X, Saven JG (2004) Advances in computational protein design. Current Opinion in Structural Biology 14 (4):487-494.

90. Samish I, MacDermaid CM, Perez-Aguilar JM, Saven JG (2011) Theoretical and Computational Protein Design. Annual Review of Physical Chemistry 62 (1):129-149.

91. Smith RD, Damm-Ganamet KL, Dunbar JB, Ahmed A, Chinnaswamy K, Delproposto JE, Kubish GM, Tinberg CE, Khare SD, Dou J, Doyle L, Stuckey JA, Baker D, Carlson HA (2015) CSAR Benchmark Exercise 2013: Evaluation of Results from a Combined Computational Protein Design, Docking, and Scoring/Ranking Challenge. Journal of Chemical Information and Modeling ASAP.

92. Wijma HJ, Janssen DB (2013) Computational design gains momentum in enzyme catalysis engineering. FEBS Journal 280 (13):2948-2960.

93. Boas FE, Harbury PB (2007) Potential energy functions for protein design. Current Opinion in Structural Biology 17 (2):199-204.

94. Boas FE, Harbury PB (2008) Design of Protein-Ligand Binding Based on the Molecular-Mechanics Energy Model. Journal of Molecular Biology 380 (2):415-424.

95. Sirin S, Pearlman DA, Sherman W (2014) Physics-based enzyme design: Predicting binding affinity and

catalytic activity. Proteins: Structure, Function, and Bioinformatics 82 (12):3397-3409.

96. Wickstrom L, Gallicchio E, Levy RM (2012) The Linear Interaction Energy Method for the Prediction of Protein Stability Changes Upon Mutation. Proteins 80 (1):111-125.

97. Mendes J, Guerois R, Serrano L (2002) Energy estimation in protein design. Current Opinion in Structural Biology 12 (4):441-446.

98. Schneider M, Fu X, Keating AE (2009) X-ray vs. NMR structures as templates for computational protein design. Proteins 77 (1):97-110.

99. Adamczyk AJ, Cao J, Kamerlin SCL, Warshel A (2011) Catalysis by dihydrofolate reductase and other enzymes arises from electrostatic preorganization, not conformational motions. Proceedings of the National Academy of Sciences of the United States of America 108 (34):14115-14120.

100. Gagné D, French Rachel L, Narayanan C, Simonović M, Agarwal Pratul K, Doucet N Perturbation of the Conformational Dynamics of an Active-Site Loop Alters Enzyme Activity. Structure 23 (12):2256-2266.

101. Bhabha G, Lee J, Ekiert DC, Gam J, Wilson IA, Dyson HJ, Benkovic SJ, Wright PE (2011) A Dynamic Knockout Reveals That Conformational Fluctuations Influence the Chemical Step of Enzyme Catalysis. Science 332 (6026):234-238.

102. Allen BD, Nisthal A, Mayo SL (2010) Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. Proceedings of the National Academy of Sciences 107 (46):19838-19843.

103. Fu X, Apgar JR, Keating AE (2007) Modeling Backbone Flexibility to Achieve Sequence Diversity: The Design of Novel α-Helical Ligands for Bcl-xL. Journal of Molecular Biology 371 (4):1099-1117.

104. Smith CA, Kortemme T (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. Journal of Molecular Biology 380 (4):742-756.

105. Lassila JK (2010) Conformational diversity and computational enzyme design. Current Opinion in Chemical Biology 14 (5):676-682.

106. Mandell DJ, Kortemme T (2009) Backbone flexibility in computational protein design. Current opinion in biotechnology 20 (4):420-428.

107. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, De Vries AH (2007) The MARTINI force field: coarse grained model for biomolecular simulations. The Journal of Physical Chemistry B 111 (27):7812-7824.

108. Bowen JP, Allinger NL (2007) Molecular mechanics: The art and science of parameterization. Reviews in Computational Chemistry 2:81-97.

109. Doruker P, Atilgan AR, Bahar I (2000) Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α-amylase inhibitor. Proteins: Structure, Function, and Bioinformatics 40 (3):512-524.

110. Berendsen H (1988) Dynamic simulation as an essential tool in molecular modeling. Journal of computer-aided molecular design 2 (3):217-221.

111. Grossman J, Towles B, Greskamp B, Shaw DE Filtering, reductions and synchronization in the anton 2 network. In: Parallel and Distributed Processing Symposium (IPDPS), 2015 IEEE International, 2015. IEEE, pp 860-870.

112. Rathore N, de Pablo JJ (2002) Monte Carlo simulation of proteins through a random walk in energy space. The Journal of chemical physics 116 (16):7225-7230.

113. Borrelli KW, Vitalis A, Alcantara R, Guallar V (2005) PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. Journal of Chemical Theory and Computation 1 (6):1304-1311.

114. Cabeza de Vaca I, Lucas MFt, Guallar V (2015) New Monte Carlo Based Technique To Study DNA–Ligand Interactions. Journal of chemical theory and computation 11 (12):5598-5605.

115. Borrelli KW, Cossins B, Guallar V (2010) Exploring hierarchical refinement techniques for induced fit docking

with protein and ligand flexibility. Journal of computational chemistry 31 (6):1224-1235.

116. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins: Structure, Function, and Bioinformatics 47 (4):409-443.

117. Gao J, Truhlar DG (2002) Quantum mechanical methods for enzyme kinetics. Annual Review of Physical Chemistry 53 (1):467-505.

118. Korkegian A (2005) Computational Thermostabilization of an Enzyme. Science 308 (5723):857-860.

119. Kellogg EH, Leaver-Fay A, Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 79 (3):830-838.

120. Dunbrack RL, Jr. (2002) Rotamer libraries in the 21st century. Curr Opin Struct Biol 12 (4):431-440.

121. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Sciences 97 (19):10383-10388.

122. Potapov V, Cohen M, Schreiber G (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Engineering Design and Selection 22 (9):553-560.

123. Estrada J, Echenique P, Sancho J (2015) Predicting stabilizing mutations in proteins using Poisson-Boltzmann based models: study of unfolded state ensemble models and development of a successful binary classifier based on residue interaction energies. Phys Chem Chem Phys 17 (46):31044-31054.

124. Karplus M, Ichiye T, Pettitt BM (1987) Configurational entropy of native proteins. Biophys J 52 (6):1083-1085.

125. Chong S-H, Ham S (2015) Dissecting Protein Configurational Entropy into Conformational and Vibrational Contributions. J Phys Chem B 119 (39):12623-12631.

126. Frappier V, Chartier M, Najmanovich RJ (2015) ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. Nucleic Acids Res 43 (W1):W395-400.

127. Frappier V, Najmanovich RJ (2014) A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. PLoS Comput Biol 10 (4):e1003569.

128. Seeliger D, Daniel S, de Groot BL (2010) Protein Thermostability Calculations Using Alchemical Free Energy Simulations. Biophys J 98 (10):2309-2316.

129. Huang X, Gao D, Zhan C-G (2011) Computational design of a thermostable mutant of cocaine esterase via molecular dynamics simulations. Org Biomol Chem 9 (11):4138-4143.

130. Joo JC, Pack SP, Kim YH, Yoo YJ (2011) Thermostabilization of Bacillus circulans xylanase: computational optimization of unstable residues based on thermal fluctuation analysis. J Biotechnol 151 (1):56-65.

131. Lee C-W, Wang H-J, Hwang J-K, Tseng C-P (2014) Protein thermal stability enhancement by designing salt bridges: a combined computational and experimental study. PLoS ONE 9 (11):e112751.

132. Pikkemaat MG, Linssen ABM, Berendsen HJC, Janssen DB (2002) Molecular dynamics simulations as a tool for improving protein stability. Protein Eng 15 (3):185-192.

133. Gribenko AV, Patel MM, Liu J, McCallum SA, Wang C, Makhatadze GI (2009) Rational stabilization of enzymes by computational redesign of surface charge-charge interactions. Proceedings of the National Academy of Sciences 106 (8):2601-2606.

134. Spector S, Wang M, Carp SA, Robblee J, Hendsch ZS, Fairman R, Tidor B, Raleigh DP (2000) Rational modification of protein stability by the mutation of charged surface residues. Biochemistry 39 (5):872-879.

135. Schweiker KL, Arash Z-A, Davidson AR, Makhatadze GI (2007) Computational design of the Fyn SH3 domain with increased stability through optimization of surface charge-charge interactions. Protein Sci 16 (12):2694-2702.

136. Borgo B, Havranek JJ (2012) Automated selection of stabilizing mutations in designed and natural proteins. Proc Natl Acad Sci U S A 109 (5):1494-1499.

137. Hendsch ZS, Thorlakur J, Sauer RT, Bruce T (1996) Protein Stabilization by Removal of Unsatisfied Polar Groups: Computational Approaches and Experimental Tests †. Biochemistry 35 (24):7621-7625.

138. Koudelakova T, Chaloupkova R, Brezovsky J, Prokop Z, Sebestova E, Hesseler M, Khabiri M, Plevaka M, Kulik D, Kuta Smatanova I, Rezacova P, Ettrich R, Bornscheuer UT, Damborsky J (2013) Engineering enzyme stability and resistance to an organic cosolvent by modification of residues in the access tunnel. Angew Chem Int Ed Engl 52 (7):1959-1963.

139. Wijma HJ, Floor RJ, Jekel PA, Baker D, Marrink SJ, Janssen DB (2014) Computationally designed libraries for rapid enzyme stabilization. Protein Eng Des Sel 27 (2):49-58.

140. Wijma HJ, Floor RJ, Janssen DB (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. Curr Opin Struct Biol 23 (4):588-594.

141. Schreier B, Stumpp C, Wiesner S, Hocker B (2009) Computational design of ligand binding is not a solved problem. Proceedings of the National Academy of Sciences 106 (44):18491-18496.

142. Morin A, Meiler J, Mizoue LS (2011) Computational design of protein-ligand interfaces: potential in therapeutic development. Trends Biotechnol 29 (4):159-166.

143. Allison B, Combs S, DeLuca S, Lemmon G, Mizoue L, Meiler J (2014) Computational design of protein-small molecule interfaces. J Struct Biol 185 (2):193-202.

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

145. Keedy DA, Georgiev I, Triplett EB, Donald BR, Richardson DC, Richardson JS (2012) The role of local backrub motions in evolved and designed mutations. PLoS Comput Biol 8 (8):e1002629.

146. Davis IW, Bryan Arendall W, Richardson DC, Richardson JS (2006) The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. Structure 14 (2):265-274.

147. Chen C-Y, Georgiev I, Anderson AC, Donald BR (2009) Computational structure-based redesign of enzyme activity. Proc Natl Acad Sci U S A 106 (10):3764-3769.

148. Frey KM, Georgiev I, Donald BR, Anderson AC (2010) Predicting resistance mutations using protein design algorithms. Proc Natl Acad Sci U S A 107 (31):13707-13712.

149. Zhou Y, Xu W, Donald BR, Zeng J (2014) An efficient parallel algorithm for accelerating computational protein design. Bioinformatics 30 (12):i255-i263.

150. Hallen MA, Keedy DA, Donald BR (2013) Dead-end elimination with perturbations (DEEPer): a provable protein design algorithm with continuous sidechain and backbone flexibility. Proteins 81 (1):18-39.

151. Lilien RH, Stevens BW, Anderson AC, Donald BR (2005) A Novel Ensemble-Based Scoring and Search Algorithm for Protein Redesign and Its Application to Modify the Substrate Specificity of the Gramicidin Synthetase A Phenylalanine Adenylation Enzyme. J Comput Biol 12 (6):740-761.

152. Leach AR (2001) Molecular Modelling: Principles and Applications. Pearson Education,

153. Shields GC, Seybold PG (2013) Computational Approaches for the Prediction of pKa Values. CRC Press,

154. Pardo I, Santiago G, Gentili P, Lucas F, Monza E, Medrano F, Galli C, Martínez A, Guallar V, Camarero S (2016) Re-designing the substrate binding pocket of laccase for enhanced oxidation of sinapic acid. Catalysis Science & Technology ASAP.

155. Young T, Abel R, Kim B, Berne BJ, Friesner RA (2007) Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. Proceedings of the National Academy of Sciences 104 (3):808-813.

156. Kiss G, Çelebi-Ölçüm N, Moretti R, Baker D, Houk KN (2013) Computational Enzyme Design. Angewandte Chemie International Edition 52 (22):5700-5725.

157. Doerr S, De Fabritiis G (2014) On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput

Molecular Simulations. J Chem Theory Comput 10 (5):2064-2069.

158. Wijma HJ, Floor RJ, Bjelic S, Marrink SJ, Baker D, Janssen DB (2015) Enantioselective enzymes by computational design and in silico screening. Angew Chem Int Ed Engl 54 (12):3726-3730.

159. Jiménez-Osés G, Osuna S, Gao X, Sawaya MR, Gilson L, Collier SJ, Huisman GW, Yeates TO, Tang Y, Houk KN (2014) The Role of Distant Mutations and Allosteric Regulation on LovD Active Site Dynamics. Nature chemical biology 10 (6):431-436.

160. Osuna S, Jiménez-Osés G, Noey EL, Houk KN (2015) Molecular dynamics explorations of active site structure in designed and evolved enzymes. Acc Chem Res 48 (4):1080-1089.

161. DuBay KH, Bowman GR, Geissler PL (2015) Fluctuations within folded proteins: implications for thermodynamic and allosteric regulation. Acc Chem Res 48 (4):1098-1105.

162. Sethi A, Eargle J, Black AA, Luthey-Schulten Z (2009) Dynamical networks in tRNA:protein complexes. Proc Natl Acad Sci U S A 106 (16):6620-6625.

163. Madadkar-Sobhani A, Guallar V (2013) PELE web server: atomistic study of biomolecular systems at your fingertips. Nucleic Acids Res 41 (Web Server issue):W322-328.

164. Lucas MF, Guallar V (2012) An Atomistic View on Human Hemoglobin Carbon Monoxide Migration Processes. Biophys J 102 (4):887-896.

165. Takahashi R, Gil VA, Guallar V (2014) Monte Carlo Free Ligand Diffusion with Markov State Model Analysis and Absolute Binding Free Energy Calculations. J Chem Theory Comput 10 (1):282-288.

166. Hosseini A, Brouk M, Lucas MF, Glaser F, Fishman A, Guallar V (2015) Atomic picture of ligand migration in toluene 4-monooxygenase. J Phys Chem B 119 (3):671-678.

167. Lüdemann SK, Lounnas V, Wade RC (2000) How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. Journal of Molecular Biology 303 (5):797-811.

168. Grubmüller H, Heymann B, Tavan P (1996) Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. Science 271 (5251):997-999.

169. Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics 10 (1):168.

170. Chovancova E, Eva C, Antonin P, Petr B, Ondrej S, Jan B, Barbora K, Artur G, Vilem S, Martin K, Petr M, Lada B, Jiri S, Jiri D (2012) CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. PLoS Comput Biol 8 (10):e1002708.

171. Senn HM, Walter T (2009) QM/MM Methods for Biomolecular Systems. Angew Chem Int Ed 48 (7):1198-1229.

172. Chaskar P, Prasad C, Vincent Z, Röhrig UF (2014) Toward On-The-Fly Quantum Mechanical/Molecular Mechanical (QM/MM) Docking: Development and Benchmark of a Scoring Function. J Chem Inf Model 54 (11):3137-3152.

173. Cho AE, Victor G, Berne BJ, Richard F (2005) Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. J Comput Chem 26 (9):915-931.

174. Fedorov DG, Nagata T, Kitaura K (2012) Exploring chemistry with the fragment molecular orbital method. Phys Chem Chem Phys 14 (21):7562-7577.

175. Jensen JH, Willemoës M, Winther JR, De Vico L (2014) In silico prediction of mutant HIV-1 proteases cleaving a target sequence. PLoS ONE 9 (5):e95833.

176. Grisewood MJ, Gifford NP, Pantazes RJ, Li Y, Cirino PC, Janik MJ, Maranas CD (2013) OptZyme: computational enzyme redesign using transition state analogues. PLoS ONE 8 (10):e75358.

177. Atkins PW (1998) Physical Chemistry. W H Freeman & Company,

178. Khersonsky O, Rothlisberge D, Wollacott AM, Dym O, Baker D, Tawfik DS (2011) Optimization of the in

silico designed Kemp eliminase KE70 by computational design and directed evolution. Journal of Molecular Biology 407 (3):391-412.

179. Genheden S, Samuel G, Ulf R (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov 10 (5):449-461.

180. van der Kamp MW, Mulholland AJ (2013) Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. Biochemistry 52 (16):2708-2728.

181. Zheng F, Fang Z, Wenchao Y, Mei-Chuan K, Junjun L, Hoon C, Daquan G, Min T, Hsin-Hsiung T, Woods JH, Chang-Guo Z (2008) Most Efficient Cocaine Hydrolase Designed by Virtual Screening of Transition States. J Am Chem Soc 130 (36):12148-12155.

182. Kamerlin SCL, Arieh W (2011) The empirical valence bond model: theory and applications. Wiley Interdiscip Rev Comput Mol Sci 1 (1):30-45.

183. Frushicheva MP, Cao J, Chu ZT, Warshel A (2010) Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase. Proc Natl Acad Sci U S A 107 (39):16869-16874.

184. Frushicheva MP, Cao J, Warshel A (2011) Challenges and advances in validating enzyme design proposals: the case of kemp eliminase catalysis. Biochemistry 50 (18):3849-3858.

185. Hediger MR, De Vico L, Svendsen A, Besenmatter W, Jensen JH (2012) A computational methodology to screen activities of enzyme variants. PLoS ONE 7 (12):e49849.

186. Hediger MR, Casper S, De Vico L, Jensen JH (2013) A computational method for the systematic screening of reaction barriers in enzymes: searching for Bacillus circulans xylanase mutants with greater activity towards a synthetic substrate. PeerJ 1:e111.

187. Hediger MR, De Vico L, Rannes JB, Christian J, Werner B, Allan S, Jensen JH (2013) In silico screening of 393 mutants facilitates enzyme engineering of amidase activity in CalB. PeerJ 1:e145.

188. Ito M, Mika I, Tore B (2014) Novel Approach for Identifying Key Residues in Enzymatic Reactions: Proton Abstraction in Ketosteroid Isomerase. J Phys Chem B 118 (46):13050-13058.

189. Steinmann C, Fedorov DG, Jensen JH (2012) The effective fragment molecular orbital method for fragments connected by covalent bonds. PLoS ONE 7 (7):e41117.

190. Steinmann C, Casper S, Fedorov DG, Jensen JH (2013) Mapping Enzymatic Catalysis Using the Effective Fragment Molecular Orbital Method: Towards all ab initio Biochemistry. PLoS ONE 8 (4):e60602.

191. Marcus RA (1993) Electron transfer reactions in chemistry. Theory and experiment. Rev Mod Phys 65 (3):599-610.

192. Blumberger J, Jochen B (2008) Free energies for biological electron transfer from QM/MM calculation: method, application and critical assessment. Phys Chem Chem Phys 10 (37):5651.

193. Wallrapp FH, Voityuk AA, Guallar V (2013) In-Silico Assessment of Protein-Protein Electron Transfer. A Case Study: Cytochrome c Peroxidase–Cytochrome c. PLoS Comput Biol 9 (3):e1002990.

194. Monza E, Lucas MF, Camarero S, Alejaldre LC, Martínez AT, Guallar V (2015) Insights into Laccase Engineering from Molecular Simulations: Toward a Binding-Focused Strategy. J Phys Chem Lett 6 (8):1447-1453.

195. Acebes S, Fernandez-Fueyo E, Monza E, Lucas M, Almendral D, Ruiz-Dueñas FJ, Lund H, Martinez AT, Guallar V (2016) Rational Enzyme Engineering Through Biophysical and Biochemical Modeling. ACS Catal ASAP.

196. Guallar V, Wallrapp F (2008) Mapping protein electron transfer pathways with QM/MM methods. Journal of The Royal Society Interface 5 (0):S233.

197. Vidal-Limón A, Águila S, Ayala M, Batista CV, Vazquez-Duhalt R (2013) Peroxidase activity stabilization of cytochrome P450 BM3 by rational analysis of intramolecular electron transfer. Journal of inorganic biochemistry 122:18-26.

198. Fox RJ, Huisman GW (2008) Enzyme optimization: moving from blind evolution to statistical exploration of sequence–function space. Trends in Biotechnology 26 (3):132-138.

199. Feng X, Sanchis J, Reetz MT, Rabitz H (2012) Enhancing the Efficiency of Directed Evolution in Focused Enzyme Libraries by the Adaptive Substituent Reordering Algorithm. Chemistry – A European Journal 18 (18):5646-5654.

200. Cui Q, Elstner M (2014) Density functional tight binding: values of semi-empirical methods in an ab initio era. Phys Chem Chem Phys 16 (28):14368-14377.

201. Christensen AS, Elstner M, Cui Q (2015) Improving intermolecular interactions in DFTB3 using extended polarization from chemical-potential equalization. The Journal of chemical physics 143 (8):084123.

202. Yilmazer ND, Korth M (2015) Enhanced semiempirical QM methods for biomolecular interactions. Computational and Structural Biotechnology Journal 13:169-175.

203. Privett HK, Kiss G, Lee TM, Blomberg R, Chica RA, Thomas LM, Hilvert D, Houk KN, Mayo SL (2012) Iterative approach to computational enzyme design. Proceedings of the National Academy of Sciences 109 (10):3790-3795