# Chapter 13

# Computational Enzyme Design at Zymvol

**Emanuele Monza, Victor Gil, and Maria Fatima Lucas**

## Abstract

Directed evolution is the most recognized methodology for enzyme engineering. The main drawback resides in its random nature and in the limited sequence exploration; both require screening of thousands (if not millions) of variants to achieve a target function. Computer-driven approaches can limit laboratorial screening to a few hundred candidates, enabling and accelerating the development of industrial enzymes. In this book chapter, the technology adopted at Zymvol is described. An overview of the current development and future directions in the company is also provided.

**Key words** Enzyme engineering, Biocatalysis, Bioinformatics, Computational enzyme design, Machine learning

## 1 Introduction

Natural enzymes are attractive templates for sustainable chemistry [1, 2] that need to be redesigned to meet industrial requirements. The most popular approach is directed evolution, where mutations are introduced randomly in the parental enzyme, mimicking natural evolution [3]. Although directed evolution is a revolutionary technique (awarded with the Nobel Prize for Chemistry in 2018), randomness and limited sequence sampling restrict the number of potential beneficial mutations. Computational techniques can drive directed evolution, recycling its results with machine learning [4] and/or selecting residue positions with high potential of improvement (hot spots) [5]. More radically, computational design can also replace directed evolution altogether in what is called rational design: in this case the sequence space screening is done in a computer, leaving only a few candidates for experimental validation [6]. The scope of this book chapter is to illustrate the principles of computational enzyme engineering adopted at Zymvol, our company, both when accompanying (hot spot prediction) or replacing (rational design) directed evolution. The main conceptual steps are described and reconducted to the existing literature.

Francesca Magnani, Chiara Marabelli and Francesca Paradisi (eds.), *Enzyme Engineering: Methods and Protocols*, Methods in Molecular Biology, vol. 2397, https://doi.org/10.1007/978-1-0716-1826-4_13,

Both our enzyme search (ES) and in silico design (ISD) pipelines combine multiple solutions from bioinformatics, protein design algorithms, and molecular modeling. The objective of ES is to find suitable biocatalysts for an input reaction on a target substrate, either as is or as a favorable template for ISD. The objective of ISD is to redesign (simulate the effect of mutations) the input enzyme to efficiently implement the target reaction over the target substrate, i.e., to achieve/improve a target chemical transformation. Properties other than activity/specificity/selectivity can be improved, like enzyme stability (to temperature, pH, solvent, etc.) and solubility.

The cornerstone of our technology is the combination of bioinformatics with protein design algorithms and empirical/physics-based simulations that model the interactions between the substrate and the enzyme. Such simulations follow a funnel like scheme with increasing complexity to gradually filter enzyme candidates (Fig. 1). In implicit solvent molecular modeling, both enzymes and ligands are fully flexible, i.e., both the protein side chains and backbone can move. On the other hand, the solvent (water) is modeled as a continuous electrostatic field, ignoring its detailed action as an ensemble of molecules [7]. This approximation is corrected in explicit solvent molecular modeling, where water is treated as an ensemble of molecules which can perturb the system dynamics or participate in the target reaction [7]. Quantum chemical algorithms can model any electronic rearrangement that occurs along the target reaction, improving accuracy [8]. Throughout the

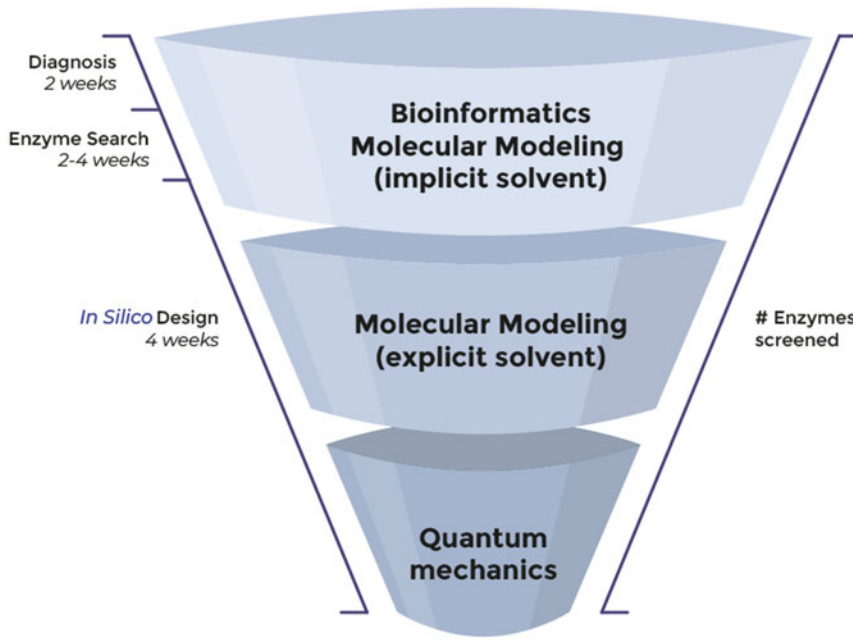


**Fig. 1** Computational workflow for enzyme engineering

pipeline simulations take into consideration not only mutations in the active site but also the potential effect of mutations located far away from the active site. Long range mutations can affect the encounters between the substrate and the catalytic residues in the active site or the migration of a substrate/product in/from the active site [9] (Fig. 2).

In ES, only bioinformatics and implicit solvent molecular modeling are adopted. In ISD, all the steps are included. However, depending on the nature and complexity of the task in hand, some of them may be excluded. For instance, quantum mechanics is not included when other parameters need to be optimized first, like binding the substrate in the right conformation to convey the desired reaction.

The program input typically comprises: a target transformation for the desired substrate and an initial enzyme sequence to be optimized (in case of ES the latter is not required). The ES output consists of 8–16 enzymes. Typically an ISD round delivers 90 enzyme variants ranked by priority, but customized solutions are available (like residue positions for saturation mutagenesis). Both ES and ISD (one round) typically take up to 4 weeks to deliver results.

In ISD, more prediction rounds (followed by experimental validation) are possible. Each iteration allows us to include experimental data into our platform, aiming at improving the accuracy of our predictions for the task in hand. Three ISD rounds are usually recommended.

For a list of definitions of concepts addressed in this chapter, *see* Table 1.

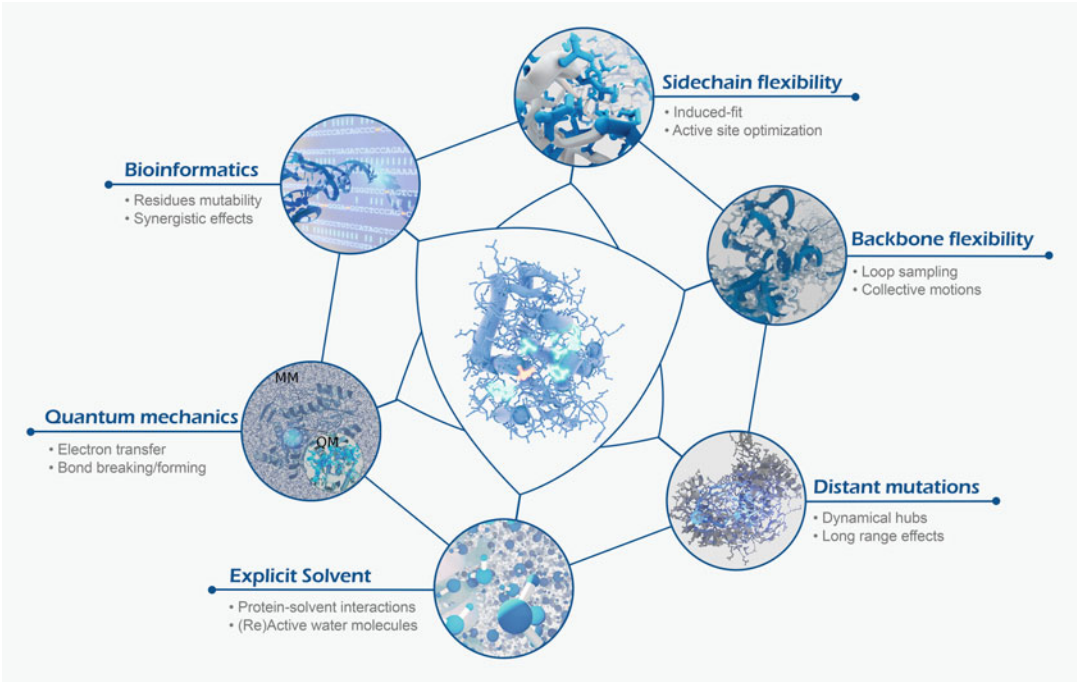


**Fig. 2** Levels of theory adopted at Zymvol

**Table 1**
**Definitions of concepts introduced in the manuscript**

| | |
|---|---|
| *Enzyme engineering related concepts* | |
| Mutagenesis | Introduction of mutations (changes of the nature of a given amino acid) in a given protein. |
| Enzyme (re)design | Modification of the enzyme sequence to reach a user-defined goal (design goal). |
| Hot spot | A residue whose mutation to any amino acid yields an improvement toward the design goal. |
| Directed evolution (design) | Thousands/millions of mutations are introduced in a protein; their effect is measured with rapid techniques (like colorimetric assays). |
| Rational design | A few mutations are introduced based on structural, sequence and mechanism knowledge. More quantitative assays are adopted to assess their effects. |
| In silico design (ISD) | Millions of mutations are virtually introduced and screened in a protein with bioinformatic and molecular modeling. The most promising are selected for experimental validation. |
| Experimental validation | Laboratorial implementation of the mutations suggested in an ISD campaign and assessment of their effects with respect to the design goal. |
| *Protein structure related concepts* | |
| Protein dynamics | The ensemble of tridimensional shapes (conformations) that a protein can adopt in its solvent. |
| Active site | Structural pocket of an enzyme where catalysis occurs. |
| Long range mutations | Mutations located far from the active site. |
| Protein backbone and side chain | All the atoms of a protein that include the peptide bond and the adjacent Cα atoms are part of the backbone. The remaining atoms are part of the side chain. |
| Near attack conformation (NAC) | Enzyme-substrate conformation that resembles the transition state of the catalytic reaction. |
| *Computational related concepts* | |
| Protein design algorithms | Algorithms designed to predict the effects of mutations on protein structural stability. |
| Bioinformatics | Analysis of large sequence databases (sequence analysis) to discover evolutionary and functional patterns. |
| Molecular modeling | Representation of a protein as a collection of atomic nuclei immersed in the energy potential of their electrons. |
| Physics-based simulations | Simulations of the behavior of a protein based on molecular dynamics, Monte Carlo simulations, and quantum chemistry. |
| Molecular dynamics | Simulations of the time evolution of a protein through the resolution of Newton's second law. |
| Monte Carlo simulations | Simulations of the dynamical behavior of a protein based on random displacements of its atoms. |
| Quantum chemistry | Simulations of a protein through the resolution of quantum mechanical equations. |

(continued)

**Table 1**
**(continued)**

| | |
|---|---|
| Implicit solvent | Representation of the solvent as a homogeneous electrostatic field. |
| Explicit solvent | Representation of the solvent as tridimensional molecules. |
| Substrate docking | Simulation of the interactions between the substrate and the enzyme active site. |
| Substrate/product migration | Entry/leaving of a substrate/product in/out of the enzyme active site. |

## 2 Bioinformatics and Molecular Modeling

In this methodological framework, two steps are included: diagnosis and mutagenesis.

In diagnosis, the design principles and mutagenesis library are the main output. The included steps of diagnosis are the following.

*Sequence analysis.* The amino acid sequence of the parental enzyme is *blasted* against UniRef90, the resulting sequences are aligned [10] and refined [10, 11]. The latter also provides an estimation of residue entropies, occupancies, and mutual information (MI) analysis. In such a way, only residues with enough entropy are retained for the final library and consensus mutations are suggested to stabilize the protein [12]. Also, potential synergistic effects are captured by MI [13].

*Protein dynamics.* The structure of the parental enzyme is prepared and its dynamical behavior is estimated with an internal protocol that merges information from normal modes [14], concoord [15], and loop sampling algorithms. A thorough analysis of the resulting structural ensemble can pinpoint potential hot spots far away from the active site. This analysis includes but is not limited to dynamical cross correlation matrix [16] and network techniques [17].

*Substrate docking.* The next step is to verify whether the substrate is capable of binding productively to the active site, i.e., respecting all the necessary catalytic contacts. If that were not the case, the first step is to introduce single mutations to alanine (Ala) to facilitate the formation. All noncatalytic residues within a certain threshold from the substrate are included in the computational mutagenesis library, provided that they are not highly conserved.

*Substrate/product migration.* Both the substrate and the product bound in the active site (in separate models) are perturbed with an in-house Monte Carlo algorithm (built within Rosetta [18])

where also the side chain and backbone of the surrounding residues are allowed to move. Simulating the exit of these molecules can pinpoint residues acting as kinetic bottlenecks, which usually display many contacts with the substrate. If they are mutable (i.e., their entropy is large enough), they are introduced in the computational mutagenesis library.

All the selected residues included in the library are allowed to mutate to populated amino acids, i.e., those that appear in the multiple sequence alignment (MSA) at those positions with enough frequency (threshold decided by the user) to preserve protein expression and robustness. The principle is somewhat similar to that behind PROSS [19] and FuncLib [20].

In the second step, mutagenesis, millions of enzyme variants are modeled and screened with an in-house protocol that includes proprietary and third-party software like Rosetta [21] and FoldX [22]. Then, the system is refined and relevant structural and energetic properties are extracted to evaluate every variant. A crude approximation is the simple substrate binding energy vs. catalytic distances plot.

In the last step, hundreds-thousands of variants are selected for the next phase. It should be noted that this is true only if the crystal structure or a high quality model of the parental enzyme is available. As a matter of fact, low/medium quality homology models have a tendency to unfold during molecular dynamics (MD) simulations [23], and the sampled ensemble might not be reliable. With bad homology models, the final list of mutants to be delivered to the lab needs to come from this single stage.

## 3 Molecular Dynamics (Explicit Solvent)

The selected variants are filtered based on MD simulations. Many features of designed enzymes can be inspected, including hydrogen bond strength, structural integrity and preorganization, solvent exposure. The main feature used to filter variants is the population of near attack conformations (NAC), i.e., structures that resemble the transition state (TS). This technique proved to be successful in tuning selectivity with high-throughput MD (HTMD) [24]. In HTMD many short simulations are run, which was shown to be more effective than traditional MD for NAC sampling [25].

As anticipated, MD is often not a viable solution in industrial applications because of the lack of a suitable crystal structure. Currently, we are evaluating alternative ways to account for NACs, based on Monte Carlo simulations.

## 4 Quantum Mechanics

Finally, chemical reactions are determined by electron rearrangements that lead to bond forming/breaking. These events cannot be accounted for by bioinformatics, molecular modeling and dynamics. Instead, quantum chemical calculations are necessary. Unfortunately, traditional techniques are as slow as informative so they are not suited to screen hundreds of variants [8]. Therefore, smart solutions are needed to decrease the resource and time burden. This passes through calculating properties, instead of TS energies, that: (1) correlate with activity; (2) converge much faster than energy during geometry optimization. One example from literature is our work with laccases [26, 27]. We developed a scoring method based on the total spin density localized on the substrate after five geometry optimization steps (the electron density roughly converged in that time) and a distance-dependent dielectric potential. Such a shortcut turned out to work well, as can be seen in Fig. 3. Testing four substrates against two laccases, a linear correlation between activity and spin density was found. In 2015, each calculation took ~4 h with six threads. With today's cloud computing facilities and better hardware, ~500 variants can be tested in 1 day for less than €500. In the past years, this solution was adopted in the successful design of a few oxidoreductases [28–31].

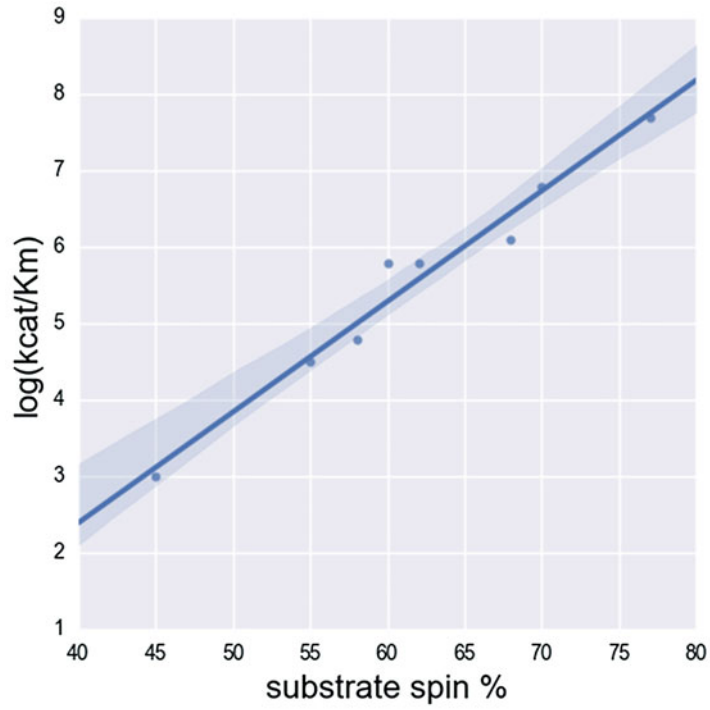


**Fig. 3** Electron density-based quantum chemical scoring of activity

## 5 Areas of Development

At Zymvol we are currently working/evaluating the following areas of development to improve our technology.

*Machine learning*. As nicely shown by Siegel and co-workers for glycosyl hydrolases in a recent machine learning (ML) study, no feature strongly correlates with the experimental data [32]. This suggests that enzymatic function is shaped by many different molecular properties, of which binding energies are not even included. This poses a serious question on the actual knowledge of how enzymes work and whether the key parameters to label activity are actually family and reaction dependent. These doubts were confirmed by ML analysis that we conducted on two data sets: a full mutagenesis study for a small protein (~60 residues) [33] and the activity data of 96 ligands against 16 esterases [34]. Although both systems can be predicted with a good degree of accuracy with their respective ML models, the selected features are distinct. Therefore, we aim at gathering thousands of data points for key enzyme families to conduct further research and develop family-dependent models. On a final note, machine learning techniques are also in the way to speed up quantum chemical estimations to unprecedented levels [35].

*Long range mutations*. It is a matter of fact that mutations that are far away from the active site can lead to remarkable increases in activity, improving catalytic preorganization in evolved enzymes [36]. Despite this success of MD in characterizing long range mutation effects [37–41], no reliable tool seems to be available for their quick detection. In our company we are developing tools that do not involve long and resource intensive MD simulations, shrinking the prediction time for hot spots from weeks to minutes. Moreover, we aim at not only singling out hot spots but actually predict the correct amino acid substitutions. This technology is also expected to have an impact on predicting the effect of distinct immobilization techniques on enzyme activity, which is of great importance in industry.

*Ancestral reconstruction*. Ancestral proteins often show exceptional properties (including promiscuity and thermal stability) due to the variegate environmental conditions that shaped our planet to what it is today. Statistical models are available to reconstruct the phylogeny of a given enzyme and trace it back to its ancestors [42].

*Quantum computing*. Finally, with a look toward a perhaps more distant future, one of the most promising suggested applications of quantum computing is solving classically intractable biochemistry problems [43]. This includes quantum mechanics calculations but also sampling problems. However, building a sufficiently large quantum computer remains a prohibitive challenge nowadays [43].

## Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no 873593. The authors acknowledge the members of the R&D team of Zymvol for their contribution to the technology: Jesus Seco, Ferran Sancho, Laura Masgrau, Ryoji Takahashi, Lur Alonso, and Marina Cañellas.

**Competing Interests**: *Authors are affiliated to Zymvol Biomodeling SL, holders of a specific technology protected as trade secret.*

### References

1. Itoh T, Hanefeld U (2017) Enzyme catalysis in organic synthesis. Green Chem 19:331–332
2. Sheldon RA, Woodley JM (2018) Role of biocatalysis in sustainable chemistry. Chem Rev 118:801–838
3. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10:866–876
4. Yang KK, Wu Z, Arnold FH (2019) Machine-learning-guided directed evolution for protein engineering. Nat Methods 16:687–694
5. Chica RA, Doucet N, Pelletier JN (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Curr Opin Biotechnol 16:378–384
6. Monza E, Acebes S, Fátima Lucas M, Guallar V (2017) Molecular modeling in enzyme design, toward in silico guided directed evolution. In: Directed enzyme evolution: advances and applications. Springer, pp 257–284
7. Skyner RE, McDonagh JL, Groom CR et al (2015) A review of methods for the calculation of solution free energies and the modelling of systems in solution. Phys Chem Chem Phys 17:6174–6191
8. Gao J, Truhlar DG (2002) Quantum mechanical methods for enzyme kinetics. Annu Rev Phys Chem 53:467–505
9. Agarwal PK (2019) A biophysical perspective on enzyme catalysis. Biochemistry 58:438–449
10. Larkin MA, Blackshields G, Brown NP et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948
11. Bakan A, Meireles LM, Bahar I (2011) ProDy: protein dynamics inferred from theory and experiments. Bioinformatics 27:1575–1577
12. Sternke M, Tripp KW, Barrick D (2019) Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. Proc Natl Acad Sci U S A 116:11275–11284
13. Penner O, Grassberger P, Paczuski M (2011) Sequence alignment, mutual information, and dissimilarity measures for constructing phylogenies. PLoS One 6:e14373
14. Demerdash O, Cui Q, Van Wynsberghe A, Li G (2005) Normal mode analysis of macromolecules: from enzyme activity site to molecular machines. In: Normal mode analysis, Chapman & Hall/CRC, pp 65–89
15. Seeliger D, Haas J, de Groot BL (2007) Geometry-based sampling of conformational

transitions in proteins. Structure 15:1482–1492

16. Hünenberger PH, Mark AE, van Gunsteren WF (1995) Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. J Mol Biol 252:492–503
17. Sladek V, Tokiwa H, Shimano H, Shigeta Y (2018) Protein residue networks from energetic and geometric data: are they identical? J Chem Theory Comput 14:6623–6631
18. Alford RF, Leaver-Fay A, Jeliazkov JR et al (2017) The Rosetta all-atom energy function for macromolecular modeling and design. J Chem Theory Comput 13:3031–3048
19. Goldenzweig A, Goldsmith M, Hill SE et al (2018) Automated structure- and sequence-based design of proteins for high bacterial expression and stability. Mol Cell 70:380
20. Khersonsky O, Lipsh R, Avizemer Z et al (2018) Automated design of efficient and functionally diverse enzyme repertoires. Mol Cell 72:178–186.e5
21. Kellogg EH, Leaver-Fay A, Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 79:830–838
22. Buß O, Rudat J, Ochsenreither K (2018) FoldX as protein engineering tool: better than random based approaches? Comput Struct Biotechnol J 16:25–33
23. Raval A, Piana S, Eastwood MP et al (2012) Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. Proteins 80:2071–2079
24. Wijma HJ, Floor RJ, Bjelic S et al (2015) Enantioselective enzymes by computational design and in silico screening. Angew Chem Int Ed Engl 54:3726–3730
25. Wijma H (2016) In silico screening of enzyme variants by molecular dynamics simulation. In: Svendsen A (ed) Understanding enzymes. Pan Stanford Publishing, pp 805–833
26. Monza E, Lucas MF, Camarero S et al (2015) Insights into laccase engineering from molecular simulations: toward a binding-focused strategy. J Phys Chem Lett 6:1447–1453
27. Lucas MF, Monza E, Jørgensen LJ et al (2017) Simulating substrate recognition and oxidation in laccases: from description to design. J Chem Theory Comput 13:1462–1467
28. Santiago G, de Salas F, Fátima Lucas M et al (2016) Computer-aided laccase engineering: toward biological oxidation of arylamines. ACS Catal 6:5415–5423
29. Giacobelli VG, Monza E, Fatima Lucas M et al (2017) Repurposing designed mutants: a valuable strategy for computer-aided laccase engineering—the case of POXA1b. Cat Sci Technol 7:515–523
30. Mateljak I, Monza E, Lucas MF et al (2019) Increasing redox potential, redox mediator activity, and stability in a fungal laccase by computer-guided mutagenesis and directed evolution. ACS Catal 9:4561–4572
31. Pardo I, Santiago G, Gentili P et al (2016) Re-designing the substrate binding pocket of laccase for enhanced oxidation of sinapic acid. Cat Sci Technol 6:3900–3910
32. Carlin DA, Caster RW, Wang X et al (2016) Kinetic characterization of 100 glycoside hydrolase mutants enables the discovery of structural features correlated with kinetic constants. PLoS One 11:e0147596
33. van der Meer J-Y, Poddar H, Baas B-J et al (2016) Using mutability landscapes of a promiscuous tautomerase to guide the engineering of enantioselective michaelases. Nat Commun 7:10911
34. Martínez-Martínez M, Coscolín C, Santiago G et al (2018) Determinants and prediction of esterase substrate promiscuity patterns. ACS Chem Biol 13:225–234
35. Qiao Z, Welborn M, Anandkumar A et al (2020) OrbNet: deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. J Chem Phys 153:124111
36. 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 48:1080–1089
37. Romero-Rivera A, Iglesias-Fernández J, Osuna S (2018) Exploring the conversion of a d-sialic acid aldolase into a l-KDO aldolase. Eur J Org Chem 2018:2603–2608
38. Maria-Solano MA, Serrano-Hervás E, Romero-Rivera A et al (2018) Role of conformational dynamics in the evolution of novel enzyme function. Chem Commun 54:6622–6634
39. Romero-Rivera A, Garcia-Borràs M, Osuna S (2017) Role of conformational dynamics in the evolution of retro-aldolase activity. ACS Catal 7:8524–8532
40. Hong N-S, Petrović D, Lee R et al (2018) The evolution of multiple active site configurations

in a designed enzyme. Nat Commun 9(1):390. https://doi.org/10.1038/s41467-018-06305-y

41. Petrović D, Risso VA, Kamerlin SCL, Sanchez-Ruiz JM (2018) Conformational dynamics and enzyme evolution. J R Soc Interface 15:20180330. https://doi.org/10.1098/rsif.2018.0330
42. Joy JB, Liang RH, McCloskey RM et al (2016) Ancestral reconstruction. PLoS Comput Biol 12:e1004763
43. McArdle S, Endo S, Aspuru-Guzik A et al (2020) Quantum computational chemistry. Rev Mod Phys 92:015003