

COMPUTATIONAL STUDIES OF THE CONFORMATIONAL LANDSCAPE OF ALLOSTERIC AND ENANTIOSELECTIVE ENZYMES

Miguel Ángel María Solano

Per citar o enllaçar aquest document:

Para citar o enlazar este documento:

Use this url to cite or link to this publication:

<http://hdl.handle.net/10803/671771>

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

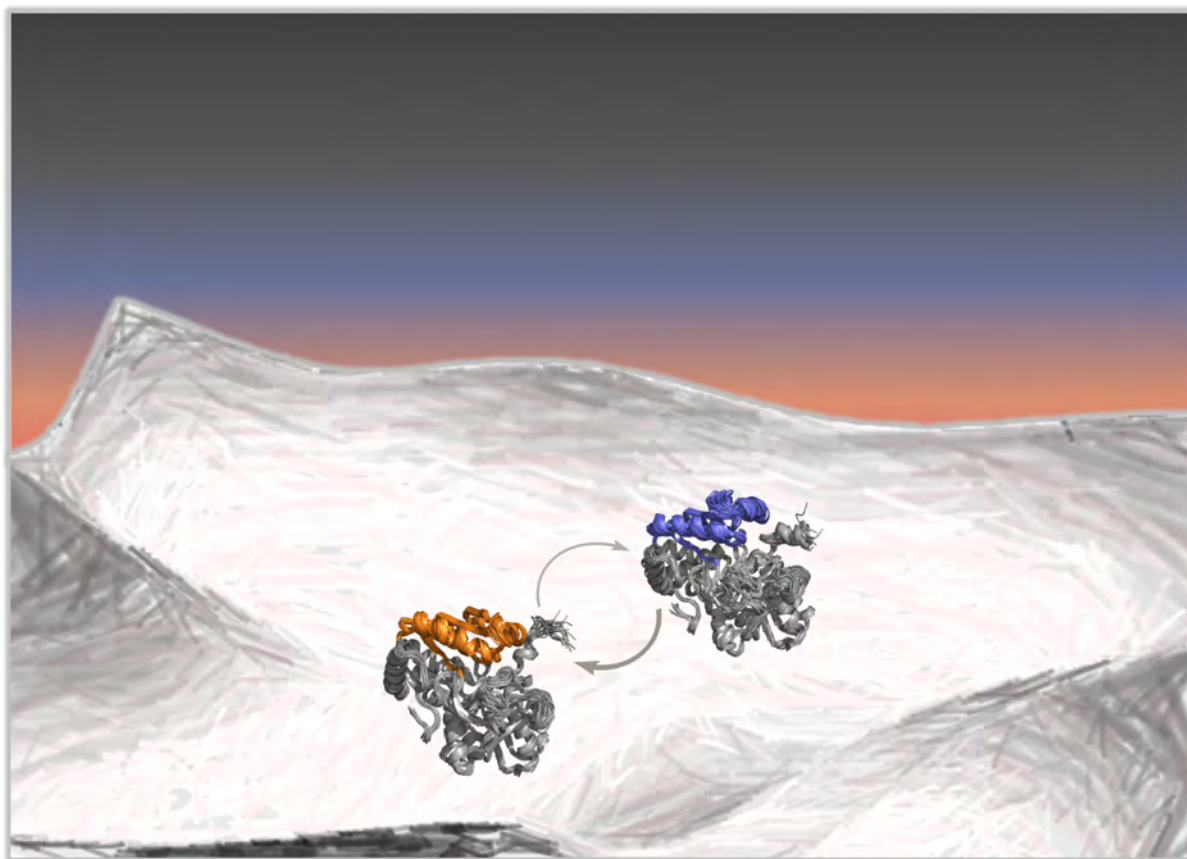
ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.



DOCTORAL THESIS

**Computational studies of the conformational landscape of
allosteric and enantioselective enzymes**



Miguel Ángel María Solano

2021



DOCTORAL THESIS

**Computational studies of the conformational landscape of
allosteric and enantioselective enzymes**

Miguel Ángel María Solano

2021

Doctoral programme in Chemistry

Supervised by: Prof. Sílvia Osuna and Prof. Marcel Swart
Tutor: Prof. Marcel Swart

Presented to obtain the degree of PhD at University of Girona



Prof. Sílvia Osuna and Prof. Marcel Swart from University of Girona,

WE DECLARE:

That the thesis entitled “Computational studies of the conformational landscape of allosteric and enantioselective enzymes”, presented by Miguel Ángel María Solano to obtain a doctoral degree, has been completed under our supervision and meets the requirements to opt for an International Doctorate.

For all intents and purposes, we hereby sign this document.

Prof. Sílvia Osuna

Prof. Marcel Swart

Girona, January 10th, 2021

*Dedicated to all scientists that I had the pleasure to
work alongside*

ACKNOWLEDGEMENTS

Most affectionate thanks to my parents for their unconditional support when I decided to embark on my scientific carrier. They have always encouraged me to make such decisions for myself and have been crucial to naturally finding my professional and personal path in the fascinating field of biochemistry.

Specially thanks to the supervisors, post-doc researchers, PhD students, Master students and undergraduate students that have contributed directly or indirectly to the projects carried out in this thesis and to my development as scientist. I would like to highlight the supervisors Francisco Garcia Cánovas, Sílvia Osuna, Marcel Swart and Roberto Chica; the post-docs Javier Iglesias, Ferran Feixas, Marc Garcia Borràs and Aron Broom; the PhD students Vanessa Ortiz, Adrià Romero, Lorenzo D'Amore, Miquel Estévez, Christian Curado and Carla Calvó; the Master student Nurzhan Mukhametzhanov and the undergraduate student Oriol Canal. Thank you so much for the knowledge transfer, the scientific discussions and the great after work time we have been through.

The projects of this thesis have been performed thanks to the financial support granted by: The Spanish MINECO for a PhD fellowship (BES-2015-074964), the Spanish MICINN for the project PGC2018-102192-B-I00 and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC-2015-StG- 679001). I also thank the university of Girona, and in particular the Institut de Química Computacional i Catàlisi (IQCC) for the material resources and facilities provided, and the Barcelona Supercomputing Center (BSC) for the computational resources. Finally, I want to highlight the support exerted by the Spanish government during the COVID 19 pandemic extending me the fellowship.

FULL LIST OF PUBLICATIONS

This Thesis is presented as a compendium of publications.

The following published articles have been included as chapters in this Thesis:

1) Maria-Solano, M. A.; Romero-Rivera, A.; Osuna, S. Exploring the reversal of enantioselectivity on a zinc-dependent alcohol dehydrogenase. *Org. Biomol. Chem.* **2017**, *15*, 4122-4129. [Chemistry, Organic, 3.564, Q1]

2) Li, G; Maria-Solano, M. A.; Romero-Rivera, A.; Osuna, S.; Reetz, M. T. Inducing high activity of a thermophilic enzyme at ambient temperatures by directed evolutions. *Chem. Commun.* **2017**, *53*, 9454-9457. [Chemistry, Multidisciplinary; 6.32; Q1]

3) Maria-Solano, M.A.; Iglesias-Fernández, J.; Osuna, S. Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution, *J. Am. Chem. Soc.* **2019**, *141*, 13049-13056. [Chemistry, Multidisciplinary; 14.70; Q1].

4) Maria-Solano, M.A.; Kinateder, T.; Iglesias-Fernández, J.; Sterner, R.; Osuna, S. Rational prediction of distal activity-enhancing mutations in tryptophan synthase [*submitted*].

Other published articles not included in this Thesis:

5) Maria-Solano, M. A.; Serrano-Hervás, E.; Romero-Rivera, A.; Iglesias-Fernández, J.; Osuna, S. Role of conformational dynamics in the evolution of novel enzyme function. *Chem. Commun.* **2018**, *54*, 6622-6634. [Chemistry, Multidisciplinary; 6.29; Q1]

6) Chen, X.; Zhang, H.; Maria-Solano, M.A.; Liu, W.; Li, J.; Feng, J.; Liu, X.; Osuna, S.; Guo, R.; Wu, Q.; Zhu, D.; Ma, Y. Efficient reductive desymmetrization of bulky 1,3-cyclodiketones enabled by structure-guided directed evolution of a carbonyl reductase. *Nat. Catal.* **2019**, *2*, 931-941. [Chemistry, Physical; 30.47; Q1]

7) Li, G.; Qin, Y.; Fontaine, N. T.; Ng Fuk Chong, M.; Maria-Solano, M.A.; Feixas, F.; Cadet, X.; Pandjaitan, R.; Garcia-Borràs, M.; Cadet, F.; Reetz, M. Machine Learning Enables

Selection of Epistatic Enzyme Mutants for Stability Against Unfolding and Detrimental Aggregation. *ChemBiochem.* **2020**, 21, 1-12. [Biochemistry & Molecular biology; 2.58; Q3]

8) Calvó-Tusell, C.; Maria-Solano, M.A.; Feixas, F.; Osuna, S. Unravelling the millisecond allosteric activation of Imidazole Glycerol Phosphate Synthase (IGPS). *In preparation*

LIST OF ABBREVIATIONS

Abbreviation	Description
ADH	Alcohol Dehydrogenases
AdK	Adenylate Kinase
aMD	Accelerated Molecular Dynamics
ATP	Adenosine Tri-Phosphate
CADEE	Computer-Aided Directed Evolution of Enzymes
cAMP	cyclic Adenosine Mono-Phosphate
CAP	Catabolite Activator Protein
CASCO	Catalytic Selectivity by Computational Design
CAST	Combinatorial Active Site Test
CDM	Cationic Dummy Model
CVs	Collective Variables
DE	Directed Evolution
DOFs	Degrees of Freedom
DSD	Differential Scanning Calorimetry
emf	electromotive force
EVB	Empirical Valence Bond
FEL	Free Energy Landscape
FEP	Free Energy Perturbation
FRESCO	Framework for Rapid Enzyme Stabilization by Computational Libraries
GAFF	Generalized Amber Force Field
G3P	Glyceraldehyde-3-Phosphate
IGP	Indole-3-Glycerol Phosphate
IGPS	Imidazole Glycerol Phosphate Synthase
ISM	Iterative Saturation Mutagenesis
ITC	Isothermal Titration Calorimetry
LJ	Lennard-Jones
LBCA	Last Bacterial Common Ancestor
MD	Molecular Dynamics
MM	Molecular Mechanics
MM-PBSA	Molecular Mechanics-Poisson Boltzmann Surface Area
MSD	Multistate Design
MSMs	Markov State Models
NACs	Near Attack Conformation
NAD	Nicotinamide Adenine Dinucleotide
NMR	Nuclear Magnetic Resonance
NPT	Isobaric-isothermal ensemble

NVE	Microcanonical ensemble
NVT	Canonical ensemble
PCA	Principal Component Analysis
PELE	Protein Energy Landscape Exploration
PLP	Pyridoxal Phosphate
PMF	Proton Motive Force
QM	Quantum Mechanics
RLS	Rate Limiting Step
RA	Retro Aldolases
RESP	Restrained Electrostatic Potential
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SDM	Site Directed Mutagenesis
SSM	Site Saturation Mutagenesis
SPE	Single Point Energy
SPM	Shortest Path Map
TI	Thermodynamic Integration
tICA	time-structure Independent Component Analysis
TrpS	Tryptophan Synthase
TS	Transition State
TST	Transition State Theory
US	Umbrella Sampling

CONTENTS

SUMMARY	1
RESUM	3
RESUMEN	5
Chapter 1. Introduction	7
Life and enzymes	8
1.1 Kinetic view.....	9
1.1.1 Some basic kinetic concepts	9
1.1.2 Michaelis-Menten equation	11
1.1.3 Steady-state theory	13
1.1.4 Multi-step enzyme cycles	14
1.1.5 Microscopic rate constants.....	16
1.2 Classical thermodynamic view	17
1.2.1 First law of thermodynamics.....	17
1.2.2 Second law of thermodynamics	20
1.2.3 Third law of thermodynamics	23
1.2.4 Gibbs energy	25
1.2.5 The art of the biochemical work	27
1.3 Statistical thermodynamic view	30
1.3.1 Fundamentals of Energy landscapes	31
1.3.2 Free Energy landscapes	34
1.3.3 Conformational free energy landscapes in proteins	35
1.4 Chemical view	37
1.4.1 Fundamentals of catalysis.....	37
1.4.2 Transition state stabilization	40
1.4.3 Role of conformational dynamics in catalysis	43
1.5 Engineering view	46
1.5.1 Overview of enzyme engineering approaches	46
1.5.2 A brief story of enzyme computational design	47
1.5.3 Engineering stereoselectivity, thermostability and allosteric properties	53
1.5.4 Ancestral enzyme properties.....	62
1.5.5 Computational design outlook	63
Chapter 2. Methodologies	65
<i>In silico</i> approaches for enzyme studies.....	66

2.1 Classical Force Fields.....	67
2.1.1 Potential energy function and parameters.....	67
2.1.2 Missing parameters in metalloenzymes.....	73
2.2 Molecular Dynamics	74
2.2.1 Newtonian dynamics	74
2.2.2 Initial velocities	76
2.2.3 Time step.....	77
2.2.4 Periodic boundary conditions and cutoff distance	78
2.3 Running a MD simulation	79
2.4 Free energy landscape construction	80
2.4.1 Dimensionality reduction of the MD data set.....	80
2.4.2 The sampling problem.....	82
2.5 Enhanced sampling techniques	82
2.5.1 Unbiased MD methods	83
2.5.2 Biased MD methods	84
2.6 Residue-by-residue correlation and proximity tools	88
Chapter 3. Objectives	91
Chapter 4. Enantioselectivity and thermoadaptation properties of alcohol dehydrogenase (ADH) enzymes	95
4.1 Exploring the reversal of enantioselectivity on a zinc-dependent alcohol dehydrogenase	97
4.2 Inducing high activity of a thermophilic enzyme at ambient temperature by directed evolution.....	121
Chapter 5. Allosteric properties and stand-alone function of tryptophan synthase (TrpS) enzymes	135
5.1 Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution.....	137
5.2 Rational prediction of distal activity-enhancing mutations in tryptoptan synthase....	169
Chapter 6. Results and discussion	207
6.1 Alcohol dehydrogenase (ADH): enantiosectivity and thermostablity studies.....	208
6.2 Tryptophan synthase (TrpS): allostery and stand-alone function studies	212
6.3 Ending Thoughts	221
Chapter 7. Conclusions.....	223
Bibliography.....	227

LIST OF FIGURES

Figure 1.1 Substrate concentration effect on the initial velocity for an enzymatic reaction obeying Michaelis-Menten kinetics.....	12
Figure 1.2 Representation of the enzyme velocity along time. The steady-state phases are also showed.	13
Figure 1.3 On the left, one dimensional energy landscape exhibiting different energy basins and energy barriers. The set of energy-basins are grouped into three different states (A , B , and C) according to activity and folding enzyme descriptors. On the right, single energy basin representation from the original function (solid green line) together with the harmonic approximation (dashed black line).....	33
Figure 1.4 Schematic representation of an enzyme free energy landscape in the <i>apo</i> state associated with an open to closed conformational exchange and the population shift towards the closed state induced by a substrate (A), mutations (B), a protein partner (C), a ligand in protein partner (D), and covalent modifications (E).....	36
Figure 1.5 Reaction coordinate diagrams. It is shown the free energy as a function of the course of the chemical reaction of the substrate conversion to product (A), together with alteration on the free energy profiles when the reaction is catalyzed by the enzyme for different mechanisms (B , C and D).....	39
Figure 1.6 Hypothetical reaction coordinate diagram for an enzyme catalyzed-reaction. The structural changes of the enzyme and substrate along the reaction are depicted.....	44
Figure 1.7 Time scales of the different fast and slow motions in proteins.....	45
Figure 1.8 Schematic representation of the <i>inside-out</i> protocol and post-optimization by means of DE rounds, the active site positions are shown as sticks while the distal mutations as blue spheres (A) together with the FEL of <i>de novo</i> design and the most efficient DE variant (B).....	49
Figure 1.9 Representation of the two main allosteric mechanisms including the conformation-based (A) and the dynamics-based (B) allostery.	59
Figure 2.1 Representation of the three layers of accuracy to study enzyme reactivity and conformational dynamics computationally.	67
Figure 2.2 Representation of the potential energy function of the bond stretching term between A and B atoms, with k_{AB} the force constant of the bond, r_{AB} the bond length and $r_{AB,eq}$ the equilibrium distance.	68

Figure 2.3 Representation of the potential energy function of the angle bending term between A, B and C atoms, with k_{ABC} the force constant of the bond, θ_{ABC} the angle and $\theta_{ABC,eq}$ the equilibrium angle.....	69
Figure 2.4 Representation of the potential energy function of the bond twisting term between A, B, C and D atoms, with V_{ABCD} the torsional force constant, n the multiplicity of the \cos function, ω_{ABCD} the dihedral angle and γ_{ABCD} the phase angle.	70
Figure 2.5 Representation of the potential energy function of the van der Waals term between A and B atoms, with ϵ_{AB} the well-depth, r_{AB} the distance between A and B atoms and σ_{AB} the interatomic distance at which repulsive and attractive forces exactly balance.....	70
Figure 2.6 Representation of the potential energy function of the electrostatic term between A and B atoms, with q_A and q_B the atomic charges of atoms A and B, respectively, r_{AB} the distance between A and B and ϵ_0 the dielectric constant (usually set to 1).....	71
Figure 2.7 Schematic view of the dimensional reduction process of multiple MDs accumulated data set.....	80
Figure 2.8 Representation of a metadynamics simulation. The repulsive potentials are deposited to the collective variable over time (from A to D) until the full free energy landscape is covered (D).....	86
Figure 2.9 Schematic view of the Shortest Path Map (SPM) construction from the MD simulations data set. The blue spheres represent the graph nodes while the lines that connect them the edges.....	89
Figure 6.1. Representation of some representative snapshots of the different conformational states sampled along the MD simulations for the TbSADH ^{W110T} and TbSADH ^{I86A} starting from the <i>pro</i> -(R) (in orange) and <i>pro</i> -(S) (in blue) orientations of 1a , respectively. The histogram of the hydride transfer distance together with the <i>pro</i> -(R)/ <i>pro</i> -(S) angle between 1 ^a and an active site residue is displayed for both variants. High and low angle (in degrees) values represent <i>pro</i> -(R) and <i>pro</i> -(S) conformations, respectively. Short hydride transfer distances (in Å) values above the dashed line indicate catalytically productive orientations.....	209
Figure 6.2 Overlay of representative snapshots for WT (A) and A85G/I86A variant (B) in the <i>apo</i> state at 30 °C. Root Mean Square Fluctuation (RMSF) values of all residues computed from the MD simulations in the <i>apo</i> state (C).....	211

Figure 6.3 On the left, Free energy landscape (FEL) associated with the COMM domain open-to-closed (O-to-C) conformational exchange of the *PfTrpS* complex (**A**), *PfTrpB* isolated (**B**) and *PfTrpB0B2* stand-alone (**C**) enzymes at A_{in} , A_{ex1} , and Q_2 reaction intermediates. The x-axis corresponds to the progression along the reference O-to-C path generated from X-Ray data, while the y-axis to the mean square deviation (MSD) distance from the reference path. On the right, overlay of the *PfTrpS* metastable conformations of the open state at A_{in} intermediate, partially closed at A_{ex1} , and closed at Q_2 , respectively (**A** left panel), the metastable conformations of the closed states at Q_2 intermediate for *PfTrpB* and *PfTrpS* (**B** left) and the metastable conformations of the closed states at Q_2 intermediate for *PfTrpB0B2* and *PfTrpS* (**C** left). The Detailed active site view of the metastable closed conformations at Q_2 intermediate for the *PfTrpS* complex (**A** right panel), *PfTrpB* isolated (**B** right) and *PfTrpB^{0B2}* stand-alone (**C** right) is also shown together with the catalytic distances (in Å) between charge-charge stabilization E104- Q_2 and proton transfer K82- Q_2215

Figure 6.4 SPM-based computational workflow for the rational design of SPM6 TrpB enzyme variant. By analyzing the conformational ensemble of the stand-alone LBCA TrpB with high catalytic activity (upper left ensemble) through the SPM, we identified positions (grey spheres, lower left structure) within allosteric pathways (black edges) in the enzyme that most contribute to the LBCA TrpB conformational dynamics in the Q_2 intermediate. Thereby the size of each edge and node corresponds to the relevance for conformational dynamics; catalytic K84 is highlighted in yellow. Excluding residues that do not participate in an allosteric pathway reduces the sequence space from 20^{393} to 20^{74} possible activity enhancing substitutions. Sequence comparison at the SPM positions between stand-alone LBCA TrpB and inefficient ANC3 TrpB reduces the sequence space to 6 mutations with respect to LBCA TrpB (lower right structure, purple residues), that were introduced into ANC3 TrpB (upper right structure, purple residues) and tested *in-vitro*.....219

LIST OF TABLES

Table 1 Definitions of various thermodynamic and kinetic stability parameters.....55

SUMMARY

Enzymes are sophisticated biomacromolecules that accelerate chemical reactions by several orders of magnitude in favor of cell demands. Such great rate acceleration comes from a precisely pre-organized active site pocket that preferentially stabilizes the transition state(s) of the reaction. Enzymes are also highly dynamic and their function is linked to its three-dimensional structure and the broad range of accessible conformations that can be sampled in solution. Because of their great catalytic power and environmental sustainability, enzymes emerge as a powerful alternative with respect to the conventional metal-based catalysts for the production of biofuels, agricultural chemicals and pharmaceutical drugs. However, natural enzymes do not cover the widespread industrial purposes and thus their function needs to be engineered. Given the vast sequence and conformational space of proteins, the rational design of enzymes for novel function is an extremely challenging task.

This thesis starts with an introduction (**Chapter 1**) that provides the reader with substantial information about the nature of enzymes and how they work from different meaningful points of view (kinetic, classical thermodynamics, statistical thermodynamics, chemical and engineering). Secondly, **Chapter 2** is focused on the fundamentals of molecular mechanics (MM) and molecular dynamics (MD) together with the different computational methodologies employed to study the conformational energy landscape of proteins. It follows with the main objectives of the thesis (**Chapter 3**) that consist of the exploration of the dynamic conformational ensemble of proteins and the study of its connection with enzyme properties such as enantioselectivity and allostery by means of computational techniques. In particular, we target the rationalization of the novel functions achieved in laboratory-evolved enzyme variants and the development of new rational design strategies focused on the enzyme conformational dynamics.

The results of the four published projects carried out along this thesis are discussed in **Chapters 4 and 5**. **Chapter 4** is devoted to the study of a zinc dependent alcohol dehydrogenase (ADH) and encompasses two published projects. In the first project (**Chapter 4.1**) the bonded-model protocol for metalloenzymes is applied to study the conformational dynamics of ADHs in order to investigate the molecular basis of the reversion of

enantioselectivity displayed by laboratory-evolved enzyme variants. The second project (**Chapter 4.2**) was performed in collaboration with an experimental group. This work focuses on the rationalization of the enhanced activity and enantioselectivity towards a non-natural substrate at ambient temperatures with little trade off in thermostability of an ADH variant evolved by Reetz and coworkers through directed evolution (DE).

Chapter 5 includes the third and the fourth projects focused on the exploration of Tryptophan synthase enzyme (TrpS: composed of TrpA and TrpB subunits) allosteric properties. In the third project (**Chapter 5.1**) enhanced sampling techniques are employed to reconstruct the free energy landscape (FEL) of Tryptophan synthase (TrpS) enzyme associated with an allosteric transition. The main goal is to decipher the origin of the loss of activity of the TrpB subunit in absence of the TrpA protein binding partner and the recovery of stand-alone TrpB activity achieved in laboratory-evolved stand-alone TrpB enzyme variants. In the last project (**Chapter 5.2**), the information obtained in **Chapter 5.1** is used to face the challenge of the rational design of allosteric properties in collaboration with Sterner and coworkers. In particular a computational strategy using our *in-house* Shortest Path Map (SPM) correlation-based tool is developed and tested for the design of TrpB stand-alone enzyme variants.

Finally, **Chapter 6** includes a brief discussion of the main results presented in **Chapters 4** and **5**, and the main conclusions of this thesis are summarized in **Chapter 7**. This thesis provides useful information to address the computational evaluation of enantioselectivity and allosteric enzymatic properties to investigate the effects induced by mutations on the conformational energy landscape of enzymes. The studies gathered in this thesis emphasize the relevance of considering the enzyme conformational dynamics in the computational enzyme engineering processes.

RESUM

Els enzims són biomacromolècules sofisticades que acceleren les reaccions químiques diversos ordres de magnitud a favor de les demandes cel·lulars. Aquesta gran acceleració prové d'una preorganització precisa de la regió activa del enzim que estabilitza preferentment l'estat o estats de transició de la reacció. Els enzims també són molt dinàmics i la seva funció està lligada a la seva estructura tridimensional i a l'àmplia gamma de conformacions accessibles que es poden mostrejar en solució. A causa del seu gran poder catalític i la seva sostenibilitat mediambiental, els enzims apareixen com una alternativa poderosa respecte als catalitzadors convencionals basats en metalls per a la producció de biocombustibles, productes químics agrícoles i medicaments farmacèutics. No obstant això, els enzims naturals no cobreixen els amplis propòsits industrials i, per tant, cal dissenyar la seva funció. Donat el gran espai de seqüència i conformacional de les proteïnes, el disseny racional dels enzims per a una nova funció és una tasca extremadament difícil.

Aquesta tesi comença amb una introducció (**Capítol 1**) que proporciona al lector una informació substancial sobre la naturalesa dels enzims i el seu funcionament des de diferents punts de vista significatius (cinètica, termodinàmica clàssica, termodinàmica estadística, química i enginyeria). En segon lloc, el **Capítol 2** se centra en els fonaments de la mecànica molecular (MM) i la dinàmica molecular (DM) juntament amb les diferents metodologies computacionals emprades per estudiar el paisatge energètic conformacional de les proteïnes. A continuació s'exposen els objectius principals de la tesi (**Capítol 3**) que consisteixen en l'exploració del conjunt de conformacions dinàmiques de proteïnes i estudiar la seva connexió amb propietats enzimàtiques com l'estereoselectivitat i l'al·lostèria mitjançant tècniques computacionals. En particular, ens orientem a la racionalització de la nova funció assolida en variants enzimàtiques desenvolupades al laboratori i al desenvolupament de noves estratègies de disseny racional centrat en la dinàmica conformacional enzimàtica.

Els resultats dels quatre projectes publicats realitzats al llarg d'aquesta tesi es discuteixen als **Capítols 4 i 5**. El **Capítol 4** està dedicat a l'estudi d'una alcohol deshidrogenasa dependent del metall zinc (ADH) i inclou dos projectes publicats. En el primer projecte (**Capítol 4.1**)

s'aplica el protocol de model enganxat per a metalloenzims per estudiar la dinàmica conformacional dels ADH per tal d'investigar les bases moleculars de la reversió de l'enantioselectivitat mostrada per variants enzimàtiques evolucionades al laboratori. El segon projecte (**Capítol 4.2**) es va realitzar en col·laboració amb un grup experimental. En aquest treball se centra en la racionalització de la millora d'activitat i enantioselectivitat en un substrat no natural a temperatures ambientals i amb poca pèrdua de la termoestabilitat d'una variant d'ADH desenvolupada per Reetz i els seus companys de treball a través de l'evolució dirigida (ED).

El **Capítol 5** inclou el tercer i el quart projectes centrats en l'exploració de les propietats al·lostèriques de la triptòfan sintasa (TrpS: composta per les subunitats TrpA i TrpB). Al tercer projecte (**Capítol 5.1**) s'utilitzen tècniques de mostreig millorades per reconstruir el paisatge d'energia lliure (PEL) de l'enzim triptòfan sintasa (TrpS) associat a una transició al·lostèrica. L'objectiu principal és desxifrar l'origen de la pèrdua d'activitat de la subunitat TrpB en absència de la seva proteïna associada TrpA i la recuperació de l'activitat autònoma de TrpB aconseguida en variants enzimàtiques TrpB autònomes desenvolupades al laboratori. En el darrer projecte (**Capítol 5.2**), la informació obtinguda al **Capítol 5.1** s'utilitza per afrontar el repte del disseny racional de propietats al·lostèriques en col·laboració amb Sterner i els seus companys de treball. En particular, es desenvolupa una estratègia computacional que utilitza la nostra eina Mapa de camins més curt (MCC) basada en la correlació per al disseny de variants enzimàtiques autònomes de TrpB.

Finalment, el **Capítol 6** inclou una breu discussió dels principals resultats presentats als **Capítols 4 i 5**, i les principals conclusions d'aquesta tesi es resumeixen al **Capítol 7**. Aquesta tesi proporciona informació útil per abordar l'avaluació computacional de l'enantioselectivitat i les propietats enzimàtiques al·lostèriques per investigar els efectes induïts per mutacions en el paisatge energètic conformacional dels enzims. Els estudis recollits en aquesta tesi emfatitzen la rellevància de considerar la dinàmica conformacional dels enzims en els processos d'enginyeria computacional d'enzims.

RESUMEN

Las enzimas son biomacromoléculas sofisticadas que aceleran las reacciones químicas varios órdenes de magnitud a favor de las demandas celulares. Esta gran aceleración proviene de una preorganización precisa de la región activa de la enzima que estabiliza preferentemente el estado o estados de transición de la reacción. Las enzimas también son muy dinámicas y su función está ligada a su estructura tridimensional y a la amplia gama de conformaciones accesibles que se pueden muestrear en solución. Debido a su gran poder catalítico y su sostenibilidad medioambiental, las enzimas aparecen como una alternativa poderosa respecto a los catalizadores convencionales basados en metales para la producción de biocombustibles, productos químicos agrícolas y medicamentos farmacéuticos. Sin embargo, las enzimas naturales no cubren los amplios propósitos industriales y, por tanto, hay que diseñar su función. Dado el gran espacio de secuencia y conformacional de las proteínas, el diseño racional de las enzimas para una nueva función es una tarea extremadamente difícil.

Esta tesis comienza con una introducción (**Capítulo 1**) que proporciona al lector una información sustancial sobre la naturaleza de las enzimas y su funcionamiento desde diferentes puntos de vista significativos (cinética, termodinámica clásica, termodinámica estadística, química e ingeniería). En segundo lugar, el **Capítulo 2** se centra en los fundamentos de la mecánica molecular (MM) y la dinámica molecular (DM) junto con las diferentes metodologías computacionales utilizadas para estudiar el paisaje energético conformacional de las proteínas. A continuación se exponen los objetivos principales de la tesis (**Capítulo 3**) que consisten en la exploración del conjunto de conformaciones dinámicas de proteínas y estudiar su conexión con propiedades enzimáticas como la estereoselectividad y el al·losteria mediante técnicas computacionales. En particular, nos orientamos en la racionalización de la nueva función alcanzada en variantes enzimáticas desarrolladas en el laboratorio y el desarrollo de nuevas estrategias de diseño racional centrado en la dinámica conformacional enzimática.

Los resultados de los cuatro proyectos publicados realizados a lo largo de esta tesis se discuten en los **Capítulos 4 y 5**. El **Capítulo 4** está dedicado al estudio de una alcohol deshidrogenasa dependiente del metal zinc (ADH) e incluye dos proyectos publicados. En el primer proyecto

(**Capítulo 4.1**) se aplica el protocolo de modelo enganchado para metaloenzimas para estudiar la dinámica conformacional de los ADH para investigar las bases moleculares de la reversión de la enantioselectividad mostrada por variantes enzimáticas evolucionadas en el laboratorio. El segundo proyecto (**Capítulo 4.2**) se realizó en colaboración con un grupo experimental. En este trabajo se centra en la racionalización de la mejora de actividad y enantioselectividad en un sustrato no natural a temperaturas ambientales y con poca pérdida de la termoestabilidad de una variante de ADH desarrollada por Reetz y sus compañeros de trabajo a través de la evolución dirigida (ED).

El **Capítulo 5** incluye el tercer y el cuarto proyectos centrados en la exploración de las propiedades alostericas de la enzima triptófano sintasa (TrpS: compuesta por las subunidades TrpA y TrpB). En el tercer proyecto (**Capítulo 5.1**) se utilizan técnicas de muestreo mejoradas para reconstruir el paisaje de energía libre (PEL) de la enzima triptófano sintasa (TrpS) asociado a una transición alostérica. El objetivo principal es descifrar el origen de la pérdida de actividad de la subunidad TrpB en ausencia su proteína asociada TrpA y la recuperación de la actividad autónoma de TrpB conseguida en variantes enzimáticas TrpB autónomas desarrolladas en el laboratorio. En el último proyecto (**Capítulo 5.2**), la información obtenida en el **Capítulo 5.1** se utiliza para afrontar el reto del diseño racional de propiedades alostericas en colaboración con Sterner y sus compañeros de trabajo. En particular, se desarrolla una estrategia computacional que utiliza nuestra herramienta Mapa de caminos más corto (MCC) basada en la correlación para el diseño de variantes enzimáticas autónomas de TrpB.

Finalmente, el **Capítulo 6** incluye una breve discusión de los principales resultados presentados en los **Capítulos 4 y 5**, y las principales conclusiones de esta tesis se resumen en el **Capítulo 7**. Esta tesis proporciona información útil para abordar la evaluación computacional de la enantioselectividad y las propiedades enzimáticas alostericas para investigar los efectos inducidos por mutaciones en el paisaje energético conformacional de las enzimas. Los estudios recogidos en esta tesis enfatizan la relevancia de considerar la dinámica conformacional de las enzimas en los procesos de ingeniería computacional de enzimas.

Chapter 1. Introduction

Life and enzymes

The emergence of biological organisms took place roughly 3.8 billion years ago. Life can be defined as the ability of an isolated entity (uni- or pluricellular) to auto-replicate and evolve along time. As we know life on Earth, the following different key components are required:

- **Code:** A simple code based on the combination of 4 different molecules (DNA) encodes all the information needed for all cell functions (replication, transcription, translation, regulation...). The protection of the code is pivotal for survival. Cells have many mechanisms to guarantee the code is transferred to the next generation in good conditions.
- **Catalysts:** The role of catalysts is also crucial. Biocatalysts (enzymes) accelerate selectively chemical reactions operating under physiological conditions in time scales compatible with cell demands (i.e. useful velocities).
- **Energy currency:** An organic compound, Adenosine Tri-Phosphate (ATP) allowing for energy transfers to drive cell processes requests.
- **Energy storage:** Disposal of large energy reserves in chemical form (e.g. glycogen or lipids).
- **Evolution:** Since all the information is in the code, the code has to undergo changes in order to evolve. The Darwinian evolutionary theory states that those changes are random and only those that confer good properties for survival are kept. Thus, living entities have to die to promote enhancement of the next generations.

This thesis focusses on the study of biocatalysts. Enzymes are vastly the combination of only 20 different amino-acids. These 20 building blocks are enough to create the most sophisticated biological machinery on earth. Billions of years of evolution from ancient life evolved enzymes to catalyze thousands of chemical reactions that allow living organisms to obtain energy from nutrients (catabolic pathways), to store the energy obtained in chemical form, to use the energy stored to synthesize macromolecules from small metabolites (anabolic pathways) and to produce an enormous array of biologically active molecules (secondary metabolism) for multiple functions such as metabolic precise control (hormones), electric signaling (neurotransmitters), defense against other organisms (drugs), among others. Considering their intrinsic natural power, it was a matter of time before we were able to extract

them from their living organisms and make them work for human purposes. It happened only a few centuries afterwards the initial steps in the scientific revolution thanks to the marriage between science and imperium. At the end of the 17th century, studies about the meat digestion by the stomach extracts described for first time the biocatalysts. In 1897 Eduard Buchner postulated that the molecules involved in sugar fermentation can work separated from the living cells. This was the end of the vitalism theories. Frederick W. Kühne was the first to name the molecules detected by Buchner enzymes.^[1] In the beginning of the 20th century, the isolation and crystallization of digestive enzymes by James Summer, John H. Northrop and Moses Kunitz ended the debate of the nature of the biocatalysts concluding that enzymes are proteins.^[2] Another important achievement was carried out by Leonor Michaelis and Maud Mentel developing the Michaelis-Mentel equation,^[3] which allowed the kinetic characterization of enzymes. Afterwards in the 50s, John Kendrew had successfully resolved the first X-ray structure of a protein (myoglobin)^[4] and James Watson alongside Francis Crick discovered the DNA structure,^[5] which was crucial for the development of genetic engineering. We are currently able to purify proteins, solve its 3D structure, characterize its kinetic parameters and generate many mutant libraries with the activity of interest for many enzymes. However, there are still many open questions behind their mode of action with no clear answer, which frustrates our efforts for engineering them.

1.1 Kinetic view

The oldest method to study enzymatic reaction mechanisms is the determination of the enzymatic reaction velocity and the mode it changes in different experimental conditions. In this section, a brief overview of enzyme kinetics provides a meaningful insight into the main kinetic parameters used to characterize enzyme catalytic efficiency.

1.1.1 Some basic kinetic concepts

The reaction rate (V) for a given chemical reaction is the velocity at which the reactants are converted into products. As for instance, A and B are converted into C in the reaction: $2A + B \rightarrow C$. The reaction rate can be related to the concentration of the reactants through the rate constant in the so-called rate law. Thus, the rate constant is a proportionally constant specific for each reaction. For the mentioned reaction, the rate law could be $k[A][B]$, where $[A]$ and $[B]$ express the concentration of the reactants A and B and their exponents corresponds to the

partial orders of the reaction (i.e. 1st order for both, A and B reactants). The overall reaction order is the sum of the partial orders for all reactants (i.e. for this case 2nd order). Note that the partial orders of reaction are not necessarily the same values that the stoichiometric coefficients, and in some cases some reactants may not appear in the rate law. For instance, if reactant B exhibits 0th order and reactant A 2nd order; the rate law is $k[A]^2$. As the rate constants, the reaction order is determined experimentally. A reactant that exhibits no dependence of the velocity respect to its concentration corresponds to a 0th order while a linear and quadratic dependence corresponds to a 1st and 2nd reaction order, respectively. The units of the rate constants depend on the reaction order as follows:

$$\text{0th order } V = k; k = \text{M} \cdot \text{s}^{-1}$$

$$\text{1st order } V = k [A]; k = \text{s}^{-1}$$

$$\text{2nd order } V = k [A]^2; k = \text{M}^{-1} \cdot \text{s}^{-1}$$

Regarding the following first order reaction:



The reaction rates of the forward and reverse reactions correspond to $k_f [\text{S}]$ and $k_r [\text{P}]$ respectively. When a chemical reaction reaches the chemical equilibrium, the forward and reverse rates become equal; $k_f [\text{S}] = k_r [\text{P}]$. Thus, for the cases where the stoichiometric coefficients are equal to the order of the reactants, the equilibrium constant (K_{eq}) can be related to the rate constants of a chemical step as:

$$K_{eq} = \frac{k_f}{k_r} = \frac{[\text{P}]}{[\text{S}]} \quad (1.2)$$

Another important statement regarding the rate constant is found in the Arrhenius equation, that is a remarkable expression that expresses the magnitude of the rate constant as a function of the temperature:

$$k(T) = A e^{\frac{-E_a}{RT}} \quad (1.3)$$

Where A is a pre-exponential factor that indicates the frequency of collisions, E_a is the activation energy, R the gas constant and T the absolute temperature.

1.1.2 Michaelis-Menten equation

The first scientist to propose that enzymes form a binary complex with the substrate was the English chemist Adrian Brown. This idea was later supported by the French chemist Victor Henry (1903), and finally Adrian Brown's concept was further studied by the German physical chemist Leonor Michaelis and his Canadian associate Maude Menten to develop the famous Michaelis-Menten equation (1913).^[3] To simplify enzyme kinetics they assumed several approximations:^[1, 6]

- Ignoring the reverse reaction by measuring initial velocities (i.e. collecting the data after *ca.* the first 60 seconds or less, when only a few percent of the product is formed).
- Negligible enzyme concentrations compared with that of the substrate (i.e. [E] of nano-molar order while [S] can be five or six orders of magnitude more).
- Postulation that in solution E forms a rapid and reversible equilibrium with S forming the ES complex, such complex is decomposed and the product is released in a slow second step of first-order rate constant (k_{cat}) (see reaction of Equation 1.4)



According to this simple enzyme reaction showed in Equation 1.4:

$$K_s = \frac{[\text{E}][\text{S}]}{[\text{ES}]} \quad (1.5)$$

and

$$V_0 = k_{\text{cat}} [\text{ES}] \quad (1.6)$$

It is also assumed that the initial or total enzyme concentration remains constant and is equal to the sum of the free and the complex form:

$$[\text{E}]_0 = [\text{E}] + [\text{ES}] \quad (1.7)$$

Therefore, if the second step is rate-limiting the global velocity is proportional to [ES] complex. The dependence of the initial velocities at different increments of [S] shows a rectangular hyperbola curve in most enzymes (see **Fig. 1.1**). At low [S], V_0 increases linearly

with [S]. In this scenario the majority of the enzyme is in the free E form. Thus, the velocity become dependent on [S] as the increments of [S] push the equilibrium towards the formation of more [ES]. However, at sufficiently high [S], V_0 tends towards a limiting value (i.e. V_{\max}) because now most of the enzyme is in the [ES] complex form. At this point, the enzyme is saturated in the ES form and further increments of [S] will not affect the velocity.

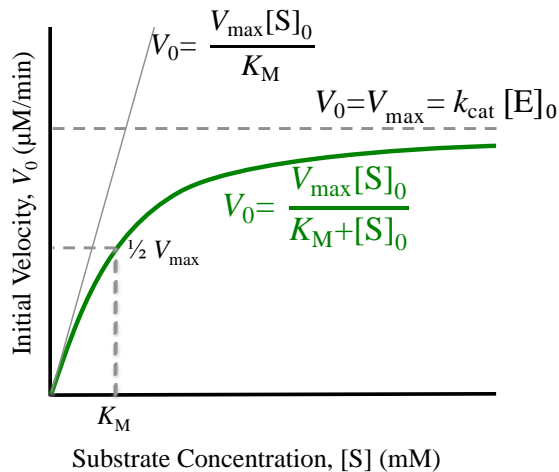


Figure 1.1 Substrate concentration effect on the initial velocity for an enzymatic reaction obeying Michaelis-Menten kinetics.

By combining Equations 1.5-7:

$$V_0 = \frac{[E]_0 [S] k_{\text{cat}}}{K_s + [S]} \quad (1.8)$$

Since the maximum velocity will be reached when $[ES] = [E]_0$ (i.e. saturation), V_{\max} can be defined as $k_{\text{cat}}[E]_0$, by substituting it in Equation 1.8, we finally obtain the Michaelis-Menten equation:

$$V_0 = \frac{V_{\max} [S]}{K_M + [S]} \quad (1.9)$$

Note that for the original Michaelis-Menten approach, $K_s = K_M$

1.1.3 Steady-state theory

In 1925 G. E. Briggs and J. B. S Haldane debated the Michaelis-Menten mechanism.^[7] The assumption of the rapid equilibrium works for many enzymes. However, such a condition is only valid when $k_{\text{cat}} \ll k_{-1}$



If k_{cat} is large enough, the equilibrium between E and S to form ES is not reached because ES is decomposed faster into E and P. The solution they proposed for such enzyme mechanisms is the steady state approach. When the enzyme is mixed with the substrate, the reaction rate increases exponentially in a pre-steady state period, which occurs at the milliseconds time-scale. This situation quickly evolves to a steady state where [ES] remains constant (i.e. enzyme velocity rate constant). The steady state scenario is maintained as long as [S] remains constant. Once the reaction evolves and [S] starts to be consumed the enzyme rate also decreases (**Fig 1.2**). Hence, by measuring initial velocities the steady-state approach is accomplished.

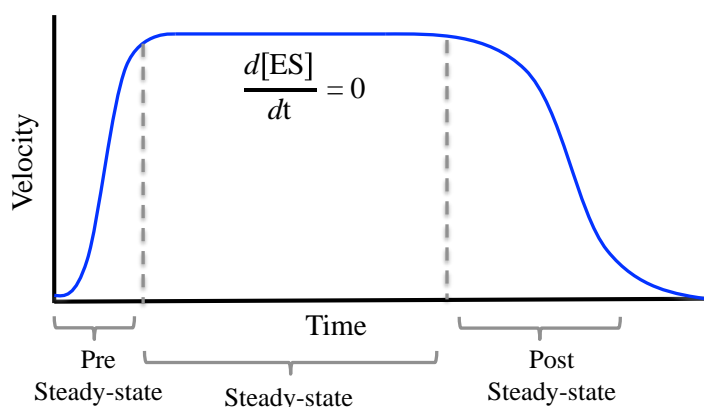


Figure 1.2 Representation of the enzyme velocity along time. The steady-state phases are also showed.

Briggs-Hadane kinetics applied the steady state to [ES], thus the rate for ES formation is equal to the rate of ES decomposition:

$$\frac{d[\text{ES}]}{dt} = 0 \quad (1.11)$$

$$\text{ES formation rate} = k_1([E]_0 - [ES])[S] \quad (1.12)$$

$$\text{ES decomposition rate} = k_{-1}[ES] + k_{\text{cat}} [ES] \quad (1.13)$$

$$k_1([E]_0 - [ES])/[S] = k_{-1}[ES] + k_{\text{cat}} [ES] \quad (1.14)$$

After realizing a set of algebraic steps over Equation 1.14, we obtain Michaelis-Menten equation again:

$$V_0 = \frac{V_{\text{max}}[S]}{[S] + (k_{\text{cat}} + k_{-1})/k_1} \quad (1.15)$$

Note that for the steady state approach:

$$K_M = (k_{\text{cat}} + k_{-1})/k_1 \quad (1.16)$$

Since K_s is equal to k_{-1}/k_1 , we have:

$$K_M = K_s + \frac{k_{\text{cat}}}{k_1} \quad (1.17)$$

As expected when $k_{-1} \gg k_{\text{cat}}$ Equation 1.16 simplifies to $K_M = K_s$ again.

1.1.4 Multi-step enzyme cycles

The enzyme mechanisms explained up to now far represent the sophisticated enzyme cycles. Often several intermediates are formed, covalently bound or non-covalently bound along the catalytic itinerary. In this regard k_{cat} and K_M are often a combination of various rate and equilibrium constants. This is the case for the following mechanism, which involves two different intermediates in the cycle:



Applying the steady-state assumption, the reader can verify that for this mechanism the Michaelis-Menten equation developed can be shown as:

$$V_0 = [E]_0 [S] \left\{ \frac{k_2 k_3 / (k_2 + k_3)}{K_S k_3 / (k_2 + k_3) + [S]} \right\} \quad (1.19)$$

In which:

$$K_M = K_S \frac{k_3}{k_2 + k_3} \quad (1.20)$$

and

$$k_{\text{cat}} = \frac{k_2 k_3}{k_2 + k_3} \quad (1.21)$$

Then in the cases that K_M is not equal to K_S :

$$K_M = \frac{[E][S]}{\sum [EX]} \quad (1.22)$$

Where $[EX]$ is the sum of all bound enzyme species. In some cases, an Enzyme-Intermediate complex (e.g. EX_2) can have a larger contribution to the K_M value than the $[ES]$, e.g. EX_1 according to reaction of Equation 1.18. Following the same approach, k_{cat} is a function of all first-order rate constants after ES formation, and with no prior information it cannot be assigned to any particular process. Any of the catalytic steps after substrate binding can be the rate-limiting step (RLS), and in some enzymes (e.g. dehydrogenases) even the product release (i.e. EP dissociation rate constant) has been found to be the RLS. Unraveling the step or combination of steps that contribute most to the reaction rate limiting, leads to useful information in order to improve enzyme efficiency. In general terms, k_{cat} is a first-order rate constant called the turnover number. It represents the maximum number of substrate molecules converted to products per active site and per unit time, or the number of times the enzyme “turns over” per unit time. On the other hand, in all cases K_M is the substrate concentration at which $V = V_{\text{max}}/2$.^[6] k_{cat}/K_M quantities are widely used to refer to the enzyme specificity and catalytic efficiency.

Note that when $[S] \ll K_M$, Michaelis-Menten equation can be expressed as:

$$V = [E][S] \frac{k_{\text{cat}}}{K_M} \quad (1.23)$$

Therefore k_{cat}/K_M turns out to be a second order rate constant that refers to the substrate and free enzyme reaction. In other words, it measures the efficiency of substrate capture by an enzyme. This is the reason why it is called the “specificity constant”. The k_{cat}/K_M parameter usually only includes the rate constants involved up to the first irreversible step of the reaction mechanism.^[8] According to the reaction mechanism of Equation 1.18 and, operating:

$$\frac{k_{\text{cat}}}{K_M} = \frac{k_1 k_2}{(k_{-1} + k_2)} \quad (1.24)$$

Notice that k_3 , which only applies in the second step of the reaction shown in Equation 1.18 is not included in the specificity constant. Thus, confirming that the k_{cat}/K_M parameter only includes rate constants up to the first irreversible step.

1.1.5 Microscopic rate constants

According to the microscopic association rate constants calculations (e.g. k_1 values), the diffusion-controlled encounter frequency of an enzyme and its substrate should be about $10^9 \text{ s}^{-1} \text{ M}^{-1}$. Most of enzymes have k_1 values in the range of 10^6 to $10^8 \text{ s}^{-1} \text{ M}^{-1}$. Enzymes as Lactate Dehydrogenases can reach values close to the diffusion controlled ($k_1 \approx 10^9$) while others as α -Chymotrypsin are far away ($k_1 = 3.4 \times 10^3$).^[6] k_{cat}/K_M values roughly correspond to k_1 values. However, the k_{cat}/K_M value is always less than k_1 , and only in particular cases k_{cat}/K_M can be approximately equal to rate constant k_1 . Enzymes with high k_{cat}/K_M values indicate that the rate limiting step for this parameter is close to the diffusion-controlled encounter of the enzyme and the substrate.

Experimentally, k_{cat} and K_M can be obtained through Michaelis-Menten curve by measuring the initial velocities in the steady-state conditions. However, in order to detect the transient intermediates formation and obtain the rate constants of the individual enzyme steps, it is necessary to measure the rate in the pre-steady state time domain (**Fig. 1.2**). To that end, rapid mixing techniques are required.

1.2 Classical thermodynamic view

Thermodynamics is that part of science that studies the conversion of different forms of energy (e.g. thermal, mechanic, electric) that takes place in natural processes. Work and heat are energy quantities that can be understood as modes of energy transfer. In thermodynamics, work covers a wide range of processes including mechanical work, electric work, surface work, magnetization work, etc. Heat is the energy transferred between two bodies at different temperatures. Heat (q) and work (w) are properties that depend on the trajectory of the process, while temperature (T), pressure (P), and volume (V) are state functions because their value only depends on the current state of the system and not on the history from which such state was reached. In other words, we cannot answer how much heat is transferred to 100 g of water in order to increase its temperature from 10 to 50 C° because we do not have information about the history of the process. In this case, the energy transferred to the water may come from heat but also from a different source such as mechanical work performed by a magnetic bar that produces friction.^[9] If all the energy were transferred from the mechanical work the heat transfer would be zero. In this context, a protein conformational exchange from state A to state B may occur in many ways and the underlying process of the exchange history is very useful to tune this bio-molecular processes. The idea of this section is to briefly explain the classical thermodynamic laws together with some knowledge about biological work in order to situate enzymes as pure thermodynamic machineries governed by enthalpy and entropy components.

1.2.1 First law of thermodynamics

According to the first law of thermodynamics, the total energy of the universe remains constant. In this context, energy can be converted from a particular form to another and transferred from a system to the surroundings and vice versa as long as the total energy does not change. Thus, the energy cannot be created or destroyed. In general, the total energy of the universe can be represented as:

$$E_{universe} = E_{system} + E_{surroundings} \quad (1.25)$$

And the energy variation as:

$$\Delta E_{universe} = \Delta E_{system} + \Delta E_{surroundings} = 0 \quad (1.26)$$

According to classical thermodynamics, the total energy of a system can be estimated as the internal energy (U) plus the kinetic and potential energies. Assuming that the system is in resting state and in absence of external fields (e.g. electric or magnetic) the total energy can be estimated only as the internal energy due to the absence of kinetic and potential energies. This is indeed the case for most of the bio-molecular processes and cases that will be covered here.

The internal energy of a particular system (e.g. a protein) includes many types of energies:

- Translational
- Rotational
- Vibrational
- Electronic
- Nuclear
- Intermolecular interactions

The first law of thermodynamics can be expressed as:

$$\Delta U = q + w \quad (1.27)$$

This indicates that a variation in internal energy (ΔU) of the system for a particular process can be calculated by the sum of the heat (q) exchange between the system and the surroundings and the work (w) executed on (or by) the system. Thus, there is a decrease in the internal energy of the system when the system performs work on the surroundings and when the heat is absorbed by the surroundings from the system. Accordingly, there is a gain in internal energy in the opposite scenario. Work has different meanings depending on its nature, as for instance mechanic, electric and expansion work.

Heat can be expressed as:

$$q = mC_e\Delta T \quad (1.28)$$

Where the specific heat capacity (C_e) is the amount of heat needed to increase 1 gram of a particular substance by 1 C° of temperature.

Another widely used thermodynamic term is the enthalpy (H). It is expressed as:

$$H = U + PV \quad (1.29)$$

Where P and V are the pressure and the volume respectively. In contrast to the internal energy (U), the enthalpy (H) is obtained by assuming constant pressure condition operating from Equation 1.27. The variation of H in a natural process can be expressed as:

$$\Delta H = \Delta U + P\Delta V \quad (1.30)$$

Notice that the difference between U and H becomes remarkable when there is expansion work during the natural process. For instance, in a chemical reaction when there is no expansion work, ΔH can be calculated by the heat generated, and according to Equation 1.30, $\Delta H = \Delta U$. However, for a chemical reaction where there is gas production, $\Delta H < \Delta U$ because part of the released internal energy is used to perform the gas expansion work. As a consequence, there is less heat released.

The standard enthalpy of reaction ($\Delta_r H^0$) is defined as the enthalpy change of a chemical reaction when 1 mol of reactants is converted into 1 mol of products in standard conditions (i.e. $P= 1\text{bar}$ and 298K). Thus, for combustion reactions (e.g. CO_2 formation) $\Delta_r H^0$ can be obtained experimentally by measuring the total heat generated in the reaction. Since the temperature during the combustion reaches much higher values than 298K , the heat released until the product is cooled up to 298K has to be considered as part of the reaction enthalpy.^[9]

$\Delta_r H^0$ can be also calculated from the summation of the molar standard enthalpy formation of the products minus the molar standard enthalpy formation of the reactants:

$$\Delta_r H^0 = \sum v \Delta_f \overline{H^0} (\text{products}) - \sum v \Delta_f \overline{H^0} (\text{reactants}) \quad (1.31)$$

The molar standard enthalpy of formation ($\Delta_f \overline{H^0}$) is the enthalpy change when 1 mol of compound is formed from the elements that constitute it (e.g. CO_2 formation from graphite and oxygen). Chemists handily assign arbitrary values of zero to the elements in their allotropic forms (e.g. graphite and oxygen). Intuitively $\Delta_r H^0$ obtained experimentally of the CO_2 combustion is equal to $\Delta_f \overline{H^0}(\text{CO}_2)$. However, many $\Delta_f \overline{H^0}$ cannot be obtained experimentally,

in these cases chemists solve this problem applying the laws developed by the German Henri Hess. The Hess laws permits to calculate the $\Delta_f \overline{H}^0$ of a reaction decomposing the target reaction in a set of reactions by which there are $\Delta_f \overline{H}^0$ values available and operating these set of chemical reactions as algebraic equations.

1.2.2 Second law of thermodynamics

The fundamental statements of the first and second law of thermodynamics took place in the mid-19th century.^[10] The first law allows the quantification of the energy exchange between the system and the surroundings and how energy can be converted between one form to another. However, the main limitation of the first law is that it cannot predict the direction of the energy exchange. At a particular set of conditions (e.g. temperature, pressure...) a natural process happens in one direction spontaneously. The unfolding of a thermophilic enzyme or the water boiling at 25 C° and 1 atm is as improbable as a ball rising spontaneously from the ground up to 1 meter. On the other hand, the unfolding of a mesophilic enzyme or the water boiling at 100 C° and 1 atm is as probable as the fall of a ball from 1 meter to the ground. In order to predict the direction of the spontaneous natural processes, a new thermodynamic function takes part: the entropy. In 1877, Ludwig Eduard Boltzmann established for first time the probabilistic basis of entropy.^[11] According to the Boltzmann equation, the entropy (S) is defined as:

$$S = k_B \ln W \quad (1.32)$$

Where W is the probability (“Wahrscheinlichkeit” in German) that a natural process occurs and k_B is the Boltzmann constant, $1.381 \times 10^{-23} \text{ JK}^{-1}$. Thus, a change in entropy only depends on the probability of the natural event changing from state 1 to state 2.

$$\Delta S = S_2 - S_1 = k_B \ln W \quad (1.33)$$

$$\Delta S = S_2 - S_1 = k_B \ln \frac{W_2}{W_1} \quad (1.34)$$

The equilibrium state is the most probable situation for an isolated system. The probability is proportional to the number of possible microstates. In this context, a system can be described macroscopically (e.g. P, V, T, U, N) and microscopically (e.g. position and velocity of each

atom). There are many microscopic states compatible with the macroscopic state of a system. W can be interpreted as the number of microstates to distribute the particles among the different energy levels compatible with the macroscopic values. Therefore, an increase in entropy is associated with an increase in the number of microstates (W), and as a consequence to an increase in the disorder of the system. The statistical thermodynamic approach will be further explained in **Chapter 1.3** focusing on the enzyme microscopic properties.

In general, Equation 1.34 is not used to calculate entropy changes for experimental purposes because the calculation of W in complex natural processes as a chemical reaction is very complicated. Instead ΔS can be easily calculated from other energy quantities as ΔH according to equation:

$$\Delta S = \frac{q_{rev}}{T} = \frac{\Delta H}{T} \quad (1.35)$$

Equation 1.35 is the thermodynamic definition of entropy.^[10] It can be estimated operating from the first law equation ($q = -w$) and developing Equation 1.34 for the case of an ideal gas expansion. According to the thermodynamic definition of entropy, the change in entropy of a system in a reversible process is determined by the heat absorbed divided by the temperature. It is noteworthy to say that this definition is only valid for a reversible process (i.e. q_{rev}). Although the entropy is a state function, the heat not. Thus, the reversible trajectory of the process has to be specified.

According to the second law of thermodynamics:

$$\Delta S_{universe} = \Delta S_{system} + \Delta S_{surroundings} \geq 0 \quad (1.36)$$

The second law of thermodynamics means that the $\Delta S_{universe}$ never decreases and becomes positive for an irreversible process (i.e. spontaneous) while is 0 for a reversible process (i.e. in equilibrium conditions). The entropy of the universe is continuously increasing and tends to a maximum value.

When the temperature of a system increases, its entropy also increases. This correlation is due to the energy input, which boosts the molecules to higher energy levels of translational, rotational and vibrational energies. As a consequence, the disorder increases at molecular level.

Since the heat from the surroundings is transferred reversibly to the system increasing the entropy infinitesimally at a constant temperature:

$$dS = \frac{dq_{rev}}{T} \quad (1.37)$$

Operating Equation 1.37 and considering constant pressure ($q_{rev} = \Delta H$), the increments in entropy as a result of the heating can be defined as:

$$\Delta S = C_P \ln \frac{T_2}{T_1} \quad (1.38)$$

Where C_p is the heat capacity, which is assumed to be independent of the temperature for a small range of temperatures.

Additionally, phase changes can also affect the entropy. For instance, at 100 C° and 1 atm, water is in equilibrium between liquid and gas phases. The change in entropy associated with the conversion to gas phase and considering constant pressure can be expressed as:

$$\Delta_{vap}S = \frac{\Delta_{vap}H}{T_b} \quad (1.39)$$

Where $\Delta_{vap}H$ and T_b are the enthalpy of evaporation and the temperature of the boiling point, respectively. Since the system is in equilibrium conditions (i.e. $\Delta S_{universe} = 0$), the gain in entropy by the system has to be equal to the loss of entropy by the surroundings due to the heat transferred to the system. On the other hand, the water does not boil at 25 C° and 1 atm because the loss of entropy by the surroundings is much higher than the gain of entropy by the system, and then $\Delta S_{universe} \ll 0$. ΔS_{system} and $\Delta S_{surroundings}$ can be calculated by imagining a set of reversible steps. For instance, the ΔS_{system} and, the $\Delta S_{surroundings}$ of the water boiling process at 25 C° and 1 atm can be calculated by dividing the process in three reversible steps (i.e. (1) heating the water from 25 C° to 100 C°, (2) phase change from liquid to gas at 100 C° (3) cooling the water steam from 100 C° to 25 C° using the equations previously detailed.

1.2.3 Third law of thermodynamics

The third law of thermodynamics was developed by chemist Walter Nernst in the beginning of the 20th century.^[12] The third law establishes that every substance has a finite positive entropy, but at zero absolute of temperature, the entropy can reach zero. This happens in the case of a pure perfect crystalline substance. Mathematically the third law can be expressed as:

$$\lim_{T \rightarrow 0} S = 0 \quad (1.40)$$

According to Boltzmann's definition of entropy (Equation 1.32), a hypothetic pure crystalline substance can only have one microstate. For this case, the crystal only has a particular arrangement of the atoms, as a consequence $W = 1$ and

$$S = k_B \ln W = k_B \ln 1 = 0 \quad (1.41)$$

The third law allows to estimate the entropy of a substance at a particular temperature as follows:

$$S_T = \int_0^T \frac{C_P}{T} dT = \int_0^T C_P d \ln T \quad (1.42)$$

Then, it is feasible to calculate the absolute entropy of a mol of substance at 298 K and 1 atm (i.e. standard molar entropy $\overline{S^0}$) by summing the amount of entropy accumulated from 0 K to 298 K. It can be done by using Equation 1.42 and also considering the entropy variations associated with the phase transitions (e.g. Equation 1.39). Subsequently, the reaction standard molar entropy ($\Delta_r S^0$) can be calculated from the standard molar entropy of the products and reactants.

$$\Delta_r S^0 = \sum v \overline{S^0} (\text{products}) - \sum v \overline{S^0} (\text{reactants}) \quad (1.43)$$

Altogether indicates that in the thermodynamic equilibrium, a particular system (e.g. an enzyme) tends to the maximum degree of disorder that the energy provided by the surroundings permits at a given temperature. A common enzyme at 298 K does not unfold (i.e. increasing its entropy to a higher level) because the heat taken from the surroundings would decrease the

entropy of the surroundings more than the gain of entropy associated with the enzyme unfolding. However, at sufficient high temperatures the entropic balance would lead the enzyme to unfold. Thanks to the thermodynamic equilibrium enzymes can efficiently explore the set of catalytic states needed for the enzyme cycle.

The temperature oscillations along the day are due to the earth translation around the sun, which is indeed the source of energy that spreads entropy in every corner. All biological entities hope for the daily energy intake of sun in order to keep working for the organisms they take part in. Additionally, biological organisms conspire against the second law of thermodynamics by building large sophisticated biological structures (lower in entropy) from small disordered molecules (higher in entropy). As for instance, the biosynthesis of proteins from amino-acids. Organisms can do that thanks to one of the most fascinating events in biological evolution, the energy “currency” in chemical form. All organisms use the Adenosine Tri-Phosphate (ATP) molecule as energy currency. In ATP higher in energy (i.e. rich in energy) phosphate bonds store large energy quantities. The energy released from the hydrolysis of ATP can be used to drive many non-spontaneous biological processes, as for instance to drive secondary metabolism pathways to yield active compounds, synthesis of biological structures, nerve impulse propagation and muscle contraction (human work).

Previous to the industrial revolution, humans lacked knowledge to convert one type of energy in another one. In essence humans and animals were the only energy conversion device available. Thus, muscle power (i.e. mechanical energy) was the key to almost all human activities. The source of energy to activate these organic muscle machines comes in the long term from the sun. Plants capture the solar energy through the photosynthesis and store it in chemical energy (i.e. food). Food is eventually converted in muscle power though the ATP generated in the cellular respiration. After industrial revolution, other energy conversion devices arose, as the steam machine, which impacted dramatically our mode of life. Since then, the energy conversion devices have been in constant development by humans. Nevertheless, the way we are coping with this development is debatable.

The sun delivers 3.766.800 exajoules per year to our planet. All plants capture only 3000 solar exajoules though the photosynthesis. All human industries and activities consume only around 500 exajoules per year, which is equivalent to the solar energy delivered by the sun in 90 minutes. And this is only solar energy. We are surrounded by other enormous sources of

energy as nuclear or gravitational.^[13] It is shocking we are not taking advantage of such massive amounts of energy. At the same time in chemical industries, we apply the concept of energy efficiently; where the energy exchanges networks operate avoiding to waste an infinitesimal piece of energy. The efficiency is so high that it seems art. It is evident we do not lack energy sources but the will and the knowledge to transform it for human purposes (e.g. food and transport).

1.2.4 Gibbs energy

Back to physics: the main limitation of the second law (Equation 1.36) is that ΔS of the system and the surroundings has to be calculated to estimate the spontaneity of a process. However, we are mostly interested in the system. Due to this fact the American physicist Josiah Gibbs developed a novel state function (G).^[14] Operating Equation 1.36 only as a function of the system and assuming constant pressure and temperature, the famous Gibbs energy is obtained:

$$G = H - TS \quad (1.44)$$

Where H and S are the enthalpy and the entropy of the system, respectively. G variations can be applied as criteria of equilibrium and spontaneity:

$$\Delta G_{system} = \Delta H_{system} - T\Delta S_{system} \quad (1.45)$$

ΔG_{system} calculation (from now on ΔG) permits to quantify the energy exchange of a spontaneous process until it reaches the thermodynamic equilibrium ($\Delta G = 0$). In a spontaneous process, the system delivers energy ($\Delta G < 0$). On the contrary a non-spontaneous process has to be forced applying energy to the system ($\Delta G > 0$). Indeed, the energy needed to force a process from A to B (e.g. unfolding of a thermophilic enzyme at 25 degrees) is the same as the energy that would be delivered in the inverse (spontaneous) process from B to A. We are forcing many non-spontaneous processes on a daily basis such as boiling water to make a cup of tea.

Equation 1.45 is divided in two main components, the enthalpy (ΔH) and the entropy ($-T\Delta S$) contributions. If $|\Delta H| \gg |T\Delta S|$ the process is enthalpy-driven and if $|\Delta H| \ll |T\Delta S|$ entropy-driven. The temperature determines the relative contribution of ΔH and ΔS . At high

temperatures the impact of ΔS is higher than ΔH . The enthalpic component in a chemical reaction can be obtained as the energy exchange as a consequence of the balance between the chemical bonds broken and the bonds formed, whereas the entropic component as the entropy gained or lost in the reaction step (e.g. gain of entropy in gas formation and/or increase in the moles of the product reaction side). The molar standard free energy variations for a chemical reaction (1 bar and 298K) can be calculated from the molar standard free energy formation of its reactants and products:

$$\Delta_r G^0 = \sum v \Delta_f G^0 (\text{products}) - \sum v \Delta_f G^0 (\text{reactants}) \quad (1.46)$$

In similarity with $\Delta_f \overline{H^0}$, an arbitrary value of zero is assigned to the $\Delta_f G^0$ of the elements in their allotropic forms. Accordingly, $\Delta_r G^0 = \Delta_f G^0$ for the formation reaction of a given compound from its elements in their allotropic forms (e.g. CO₂ formation). $\Delta_r G^0$ can be calculated from $\Delta_r H^0$ and $\Delta_r S^0$ values obtained in Equations 1.31 and 1.43 respectively, using them in the following Gibbs equation:

$$\Delta_r G^0 = \Delta_r H^0 - T \Delta_r S^0 \quad (1.47)$$

Physicochemists define the standard state of an ideal solution when all reactants and products are at 1M concentration, at 1 bar pressure and 298 K temperature ($\Delta_r G^0$). Thus, the free energy change for the (A +B → C +D) reaction is given by:

$$\Delta_r G = \Delta_r G^0 + RT \ln \frac{([C/1M])([D/1M])}{([A/1M])([B/1M])} \quad (1.48)$$

Where R is ideal gas constant. Accordingly, when all reactants are in 1 M concentration $\Delta_r G = \Delta_r G^0$. Thus, $\Delta_r G^0$ can be interpreted as the free energy variation of 1 mol of reactants evolving to products or vice versa until the equilibrium is reached from standard conditions assigned as initial state. $\Delta_r G^0 \ll 0$ indicates the reaction would evolve spontaneously towards products formation and $\Delta_r G^0 \gg 0$ indicates the reaction would evolve spontaneously towards reactants formation. When a reaction reaches the chemical equilibrium:

$$0 = \Delta_r G^0 + RT \ln K_{eq} \quad (1.49)$$

$$\Delta_r G^0 = -RT \ln K_{eq} \quad (1.50)$$

Where K_{eq} is the reaction ratio in the equilibrium and can be related with $\Delta_r G^0$ as follows:

$$K_{eq} = e^{-\frac{\Delta_r G^0}{RT}} \quad (1.51)$$

In this context $\Delta_r G^0$ can be interpreted as a constant that indicates not only the direction of the process but the driving force.

For the protein unfolding process (N → U):

$$\Delta_u G^0 = -RT \ln K_{eq} = -RT \ln \frac{U_{eq}}{N_{eq}} \quad (1.52)$$

Then $\Delta_u G^0$ can be estimated by determining the U/N ratio in the equilibrium. The U/N ratio of a protein can be for instance monitored experimentally through circular dichroism, tryptophan fluorescence or changes in tyrosine absorbance.^[15] The entropic $\Delta_u S^0$ and the enthalpic $\Delta_u H^0$ terms associated with the protein unfolding can be estimated experimentally by the differential scanning calorimetry (DSC) technique. DSC measures the heat supplied to the system at constant pressure by gradually scanning a range of temperatures over time.^[16]

1.2.5 The art of the biochemical work

Further operating Equation 1.45, using first and second law equations, assuming reversible trajectories and constant P and T, ΔG can be expressed as:

$$\Delta G = w_{rev, no PV} \quad (1.53)$$

Where ($w_{rev, no PV}$) is the maximum work including all types of work (e.g. electric, superficial...) except the expansion work. Thus, negative ΔG values permit to quantify the maximum quantity of useful work (energy) delivered by the system in a spontaneous process. On the other hand, positive ΔG values (non-spontaneous process) provide the minimum work

that has to be invested in the system to force the course of a non-spontaneous process. However, since the natural processes occur spontaneously (i.e. irreversibly) the real work that can be obtained is always lower. This is the case of energy obtained from the combustion of fossil fuel. The thermodynamic efficiency is low because the combustion is a highly irreversible process that occurs in one step, and not in infinitesimal steps (i.e. reversible process). In this context, most of the energy is released in the less efficient way (i.e. heat). The heat generated by the combustion reaction can be coupled to a machine heat (e.g. thermal machine of Carnot) to convert the heat into mechanical work. However, this process is subjected to conversion energy limitations. On the other hand, there is a much more efficient way to obtain energy from a combustion reaction by performing the reaction into a fuel cell. In this case the reaction occurs in a more reversible way obtaining useful electric work. The electron flux of the reaction from the cell electrodes (anode to cathode) can be coupled to an electric motor to convert electric work into mechanical work. For instance, in the propane-oxygen fuel cell the efficiency that can be obtained is up to 70 %, which is roughly double the work obtained in an internal combustion engine. The electromotive force (emf) of a particular reaction can be estimated by the Nernst equation. For the $(A + B \rightarrow C + D)$ reaction it is given by:

$$E = E^0 - \frac{RT}{\nu F} \ln \frac{([C/1M])([D/1M])}{([A/1M])([B/1M])} \quad (1.54)$$

Where ν is the stoichiometric coefficient, F is the Faraday constant (i.e. the charge that carries 1 mol of electrons), E is the observed emf and E^0 is the standard emf of the cell (i.e. 298 K and all products and reactants at 1 M concentration). In the case of a reversible cell at a given temperature and pressure, $-\nu FE$ is the maximum work that can be obtained, which is indeed the decrease in Gibbs energy by the system:

$$\Delta_r G = -\nu FE = W_{electric, max} \quad (1.55)$$

Electrochemical measurements provide a more direct determination of $\Delta_r G$ (or $\Delta_r G^0$) of a process. The combustion of glucose in air is also a highly irreversible reaction. As a consequence, the energy is released in heat form in one step and the amount of energy is far from the maximum that could be obtained. But this is not the case when biological efficiency takes place. Organisms divide the combustion reaction into multiple steps aided by enzymes. In such a way the process becomes more reversible, and then a large number of energy currency

can be obtained (i.e. ATP synthesis from ADP and P_i). How cells convert the electric work obtained from the set of oxidations of glucose into chemical form is a fascinating event called oxidative phosphorylation. Most of the ATP (90 %) is obtained in the terminal respiratory chain through the oxidative phosphorylation. In this process, an electron flux from the reduced coenzymes formed in the Krebs cycle (NADH) lead its electrons to the oxygen through an ensemble of acceptors called electron transport chain, which split the redox reaction in several steps. As the electrons flux downstream along the electron transport chain (i.e. redox steps aided by enzymes), most of the free energy released by each redox reaction steps are used to expel hydrogen ions (H^+) from inside a compartment (e.g. mitochondrial matrix) towards another (e.g. mitochondrial intermembrane space) through the membrane that separates them. This process generates a chemical and electrical gradient difference since the $[H^+]$ increases in the mitochondrial intermembrane space, i.e. where they are being pumped into, which generates proton-motive-force (PMF). When the $[H^+]$ expelled flow back spontaneously to the mitochondrial matrix because of the gradient, the energy released is available to perform work; this $[H^+]$ gradient is analogous to the electric work that is performed in a battery. In the biological case, the $[H^+]$ flux is coupled to an enzyme called ATP synthase, which uses the free energy released of the transport of 4 H^+ in favor of a gradient to form one ATP molecule from ADP and P_i . Another interesting event is how the release of chemical energy can be used to perform mechanical work in the muscle. Myosin is an ATPase that uses the energy released from the ATP to perform a conformational change that triggers a power stroke from the myosin head on the thin muscle filaments leading to the contraction that produces human mechanical work.

We have seen so far how non-spontaneous processes can be performed by supplying the energy required to the system allowing living entities their daily duties. Nevertheless, there is a highly remarkable question to ask:

Why does the second law of thermodynamics immediately not reestablish the thermodynamic equilibrium by breaking the ordered macromolecules built and the ATP phosphate bonds, thus messing it all up?

The non-spontaneous events that organisms force, do not flow backwards immediately (spontaneous events) because of the fact that a process that is highly spontaneous does not mean it has to happen very fast. Fortunately, time-scale matter. A highly spontaneous process

can take long periods of time to occur (e.g. months). The ATP phosphate bond is high in energy and stores a considerable amount of energy. However, the phosphate bond does not break spontaneously releasing its energy because the ATP molecule is very stable.

We have seen how the oxidation of glucose is much more efficient in terms of useful work available when the reaction is divided in a set of steps aided by enzymes but the main role of enzymes is to accelerate the chemical reaction steps. Enzymes accelerate chemical reactions several orders of magnitudes allowing living organisms to dispose of biologically active compounds and energy quantities to drive biological processes in time-scales compatible with life. The disposal of energy is achieved accelerating for instance the combustion of glucose and the ATP hydrolysis reactions. The combustion of glucose in the absence of enzymes may take years, which can be easily tested by exposing a bag of sugar in presence of oxygen. Through the action of enzymes, we can dispose of the glucose chemical energy in seconds. This energy can be used to think, see and move. How enzymes operate to accelerate chemical reactions will be seen in **Chapter 1.4**.

1.3 Statistical thermodynamic view

Statistical thermodynamics is the discipline that links microscopic properties with the macroscopic properties of matter. A macrostate is a condition in which a particular value of many properties as P , V , T , n , H , S , G , U are assigned. A given macroscopic property (e.g. internal energy U) can be interpreted as a time-averaged quantity of an ensemble of microstates sampled over time $\langle U \rangle$. According to the ergodic hypothesis,^[17] if a system evolves over long periods of time, it passes through all accessible microstates in statistical equilibrium. In other words, the microstates forming an ensemble include all past and future microstates that can be explored, and are statistically distributed as relative populations based on its probability to be sampled. In this section, the importance of the microscopic view in the enzyme design field will be discussed. The underlying characteristics of the microscopic conformations of enzymes and its relative population distributions (i.e. conformational ensembles) is of high relevance to understand and tune enzyme activity.

1.3.1 Fundamentals of Energy landscapes

The energy landscape of a complex molecule, for example an enzyme, is a visual representation of the potential energy (U) as a function of the enzyme microstates or configurations (i.e. set of atomic Cartesian coordinates) so that any enzyme configuration leads to a potential energy value $U(x)$; see **Fig. 1.3**. Do not confuse the potential energy with the internal energy, which is a macroscopic property. The Boltzmann factor connects the probability of sampling a particular configuration (x) with its associated energy in a simple exponential equation:

$$pdf(x) \equiv p(x) = \frac{e^{\frac{-U(x)}{k_B T}}}{\int_{\nu} dx e^{\frac{-U(x)}{k_B T}}} \quad (1.56)$$

Where $pdf(x)$ is the probability density function as a function of x , which is equivalent to $p(x)$, k_B is the Boltzmann constant and T the absolute temperature.^[16] The numerator of the equation is the Boltzmann factor, which is a weighting factor proportional to the probability density. In order to obtain the exact probability density, the normalizing constant has to be present, which is the integral appearing in the denominator and represents the summation of all probability weights in the region of interest (ν); thus, the probabilities of all accessible microstates must add up to 1. The normalizing constant is also named the partition function, which is denoted by Z or Q . According to the energy and probability relationship, the probability of a microstate decreases as the energy increases. In other words, lower energy microstates are more likely. The temperature plays a key role in a given population distribution. The relative probability of all microstates becomes equal as the temperature increases. The physical meaning is that configurations that are unlikely at one temperature become more likely as T increases and as a consequence lower in energy configurations become less likely.^[16]

Consider a mol of enzyme in a box of water at 0 K. According to the third law of thermodynamics, only one enzyme configuration is likely. However, there are thousands of microstates super high in energy waiting for an increase of temperature to be sampled. As the temperature reaches 298K, the configuration that was extremely low in energy at 0 K, now is sharing its selfish relative population with another multitude of configurations the enzyme can adopt. At this point, the enzyme is mostly maintaining its native conformational ensemble being able to perform its catalytic itinerary. When going further to 393 K, another set of microstates starts to become likely. Those microstates correspond to unfolded configurations.

So that the native configurations become less likely due to the increase in probability of these new unfolded microstates that were high in energy at the previous temperature. At this stage, one might think that to fold an enzyme at 393K requires free energy, so work is needed. At extreme temperature conditions, the entropy becomes so powerful thousands of microstates where enzymes are fragmented (i.e. chemical bonds are broken) are sampled. It generates a massive number of microstates that tend to be equally likely, and of course the initial super lower in energy coordinates of the configuration at 0 K is also sampled, although not in the same form (i.e. now it is completely fragmented) and not with the same probability (i.e. now it is only an infinitesimally relative population of the whole thing).

Back to the Boltzmann factor, it is worth mentioning that $U(x)$ is not the only energy term that contributes to the total energy. The total energy (E) is indeed the sum of the potential energy (U) and the kinetic energy (KE):

$$E = U(x) + KE(v) = U(x) + (m/2)v^2 \quad (1.57)$$

Where m is the mass and v the velocity of a given particle. In general, kinetic energy is the energy that a particle possesses due to its motion, while potential energy is the energy associated with the particle position, which is subjected to forces. Interestingly, 400 years ago Frances Bacon stated that heat is motion. Kinetic energy makes enzymes move away from the minimum in energy landscapes, while the restoring force due to the potential energy landscape moves them back toward the minimum in energy.

For simplicity, the kinetic energy (KE) is excluded from the total energy (E). Notice that the full Boltzmann factor is factorizable:

$$e^{-E/k_B T} = e^{-\frac{U(x)}{k_B T}} e^{-\frac{(\frac{m}{2})v^2}{k_B T}} \quad (1.58)$$

Therefore, the first factor depends only on the variable x and the second only on v , which means that the variables are statistically independent. The consequence is that the distribution of velocities does not affect the distribution of positions, and Equation 1.56 is indeed correct and **Fig. 1.3** a realistic representation of the energy landscape.^[16]

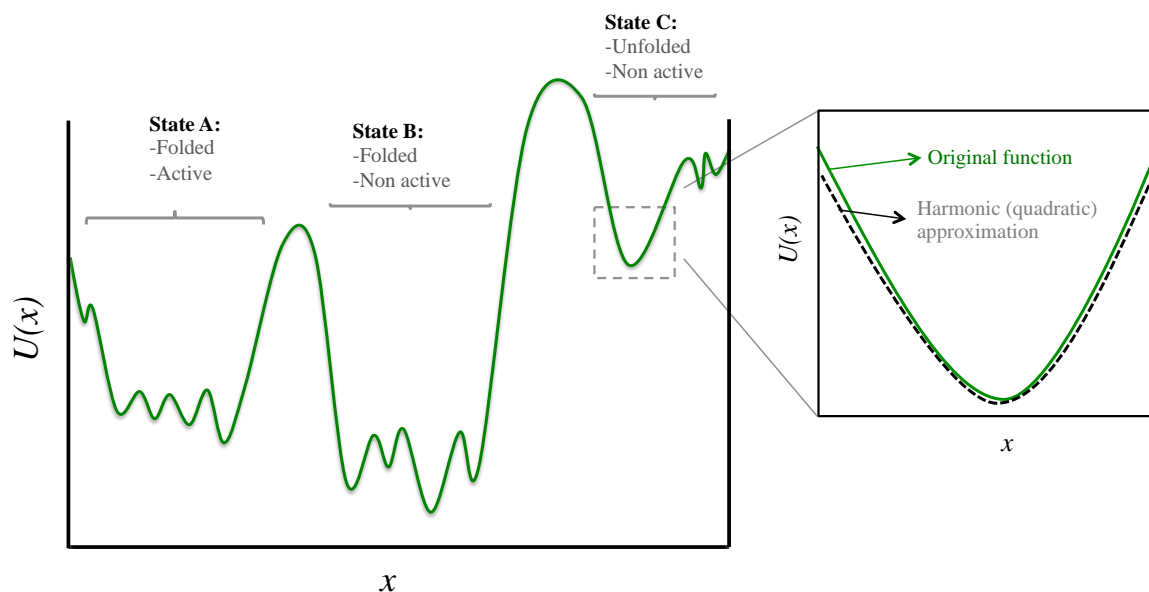


Figure 1.3 On the left, one dimensional energy landscape exhibiting different energy basins and energy barriers. The set of energy-basins are grouped into three different states (A, B, and C) according to activity and folding enzyme descriptors. On the right, single energy basin representation from the original function (solid green line) together with the harmonic approximation (dashed black line).

The energy landscape of **Fig. 1.3** is simplified in three different states (A, B, and C). There is some controversy about the meaning of state. Strictly a state can be defined as a group of configurations belonging to a single-well or single basin. However, in a more practical way, a state is usually referred to a group of similar energy basins. Accordingly, one can define as many states as appropriate to describe a particular process under study. For instance, if we aim to study the enzyme unfolding process, states A and B of **Fig. 1.3** should be compiled into the folded state and state C would be assigned as the unfolded state. In contrast, if we aim to estimate catalytic activity (i.e. active/inactive population distribution) states B and C should be grouped as non-catalytic and state A as catalytic.

Regarding the potential energy function, a particular enzyme configuration (i.e. specific set of coordinates) has associated with it a potential energy value. However, a state consists of multiple configurations, each one with its own set of coordinates, and hence its own potential energy value. In this context the probability density of a state (e.g. state A) can be estimated by adding up (i.e. integrating) all the probability densities that encompasses the region of the state (v_A):

$$p_A = \int_{v_A} dx p(x) = \frac{\int_{v_A} dx e^{-\frac{U(x)}{k_B T}}}{\int_v dx e^{-\frac{U(x)}{k_B T}}} \quad (1.59)$$

In order to operate $U(x)$ has to be approximated to a particular function. The simplest case is the treatment of a smooth energy basin using the harmonic (quadratic) function (**Fig. 1.3**, on the right), in which we can approximate the potential energy (solid green line) near the minimum by a simple “harmonic” that is, quadratic potential (dashed black line). However, the space region that encompasses a state is not straightforward, being the boundaries of a state very difficult to define. In addition, typical energy landscapes are complex and possess many barrier-separated basins complicating such efforts, so that in practice this sort of operations is not feasible.

1.3.2 Free Energy landscapes

The free energy of a microstate ($G(x)$) can be defined as the energy whose Boltzmann factor gives its correct relative probability density. Thus, the free energy of a microstate can be related with or approximated by its probability density via:

$$G(x) \approx -k_B T \ln p(x) \quad (1.60)$$

Accordingly, the Boltzmann factor for the energy of a single configuration tells you its relative probability (compared to another configuration), and the Boltzmann factor of the free energy of a state (e.g. state A) indicates its relative probability compared to other states (i.e. $G_A = -k_B T \ln p_A$).^[16] After all, probability (relative populations) is an observed behavior of the system described statistically. Thus, as long as we believe in probabilities, the free energy landscape approach is safe. However, it has been highly debated if the use of free energy landscape term for the analysis of population distributions that, for any reason (e.g. lack of enough sampling), is not the population distribution at equilibrium conditions (i.e. relative

probabilities not correctly associated). This is indeed the case for the sampling problem (see **section 2.4.2**). In this scenario, the lack of convergence in the energy landscape leads to imprecise energy values. In agreement with Daniel M. Zuckerman opinion,^[16] I consider that to say that a free energy landscape is trustable when it is constructed from imprecise population distributions is the same that to say that the free energy of a chemical reaction is reliable when the enthalpy changes (i.e. heat released in the process) are not measured with precision. In other words, who would trust the enthalpy values provided by a calorimetric pump that does not measure rigorously the heat?

Nevertheless, although non-equilibrium analysis does not provide the accurate free energy differences between states, they are very useful to sample the major states of the energy landscapes and to observe trends in energy differences. In those cases, and as we will see in some studies included in this thesis, just naming it as energy landscape or conformational population analysis are more appropriate terms.

1.3.3 Conformational free energy landscapes in proteins

In recent years, the population shift concept originated from the Monod-Wyman-Changeux model of allostery^[18] has become more popular than the induced fit model. Recently, Kovermann and coworkers provided evidence for a conformational selection pathway in the adenylate kinase (AdK) enzyme.^[19] It is worth mentioning that this enzyme was usually used as a model example of induce fit. As shown by X-ray crystallography, Adk adopts an open conformation in absence of the ligand, whereas a catalytically competent closed conformation is required for catalysis. According to the conformational selection model, this high in energy closed conformational state should also be visited in the absence of ligand, although with a lower frequency. By introducing a disulfide bond, they succeeded in trapping AdK in a closed conformation in the *apo* state. The X-ray structure provided the definitive proof of the closed conformation of the enzyme being also sampled in the absence of any ligand, thus highlighting that higher in energy functionally relevant states are visited even in the *apo* state. Similar to substrate binding, introduction of mutations to the enzyme sequence, protein-protein interactions, allosteric ligands and covalent modifications (e.g. phosphorylation) can induce a shift in the populations of the pre-existing conformational states. In **Fig. 1.4** is shown the Free Energy Landscape (FEL) of an enzyme that can sample open and closed populations among

others in *apo* state and how a population shift is induced by an external factor towards the closed conformational state, so that the enzyme activity is modulated.

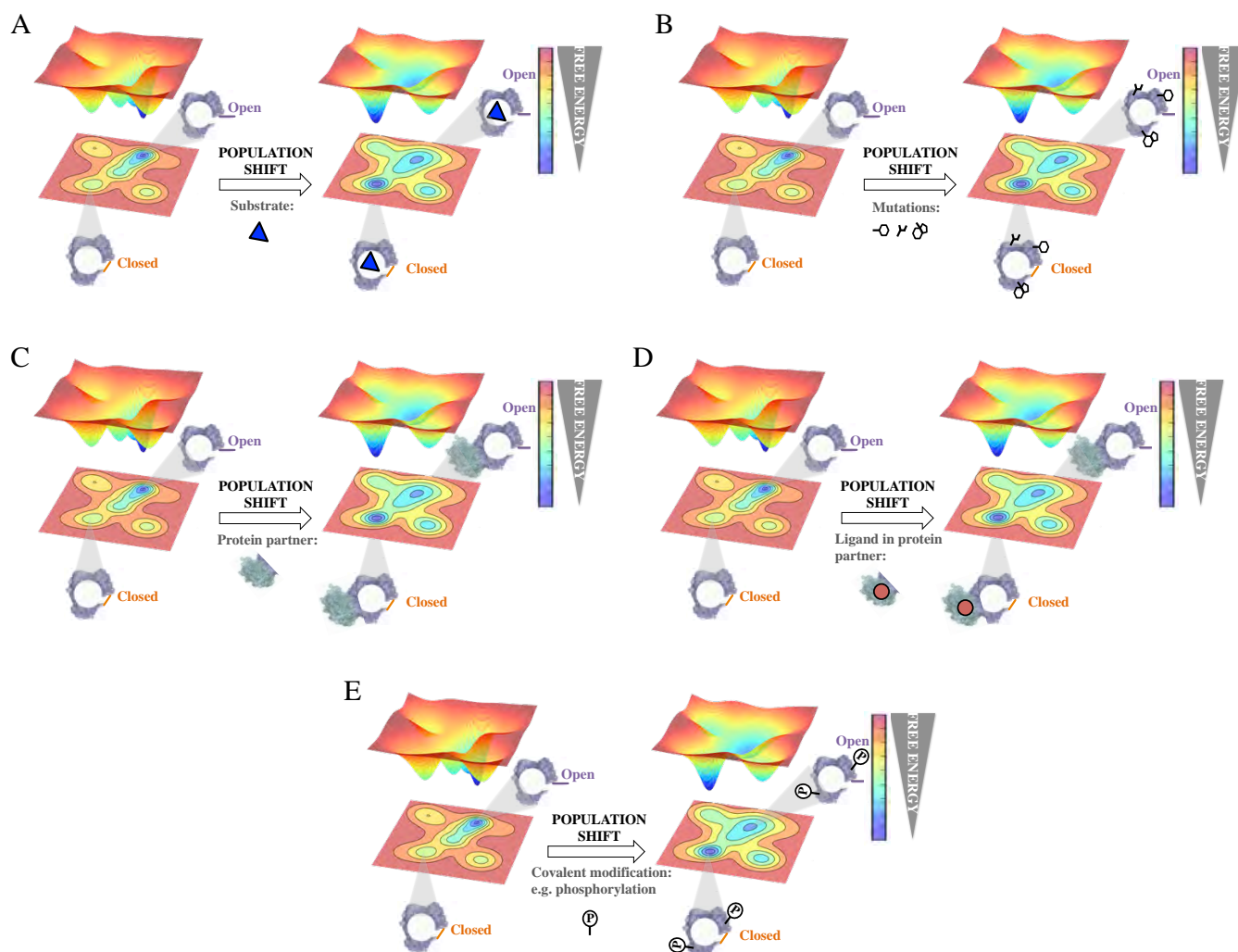


Figure 1.4 Schematic representation of an enzyme free energy landscape in the *apo* state associated with an open to closed conformational exchange and the population shift towards the closed state induced by a substrate (A), mutations (B), a protein partner (C), a ligand in protein partner (D), and covalent modifications (E).

High in energy conformational states relevant for substrate binding can be additionally important for conferring the enzyme the ability to accelerate additional promiscuous reactions,^[20] or for the enzyme evolution towards novel function.^[21] The effect of mutations on the relative enzyme populations was elegantly demonstrated by a recent example by Tokuriki and Jackson through an impressive collection of X-ray structures.^[21b] They demonstrated that the change in function from a phosphotriesterase into an arylesterase is achieved by gradual population of pre-existing conformational states, i.e. a population shift occurs along the

evolutionary pathway. Their study established that minor states that conferred the natural enzyme some arylesterase activity were gradually stabilized to become major states in the evolved arylesterases.^[21b] A similar finding was obtained by Jackson in evaluating how ancestral binding proteins evolved into specialist binders.^[22] An ancestral arginine-binding protein was crystallized in complex with L-arginine and L-glutamine revealing that the promiscuous binding of L-glutamine was possible due to alternative conformational states. These alternative conformational states were further populated along evolution to produce the contemporary L-glutamine specific protein binders. These studies support the idea that the underlying principle that guides enzyme evolution lies in the population shift of the conformational states that pre-exist in solution. The effect of introducing mutations to the enzyme sequence for their evolution towards new functions and novel substrate scope has a high similarity to substrate binding and allosteric regulation processes.^[21b, 21c, 23] In all cases, a redistribution of the populations of the conformational states exists, but in the particular case of enzyme evolution this population shift should favor the catalytically competent conformational states for the new target reaction. The allosteric properties of enzymes are further explained in **section 1.5.3**.

1.4 Chemical view

1.4.1 Fundamentals of catalysis

As discussed in the previous sections, enzymes are essential for living organisms; in this section, their mode of action will be discussed. All enzymes are proteins with the exception of a small set of RNA molecules. The catalytic activity of proteins is associated with its primary (amino acid sequence), secondary (alpha helix, beta sheets, random coil structures), tertiary (3D native structure, e.g. globular, filamentous) and quaternary structures (assembly of protein subunits). Some enzymes require additional chemical compounds to be active, called cofactors. The cofactor can be either one or several metallic ions (e.g. Cu^{2+} , Fe^{2+} , Zn^{2+} , Mn^{2+}) or complex organic molecules such as the Nicotinamide Adenine Dinucleotide (NAD) or the Pyridoxal Phosphate (PLP). Such organic compounds are called coenzymes and most of them are vitamin derivatives. Some enzymes require both, a coenzyme together with the metal ion (e.g. heme group). A prosthetic group refers when the cofactor is covalently or tightly bound to the

enzyme. The complete and active enzyme including the coenzyme and/or metal ions is called holo-enzyme while the enzyme in absence of the cofactor is called apo-enzyme.

In a general view, enzymes work as any other catalyst. They do not alter chemical equilibrium. Thus, the equilibrium constant and the reaction free energy exchange remain unaffected. What catalysts do is to accelerate the speed of the chemical reactions in order to reach the equilibrium faster. A reaction coordinate diagram is a representation of the free energy changes over the course of a reaction defined at certain conditions (e.g. 298 K, partial pressure of each as at 1 atm and concentrations of solutes at 1 M). In this context changes in the free energy can be estimated from experimentally determined equilibrium constants using the Gibbs equation (see Equation 1.50). In favorable cases, barrier heights between successive chemical species can be determined from studies of temperature dependence or from kinetic isotope effect data.^[8] In **Fig. 1.5-A** is represented a simple case where the substrate is transformed into a product in one single step. As shown in the diagram, there is an energy barrier to overcome in order to evolve the reaction towards product formation. In this regard, the substrate needs energy quantities to climb up to higher-in-energy configurations allowing for the reactant groups alignment, the formation of unstable transitory charges, chemical bonds rearrangement and other transformations. Once the reaction reaches the highest in energy configuration (i.e. transition state) the reaction coordinate can fall towards the substrate or product formation with the same probability. The transition state is not a chemical species with significant stability; instead it is a fleeting moment when the charge and bonds rearrangements reach the utmost unstable situation. The rate of a reaction can be estimated through the activation energy (ΔG^\ddagger) values: high energy barriers yield slow reactions (low rates). The reaction rates can be boosted by increasing the temperature, in this way a higher number of molecules have enough kinetic energy to overcome the activation energy. Alternatively, the activation energy can be decreased aided by the action of catalysts, such as enzymes (**Fig. 1.5-A-B**)

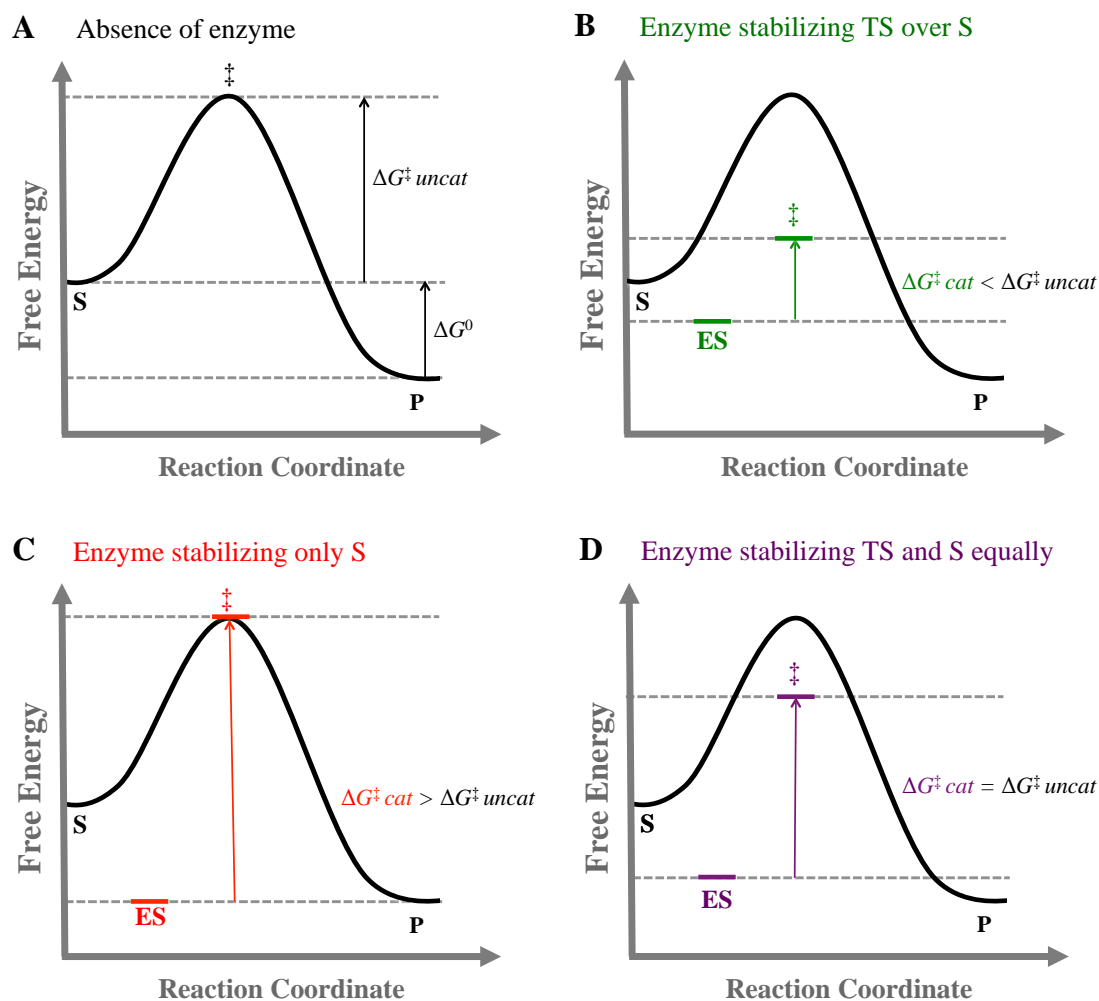


Figure 1.5 Reaction coordinate diagrams. It is shown the free energy as a function of the course of the chemical reaction of the substrate conversion to product (A), together with alteration on the free energy profiles when the reaction is catalyzed by the enzyme for different mechanisms (B, C and D).

In 1930 Eyring, Evans and Polanyi developed the Transition State Theory (TST) based on statistical mechanics justification. The theoretically constructed Eyring equation is more useful compared to the empirical Arrhenius equation (see Equation 1.3) and expresses the effect of the temperature on the reaction rates as:

$$k = \frac{\kappa k_B T}{h} e^{-\Delta G^\ddagger / RT} \quad (1.61)$$

Where k_B is the Boltzmann constant, h is the Planck constant, κ the transmission coefficient (often assumed to be ~ 1 for condensed phase reactions) and ΔG^\ddagger the activation free energy.

1.4.2 Transition state stabilization

Enzymes perform catalysis on a specific region of its tridimensional structure that geometrically consist of a cavity or pocket called active site. The fundamental question that arises is how enzymes decrease the activation energy. The answer can be split in different strategies that cooperate can together, the binding energy and the covalent rearrangements:

(i) **Binding Energy:** The enthalpic contributions of the non-covalent interactions between the enzyme and the substrate provide the main source of free energy used by enzymes to decrease the activation energy. In 1946 Linus Pauling proposed the idea that the active site geometry and charge distribution is precisely complementary to the transition state rather than the substrate. The substrate would bind in the active site forming the ES complex aided by some non-covalent interactions but the total non-covalent interactions will be formed when the substrate reaches the TS. In this context, the enzyme active sites balance the substrate and TS stabilization efficiently, i.e. stabilizing the TS over the S (**Fig. 1.5-B**). A way to do so is to stabilize regions of the substrate that most resemble the TS configurations while the geometries and charge differences between them are optimized for the TS. Note that if the active site was designed to be complementary only to the substrate, the ES complex would be highly stabilized improving K_M considerable, but the energy barrier to reach the TS would increase proportionally in decline of k_{cat} (**Fig. 1.5-C**). On the other hand, if S and TS are equally stabilized by the enzyme, little advantage would be gained because the activation energy would remain the same as occurring in the absence of the enzyme (**Fig. 1.5-D**). Warshel and coworkers reported that the main contribution for enzyme catalysis arises from the electrostatic stabilization of the TS.^[24] In this context, active sites provide a specific 3D structure with local charged groups that through strong Coulombic interactions stabilize ionic and polarized transition states. Note that stabilization of the very same transition state in bulk water would require a substantial thermodynamic penalty, referred to as a reorganization energy, for water molecules to be arranged in a manner that stabilizes ionic transition states.^[8]

The multiple non-covalent interactions formed between the substrate and the enzyme active site provide a substantial driving force for the catalysis. Equation 1.61 permits to calculate that ΔG^\ddagger has to decrease *ca.* 5.7 kJ/mol to accelerate a first order reaction a factor of 10 in the cell conditions. A single non-covalent interaction is estimated to provide among 4-30 kJ/mol. Thus, the global energy that arises from multiple non-covalent interactions is enough to decrease the

energy barriers by the 60-100 kJ/mol required to explain the great raise of rates observed in many enzymes.^[1] Besides, the binding energy also confers the enzyme its high specificity. The specificity is the ability of the enzyme to discriminate between different substrates. The specificity arises from the multiple non-covalent interactions between the enzyme active site and its specific substrate.

The sum of the unfavorable activation energy (positive) and the favorable binding energy (negative) results in a lower net activation energy. The main thermodynamic factors that contribute to the activation energy are the substrate distortions, entropy reduction, desolvation and catalytic groups alignment. All of them are paid by the favorable enthalpy of the binding energy:

- **Substrate distortions:** The free energy from non-covalent interactions formed during the TS formation compensates thermodynamically any substrate distortion; such as unstable electronic redistributions the substrate has to undergo to react. So that the energy required for the distortion is paid for by the binding energy.
- **Substrate entropy reduction:** In solution, the productive collisions between reactants are rare events. The binding energy constrains the degrees of freedom of the substrates in the active site and properly orients their reactant functional group aligning their molecular orbitals. In 1971 Page and Jencks demonstrated that the motion restriction going from a bimolecular to a unimolecular reaction involving an ester and a carboxylate group to form an anhydride yield rate increments of many orders of magnitude.^[25]
- **Substrate desolvation:** The interactions between the enzyme and the substrate replace most or all the hydrogen bonds that take place in solution between the substrate and the water molecules.
- **Enzyme conformational changes:** The binding energy can also induce enzyme structural rearrangements in order to increase the catalytic properties. These conformational changes may occur in small regions close to the active site pocket but also distal domains can be modulated. The conformational change can approach protein

functional groups and properly orient them towards the substrate and also can provide additional non-covalent interactions to stabilize the TS. This idea was initially postulated by Daniel Koshland as the induced fit model. However as discussed in **section 1.3.3** the conformational selection model is becoming more popular.

(ii) Covalent rearrangements: They generally involve transient covalent bonds between the enzyme functional groups and the substrate or the group transfers from or towards the substrate. These covalent interactions decrease the activation energy providing an alternative reaction pathway lower in energy, therefore boosting the reaction rate. Among the main covalent rearrangement mechanisms are the acid/base catalysis, the covalent catalysis and the catalysis by metal ions.

- **Acid/base catalysis:** In many chemical reactions some intermediates formed along the reaction coordinate may undergo charge instabilities that quickly decompose them back to its constituent reactant species. Proton donor/abstraction can stabilize these charge instabilities forming chemical intermediates that favor the product formation. This proton transfer may occur between the ionized water molecules and the substrate. However, if the proton transfer between the water and the intermediate is slower than the intermediate decomposition into its reactants, only a small fraction of the intermediates will be stabilized. At this point, enzymes strategically place amino-acid side chains in the active site that can act as proton donor/abstraction enhancing the reaction velocity considerably (from 10^2 to 10^5 orders of magnitude). The pK_a values of these catalytic residues is key for its action. Note that the protonation state of these groups is critical to perform the proton donor/abstraction, which causes pH dependence. It is worth mentioning that the pK_a values of the active site residues may differ significantly with respect to the pK_a values in solution. The active site environment modulates the pK_a : when acids and bases are placed into a hydrophobic environment the non-charged species are stabilized. Thereby acids exhibit higher pK_a favoring the protonated form of the carboxylate group ($-\text{COOH}$) whereas bases exhibit lower pK_a favoring the deprotonated form of the amine groups ($-\text{NH}_2$). The same trend occurs when in the vicinity of acids and bases are located like charged residues (i.e. negative in the case of the acids and positive for the bases). However, when in the vicinity opposite charged residues are placed, acids exhibit lower pK_a favoring the deprotonated

form of the carboxylate group ($-\text{COO}^-$), whereas bases exhibit higher $\text{p}K_{\text{a}}$ favoring the protonated form of the amine groups (NH_3^+).

- **Covalent catalysis:** It involves a transitory covalent interaction between the enzyme machinery and the substrate. In this regard, the enzyme modifies the course of the chemical reaction adding new chemical steps. These new steps provide a lower in energy pathway to the non-catalyzed reaction. Several side chain residues and cofactors can act as nucleophiles to form covalent bonds with the substrate. Finally, the covalent formed complex undergoes an extra chemical step in order to regenerate the free form of the enzyme.
- **Metal ions:** This strategy is difficult to classify because metal ions catalyze chemical reactions in many ways. Metal ions can be placed in the active site by coordination with enzyme residues or by taking part in coenzymes. In some cases, their mode of action can be simply attributed to the binding energy through ionic interactions (e.g. Mg^{2+} in kinases). In other enzymes they fix the substrate in the active site properly orienting it for catalysis, but also the coordinating bond between the metal ion and the substrate alters the electronic properties of the substrate improving its tendency to react, as for instance in zinc dependent Alcohol Dehydrogenases (ADH). Metal ions also transiently switch oxidation states during the catalytic cycle as in case of the iron in P450s or the copper in Tyrosinases. Furthermore, metal ions can be just part of the protein scaffold stabilizing distal regions of the protein, as is the case for Na^+ in tryptophan synthase, where the lack of the cation hampers dramatically the catalytic activity.

1.4.3 Role of conformational dynamics in catalysis

The ability of enzymes to visit different thermally accessible conformations has been explained previously in **section 1.3.3**. Here we focus on the link between enzyme dynamics and catalysis, which has been highly debated. ^[26] In 2002 Hammes-Schiffer suggested^[26a] that when the substrate binds the enzyme, it becomes an integral part of it. In this way both, the substrate and the enzyme experience simultaneous conformational changes affected by each other along the reaction coordinate. It becomes more obvious in multistep mechanisms. Multi-step enzyme cycles require that enzymes optimally stabilize multiple transition states. In this context, the ability of enzymes to adopt different catalytic conformations along the reaction pathway is

pivotal. As expected, this dynamic ability plays a key role not only in the chemical step but also in enzyme regulation, inhibition, substrate binding and product release.

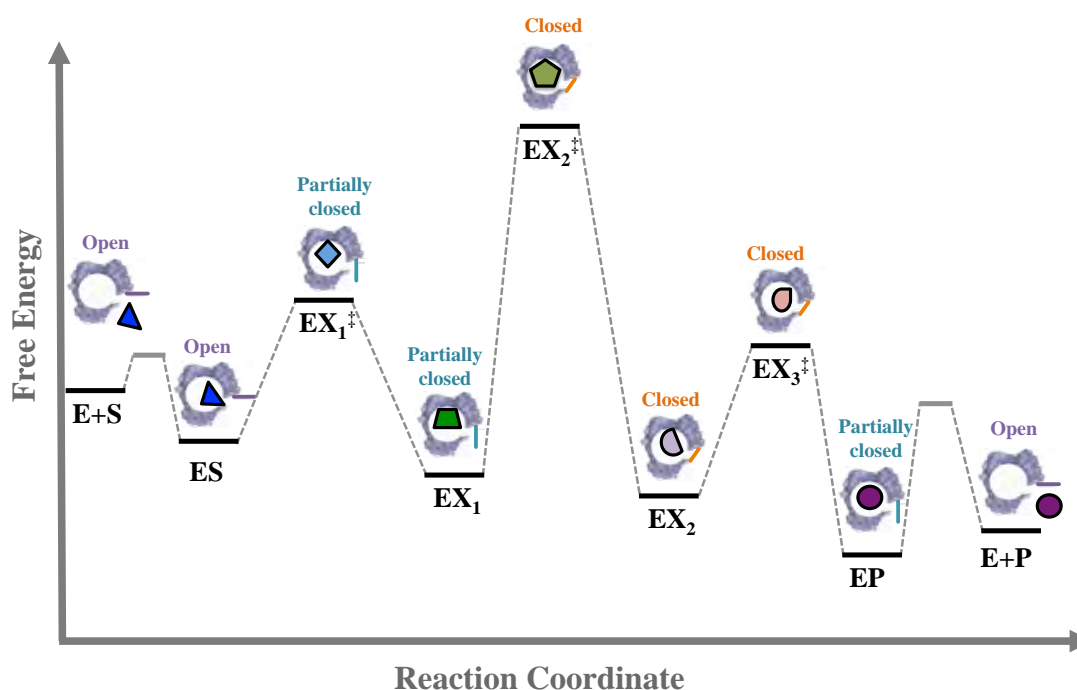


Figure 1.6 Hypothetical reaction coordinate diagram for an enzyme catalyzed-reaction. The structural changes of the enzyme and substrate along the reaction are depicted.

Reaction coordinate diagrams are valuable tools to interpret the enzyme catalytic cycles involving substrate binding, intermediates/TS structural evolution and product release. At difference with TS, intermediates have significant stability, although they usually have a short life-time. In **Fig. 1.6** a multi-step reaction coordinate shows the structural evolution of the enzyme and the substrate highlighting the catalytic enzyme conformation at each reaction step (i.e. open-to-closed conformational exchange). Herein the relative stabilities of the catalytic open-to-closed conformations at each reaction step is essential to efficiently optimize the catalytic pathway. Unstable catalytic conformations with higher energy barriers associated (e.g. adoption of closed state in **Fig. 1.6**) can contribute to the rate-limiting step of the reaction. The rate-limiting step on the catalytic cycle is the one with the highest energy barrier of the diagram, which often is the chemical step although in some cases conformational change can also be limiting. However other chemical steps with similar energy barriers can be rate-contributing.

Motions that occur in enzymes can display a variety of time scales.^[27] Bond vibration (10-100 fs) and side-chain rotations (ps to μ s) take place on the shortest time scales, whereas loop

motions, often key for substrate binding and product release, occur on the nanosecond up to millisecond time scales. On the longest time scales, slow domain motions and allosteric transitions can take place (μs to s),^[28] see **Fig. 1.7**. All these motions can precede or occur after the chemical step, and indeed in some natural and laboratory-evolved enzymes conformational changes have been found to be rate-limiting.^[29] Many examples have been provided in the literature highlighting the importance of engineering flexible loops for novel function.^[30] Recent studies based on the analysis of static X-ray structures along evolutionary pathways and in ancestral protein reconstruction,^[21b, 22] nuclear magnetic resonance (NMR) experiments,^[29b, 31] and computational studies based on Molecular Dynamics (MD) simulations^[21a, 27c, 32] have provided further support of enzymes as an ensemble of thermally accessible conformations, whose populations can be tuned. All these evidences highlight the crucial role of the enzyme conformational dynamics for its function.

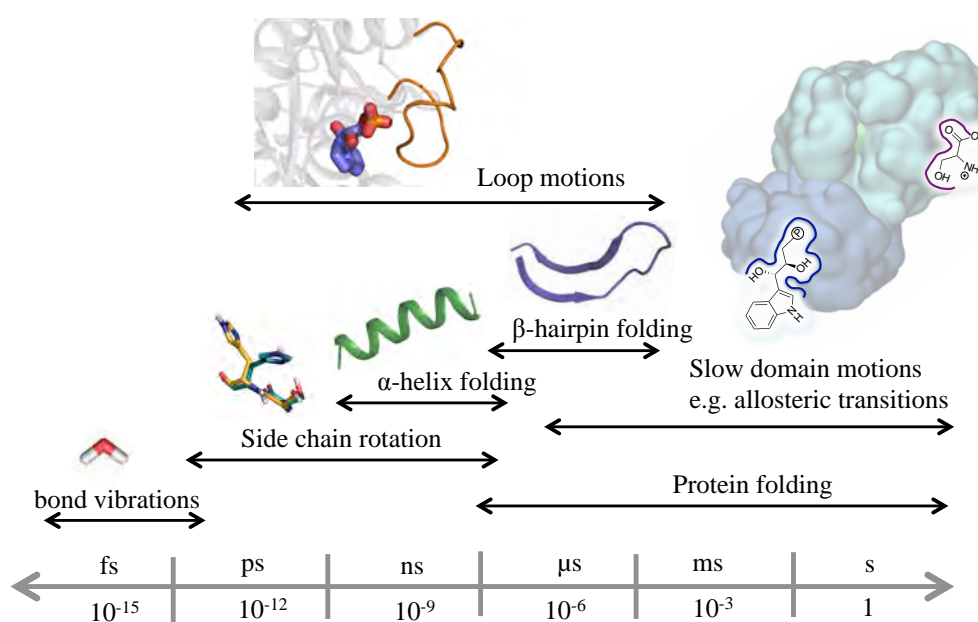


Figure 1.7 Time scales of the different fast and slow motions in proteins.

1.5 Engineering view

The use of biocatalysts in industry has expanded significantly over the last few decades. The employment of enzymes has impacted considerably many industries, such as pharmaceutical, food and biofuel. Enzymes arise as a potential alternative with respect to traditional catalysts for many reasons:

- **Great catalytic power:** often far higher than synthetic or inorganic catalysts. Enzymes dramatically accelerate chemical reactions; the increments on rate achieved oscillate from 5 to 26 orders of magnitude.^[1, 33]
- **High degree of specificity and stereoselectivity towards its substrates:** discriminating easily among substrates with very similar structures and when acting on pro-chiral substrates, precisely yielding the optically pure stereoisomer requested.
- **Environmentally and economic sustainable:** enzymes are produced from inexpensive renewable resources and are themselves biodegradable. They work under aqueous solutions and mild conditions of temperature, pressure and pH. In addition, they are easily amenable to economic modeling. The cost of production is stable at difference with catalyst based on metals like rhodium, whose market price leads to large fluctuations due to its scarcity and competing demand by other industries.^[34]

The main handicap is that for many industrial purposes there is no a natural enzyme that efficiently accelerates the targeted reaction or in some cases the reaction itself can be performed but the substrate scope needs to be expanded to meet industrial requirements. To that end, natural enzymes have to be engineered. Given that enzymes are large systems with high intrinsic degrees of freedom, they have an enormous potential to exploit the development of novel enzyme function.

1.5.1 Overview of enzyme engineering approaches

Enzyme engineering approaches can be divided into different strategies.^[34] The oldest consists of the variation of the reaction conditions together with kinetic studies, which allowed optimizing wild type enzymes towards the production of natural compounds. More recent strategies focus on rational approaches based on mutagenesis techniques. Site-directed mutagenesis (SDM) is a purely rational technique where the positions of the amino acids

subjected to mutagenesis and the nature of the amino acid substitution is selected by prior structural or functional knowledge. These strategies allowed the broadening of the substrate scope of many enzymes to obtain non-natural compounds.^[34] A striking approach that dramatically accelerated the pace of biocatalyst optimization (although with a high cost associated) is based on Directed evolution (DE) techniques, which were pioneered by Pim Stemmer and Frances Arnold.^[35] DE is a purely random technique that mimics Darwin's theory of natural selection evolution by inducing multiple random mutations on the enzyme sequence space. Posterior screening of the multiple variants generated allows for selection of the best hits to be subjected to a new DE round. The process is over when the desired enzymatic property (e.g. activity) is achieved. Other successful strategies are semi-rational approaches, that emerged as the combination of random methods (e.g. DE) with elements of rational design (e.g. prior knowledge),^[36] as for instance site-saturation mutagenesis (SSM) and iterative saturation mutagenesis (ISM). In SSM all the set of natural amino-acids are randomly tested for each rationally chosen position subjected to mutagenesis, and ISM methodology, that was developed by Manfred Reetz and coworkers, is a more sophisticated approach where iterative cycles of SSM on the chosen sites are performed, which often are sets of one, two or three amino acid positions.^[37] Another powerful semi-rational approach that aims to decrease the sequence space that DE has to search consists of taking advantage of the vast protein sequences, structures and functional information deposited in the databases to rationally guide the design process (i.e. selection of hotspots and creation of 'small but smart' libraries). Many nice bioinformatics tools have been developed to that end as for instance *HotSpot Wizard*,^[38] *ProSAR*,^[39] *SCHEMA* and *ARSA*. These tools have been successfully applied to (re)design enzymes for industrial approaches reducing screening efforts^[40] and have the great advantages of being extremely fast and easy to use, so that the tool is not limited to experts only.^[41] At a high rational level, computational strategies by means of computer modeling and thermodynamic calculations arose as rational approaches for the mutation prediction and evaluation.^[32b]

1.5.2 A brief story of enzyme computational design

Initial attempts to computationally engineer enzymes towards non-natural reactions or substrates were based on protocols that (re)designed the active site of some natural protein scaffolds by means of computational design strategies, in which a selected subset of active site

residues are subjected to mutagenesis while treating most of the enzyme protein as rigid.^[42] Despite the initial successes, computationally designed enzymes present quite low catalytic activities,^[42a] and needed to be further evolved by means of experimental techniques such as DE.^[43] Combining computational protocols and DE techniques (i.e. semi-rational approaches) has been shown to be a great strategy in designing enzymes for a broad scope of challenging transformations.^[44] The origin behind the low activities of computational designs has been attributed to the overly restrictive definition of active site residues,^[45] the imperfect realization of the ideal arrangement for TS stabilization,^[46] and the tendency to consider only the chemical steps while overlooking essential dynamic conformational changes for substrate binding and product release.^[47]

A more ambitious strategy is *de novo* computational-design, by which the whole enzyme active site is generated from scratch in an inert protein scaffold. This task is much more challenging than the (re)design since no advantage is taken from the natural protein scaffold. However, it provides a workflow by which the target design can be *a priori* any desired reaction that the user is interested in. One of the most successful approaches for *de novo* designs is the *inside-out* protocols that combine computational design software as for instance *Rosetta*^[48] with the *theozyme* concept.^[49] In this context, initial geometry optimization of the transition state including protein functional groups (i.e. involved in binding and catalysis) by means of quantum mechanics calculation provides the idealized three-dimensional model of a minimal active site, also called *theozyme*. This geometry is then placed into a protein scaffold using for instance *RosettaMatch*,^[50] *ORBID*^[42b] or *scaffold select*^[51] software. In addition a computational design simulation is performed (e.g. *RosettaDesign*)^[42a] to search within the protein sequence of the residues encompassing the pocket where the *theozyme* has been placed to optimize the packing between the transition state, the functional side chains and the nearby residues.^[52] The resulting designs can be experimentally tested. Moreover, iterative analysis based on X-Ray and MD data can be meaningful to rationalize and guide the improvement of previous designs.^[53]

These protocols have been shown to be extremely useful for designing new enzyme variants, based on different scaffolds, achieving some initial activity for some non-natural reactions including Kemp elimination,^[53-54]retro-aldol,^[44a, 44f] Diels-Alder,^[55] ester hydrolysis,^[56] and Morita-Baylis-Hillman^[57] reactions. However, as in the case of (re)designed natural enzymes, the catalytic activities of *de novo* designed enzymes are still orders of magnitude lower than those of natural enzymes,^[33, 58] thus requiring the employment of DE techniques in order to boost the catalytic activities by several orders of magnitude (see **Fig. 1.8**).

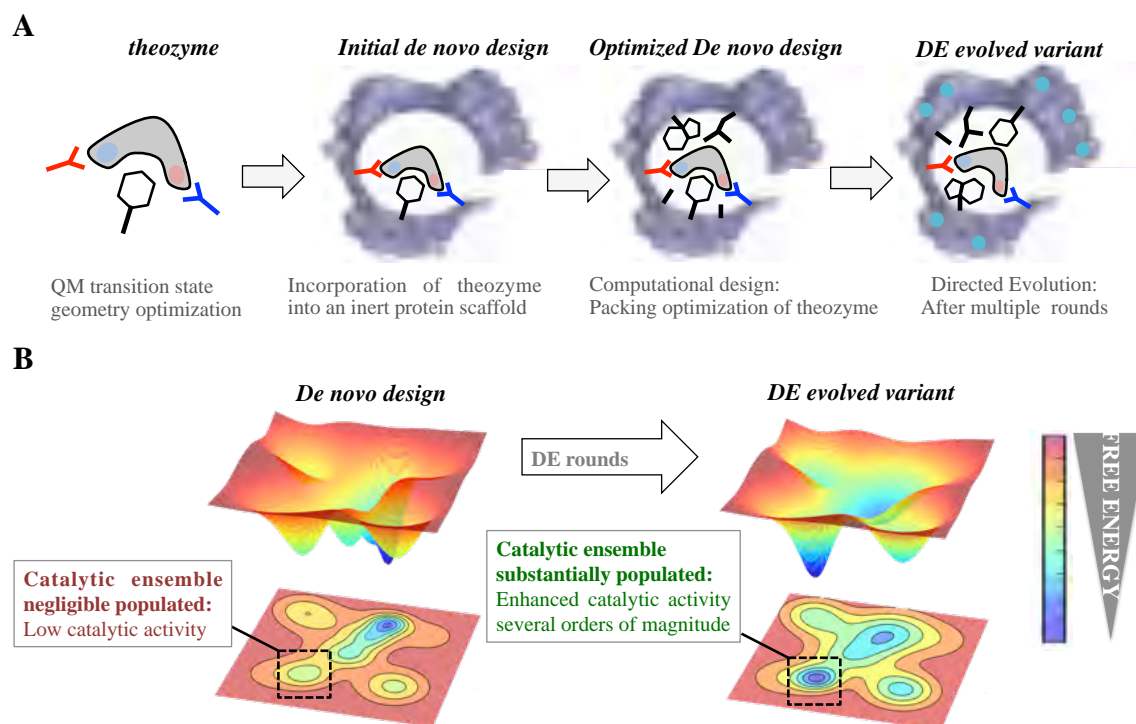


Figure 1.8 Schematic representation of the *inside-out* protocol and post-optimization by means of DE rounds, the active site positions are shown as sticks while the distal mutations as blue spheres (**A**) together with the FEL of *de novo* design and the most efficient DE variant (**B**).

One of the most representative cases of *de novo* computationally-designed enzymes was the creation of Kemp eliminases, which catalyze a proton abstraction from a carbon by a base. The *inside-out* Kemp eliminases exhibited quite low activities, due to the lack of precision to generate the perfect arrangement of the active site for catalysis.^[44e, 59] The different computational designs were further optimized through DE, making use of iterative design protocols that yielded new variants exhibiting higher activities.^[44b, 44d, 53, 60] The most efficient DE-evolved variant was obtained by Hilvert and coworkers and took 17 rounds of DE on an *in silico* designed variant with already substantial activity.^[44e] In a recent study, the room-temperature X-ray crystallography experiments show that the evolution of the *in silico*

designed Kemp eliminase towards the most efficient DE evolved variant is attributed to a rigidification of active site residues shifting conformational ensembles towards catalytically productive substrates. In addition, they highlight the fact that multistate approaches in the computational design field may be useful in order to save DE rounds.^[61]

Another highly proficient Kemp eliminase reported so far was recently created by Kamerlin, Sanchez-Ruiz, and co-workers using an alternative approach. They showed that through a single hydrophobic-to-ionizable mutation an ancestral β -lactamase substantial levels of a Kemp eliminase activity was achieved by assisting the generation of a new active site.^[62] It was striking that with only 1-2 mutations this new variant showed catalytic efficiency comparable to those of natural enzymes, even in the absence of a single DE round. Of particular interest is that such high activities were achieved mainly due to the modulation of conformational flexibility of the ancestral enzyme. In a recent study the efficiency of this *de novo* minimalist design containing only two mutations has been enhanced using a *Funclib*, which combines phylogenetic analysis and *RosettaDesign* to rank enzyme variants with multiple mutations that do not alter negatively enzyme stability. Their best ranked variant showed only one order of magnitude lower in efficiency than the most efficient DE-evolved variant.^[63]

A nice example of the importance of enzyme conformational dynamics and the population shift concept for acquiring new function was reported for retro-aldolases (RA). The *inside-out* protocol was applied for generating these mechanistically complex RA enzymes.^[44f] The designed RAs catalyze the cleavage of methodol substrate by a multistep reaction involving a Schiff base intermediate, between the catalytic lysine and the substrate. Hilvert and co-workers applied DE on the computationally designed RAs to enhance their modest activities towards methodol cleavage. One of the most important mutations was the introduction of a new catalytic lysine in the binding pocket in the second evolved variant (RA95.5). The introduced mutations completely remodeled the active site allowing a better positioning of the Schiff base intermediate for catalysis. Recently, a highly active RA variant (RA95.5-8F) was generated after multiple rounds of DE, which exhibits comparable activities to those of natural class I aldolases.^[64] RA95.5-8F features a sophisticated catalytic tetrad responsible for the enhanced efficiency of the enzyme. This series of studies shows the great power of DE in converting the original computational designs into highly proficient enzymes reaching activities similar to those of natural enzymes. It is worth mentioning that the rate limiting step of RA95.5-8F was

shifted respect to less evolved variants along the DE evolution pathway, as observed experimentally^[65] and by QM/MM calculations over the full multi-step reaction mechanism.^[66]

In order to elucidate the conformational dynamics of the different RAs variants generated along the evolutionary pathway, Osuna and co-workers computationally analyzed them by means of microsecond timescale MD simulations.^[21a] The conformational ensemble of the variants was explored through the application of the PCA technique to the MD simulations. By measuring the distance between the base and the Schiff base intermediate in the different conformational states sampled along the MD simulations, the major conformational states were sampled to distinguish among catalytically inactive and active conformations. The least active variant (i.e. the computational design RA95.0) sampled only a few catalytically active conformations. The population of the catalytically active conformational states was raised along the evolutionary pathway. The most prominent shift was observed for the last evolved variant showing that all the conformations explored were catalytically competent (RA95.5-8F). The analysis of the conformational landscape of the variants highlighted that the conformational heterogeneity of the computational and less evolved variants was tuned to progressively stabilize the catalytically active conformational sub-states, which become major in the most evolved variants. Interestingly, the RA intermediate variants that exhibit a high degree of conformational flexibility were found to be highly promiscuous.^[44i, 44j]

Advances in the available biophysical techniques and computational tools have contributed to a deeper understanding of the conformational dynamics of enzymes and their key role for function.^[32a, 67] The above-mentioned examples further confirm that the explicit consideration of the dynamic conformational ensemble of proteins in the computational design of novel enzyme function could greatly aid the community.^[68] In this context, extensive MDs simulations and MD-based enhancing sampling techniques have been shown to be very efficient to evaluate the conformational dynamics of enzymes and have been used to rationalize how mutations affect the catalytic activity of enzymes.^[21a, 27c, 32c, 69] In addition, MD data can provide meaningful information to guide enzyme design processes accounting for active site and distal positions (e.g. loop engineering).^[69b, 70] However, these strategies are computationally too expensive for the evaluation of a large set of variants, and when combined with QM/MM calculations or QM/MM MD simulations (i.e. linking the conformational dynamics with the chemical step)^[71] the computational cost is much higher. Besides, the

rational identification of the potential hotspots and the nature of the substitutions based on these thermodynamic calculations is often not straightforward and very challenging.

Some recent enzyme design protocols that attempted to account for protein flexibility at low computational cost were restricted to active site residues. The *inside out* protocol implemented short MD simulations at the end of the design process to identify and rank the best enzyme mutants based on how well the *theozyme* geometry was maintained in the MD runs.^[72] Another approach is the computational multistate design (MSD), which performs a computational design calculation over an ensemble of conformations rather than a single structure and then a combination function is applied to obtain the ranked sequences as a single score over all ensemble members (e.g. Boltzmann-weighted average). Finally, the top ranked sequences are used to generate combinatorial libraries of reduced size.^[73] This strategy can be applied for ensembles at different stages of the reaction pathway accounting for multi-step mechanisms.^[74] It is worth mentioning that this approach can be also implemented for more sophisticated combining functions, as for instance *Negative design* (i.e. combination function that leads to the stabilization of a target ensemble over another one). Other strategies are based on frameworks that combine *RosettaDesign* with high-throughput MD simulations to increase the conformational sampling to evaluate near attack conformation (NACs) frequencies. In this regard CASCO (Catalytic Selectivity by Computational design) developed by Janssen and similar workflows has been applied to engineer enzyme enantioselectivity.^[75] A similar strategy based on MD screening, also developed by Janssen, has been used to enhance thermostability: the FRESCO (Framework for Rapid Enzyme Stabilization by Computational libraries).^[76] Another strategy based on PELE (Protein Energy Landscape Exploration) calculations has achieved a great goal in the enzyme design field by engineering an additional active site on a natural protein. This new variant encompassing two active sites performed enhanced catalytic properties towards the natural reaction and proved the potential to design non-natural reactions, which can be exploited for cascade reactions.^[77]

A completely different approach regarding the strategies have been explained so far is the CADEE (Computer-Aided Directed Evolution of Enzymes) workflow developed by Kamerlin and co-workers. In this study, the authors present a pedagogical example of how the reaction barriers of a large number of variants can be estimated by means of an empirically-based QM/MM description of the reactivity using valence-bond theory (empirical valence bond (EVB) approach).^[78]

Most of the computational approaches reported in the literature based on thermodynamic calculations to deal with a large number of mutation predictions are restricted to the active site engineering, as for instance some of the above mentioned (e.g. MSD, CASCO, and EVB). However many examples have been provided in the literature demonstrating that mutations located at remote positions from the active site can have a large impact on the catalytic activity of the enzyme.^[32c] ^[64, 79] For instance, the effect of distal mutations has been nicely demonstrated experimentally and computationally in cyclophilin A.^[80] Indeed, no correlation is found between the influence of a given mutation on the catalytic constant of the enzyme and its proximity to the active site.^[81] Due to the broad sequence space of proteins, the computational prediction of distal mutations has been proven to be challenging.^[32c, 79] The key role exerted by remote mutations on the active site of the enzyme suggests that allostery (i.e. regulation of enzyme function by distal regions) is an intrinsic characteristic of enzymes,^[82] which might be exploited for enzyme evolution.^[21a] Recently Osuna and co-workers have shown that correlation-based tools usually employed for elucidating allosteric processes can be successfully applied in the enzyme design field, identifying key distal positions that might influence the enzyme activity.^[21a] The DynaComm.py python code developed by Osuna's group generates the shortest path map as output (see **Chapter 2.6**), which provides a residue network accounting for its role in conformational dynamics based on correlated motions. By comparing the SPM analysis with the positions mutated along the RA evolutionary pathway, most of the mutation points introduced in the different DE rounds were identified. The predictive power and applicability of SPM in the enzyme design field is assessed in **Chapter 5**.

1.5.3 Engineering stereoselectivity, thermostability and allosteric properties

(i) **Stereoselectivity:** This property adds an extra dimension to chemical specificity in biological systems. The selective formation of only one stereoisomer product from *pro*-chiral substrates is one of the most sophisticated tasks performed by enzymes. Given the importance of stereoselectivity in biology (i.e. only one stereoisomer has biological activity while the others can be toxic) natural enzymes evolved to be highly stereoselective. Reversing the stereoselectivity of a natural reaction or to induce stereoselectivity from scratch targeting a non-natural substrate requires the introduction of mutations. These changes in the protein sequence induce a *re*-shape in the active site pocket and a shift in the conformational native

ensemble (i.e. free energy landscape) in order to preferentially favor the formation of the desired stereoisomer. A powerful experimental method to enhance stereoselectivity and to expand substrate scope consists of a semi-rational DE approach applying iterative saturation mutagenesis (ISM) on a reduced set of relevant active site amino acids chosen (combinatorial active site test (CAST) sites), which is referred as iterative CASTing.^[37, 83] Second-sphere and distal mutations can also lead to a re-shaped binding pocket through allosteric effects.^[37a] Computational QM/MM calculations and MD simulations are promising tools to discern the factors governing the improvement in enzyme enantioselectivity at the molecular level.^[32b] Most of the computational evaluation studies are based on quantifying the frequency of the different catalytically productive orientations (e.g. *pro-(S)* and *pro-(R)*), which can be done by monitoring some selected angles and distances between the substrate and important active site residues along the MD simulations.^[75a, 84] By combining computational design with short MD simulations, Janssen and Baker successfully (re)designed the active site of an epoxide hydrolase obtaining enhanced enantioselectivities through the CASCO workflow^[75a] and in a later study Janssen and co-workers redesigned the enantioselectivity towards an hydroamination reaction using a similar workflow.^[75b] Recent studies have shown that the analysis of enzyme structure flexibility (through root mean square fluctuation (RMSF) analysis) along MD simulations can be used to identify key functionality in loop regions adjacent to the binding pocket.^[84c, 85] By modulating the conformational dynamics of these loops the reversal of enantioselectivity can be achieved.^[85b] The pivotal role of enzyme conformational dynamics towards novel enantioselectivity is assessed in **Chapter 4**.

(ii) Thermostability: Even if the novel function is achieved (e.g. stereoselectivity for a non-natural substrate), there is another parameter that is essential for industry purposes. Thermostability co-determines the feasibility of applying an enzyme in an industrial process. High stability is generally considered an economic advantage because of reduced enzyme cycles.^[86] Thermostability properties consist of the ability of enzymes to keep the native conformational ensemble over time avoiding denaturation. Irreversible thermal denaturation usually comprises a reversible unfolding step followed by an irreversible step involving aggregation and/or proteolysis. The deactivation step pulls the folded/unfolded equilibrium towards the deactivated conformations decreasing the enzyme reaction rate along time until all enzymes are deactivated. Thermostability studies strongly suggest that the unfolding processes that make a protein amenable to deactivation are partial/local rather than global in character.^[87] It is worth mentioning that this partial unfolding events primary involve surface-located parts

of the protein. When designing mutations for stabilization against irreversible processes, one may use the same reasoning as for stabilization against reversible, unfolding.^[86] However, when focusing on the unfolding process, only mutations that increase the ΔG of unfolding, stabilizing the local regions whose unfolding events triggers deactivation, would contribute to decrease denaturation rates.^[86] Several strategies have been identified to confer protein stabilization, such as entropic stabilization” (rigidification) by Gly \rightarrow Ala, Xxx \rightarrow Pro mutations, the introduction of disulfide bridges, “helix capping” by introducing residues that interact with the alpha-helix dipole, other types of helix optimization, the introduction of salt bridges and the introduction of clusters of aromatic–aromatic interactions.^[86] Comparison between thermophile and mesophilic enzymes together with DE studies indicates that Nature has employed many different structural strategies for obtaining high stability.^[88]

When assessing thermostability enhancement, it is noteworthy to distinguish between thermodynamic and kinetic parameters. The different approaches for the protein stability evaluation often leads to confusion and ambiguity. Thermodynamic stability is purely linked to the tendency of a protein to reversibly unfold, whereas the kinetic stability accounts for the deactivation process, which is often affected by the partial unfolding step, i.e. the thermodynamic stability contributes to the kinetic stability (see **Table 1**).^[15]

Table 1. Definitions of various thermodynamic and kinetic stability parameters.

Stability parameters	Definitions
Thermodynamic	
Free energy of unfolding (ΔG_u)	Change in Gibbs free energy going from the folded to unfolded state
Melting temperature (T_m)	The temperature at which half of the protein is in its unfolded state
Unfolding equilibrium constant (K_u)	The concentration of unfolded species divided by the concentration of folded species
Half-concentration ($C_{1/2}$)	The concentration of denaturant needed to unfold half of the protein (chemical equivalent of T_m)
Kinetic	
Observed deactivation rate constant ($k_{d,obs}$)	Overall rate constant for going from native to deactivation species
Half-life ($\tau_{1/2}$)	Time required for residual activity to be reduced by half
Temperature of half- inactivation (T_{50})	Temperature of incubation to reduce residual activity by half during a defined time period
Optimum temperature (T_{opt})	Temperature leading to highest activity
Total turnover number (TTN)	Moles of product produced over the lifetime of the catalyst

One of the most efficient computational protocols to enhance thermostability is FRESCO (Framework for Rapid Enzyme Stabilization by computational libraries)^[76] which was developed by Janssen and coworkers. FRESCO predicts mutations that stabilize the enzyme towards unfolding processes (i.e. increments in ΔG_u) and generates small mutant libraries requiring far less screening than conventional directed evolution methods.

The main disadvantage of enhancing thermostability in a mesophilic enzyme is that generally it is accompanied by a tradeoff in activity. Evolution tuned the thermostability of enzymes although upside down (i.e. enhancing activity with a tradeoff in thermostability). Ancient life most probably existed in hot environments (hot-start hypothesis).^[89] Ancestral enzymes had to cope with high temperatures, being thermophilic. After earth cooling, thermophilic enzymes had to catalyze reactions at low temperatures, and as a consequence the catalytic activity dropped dramatically. Cold adaptation consists of the evolutionary mechanisms that drove ancestral enzymes to become mesophilic. In other words, enzymes discount thermostability properties in order to be efficient at low temperatures. How catalytic efficiency adapts to temperature changes is currently poorly understood. However some thermodynamic insights have been reported, as for instance in the cold adaptation process of Adenylate kinase, whose rate limiting step was previously reported to be an enzymatic conformational change.^[90] These studies show how Adk cold-adapted enzymes present a lower energy cost associated with the rate limiting enzyme conformational transition (i.e. lower activation energy barriers) at low temperatures.^[89b, 91] In particular Hilser and co-workers reported that cold adaptation can be achieved by introducing Gly mutations on the protein surface, which increase the fluctuations of these regions (entropy tuning changes). These changes are allosterically propagated towards the active site modulating the enzyme activity.^[91] This dynamically-tuned allosteric mechanism provides insights into the previous studies in Lactate dehydrogenase, where the cold-adapted species presented identical active site whereas Gly was shown to be prevalent at surface sites.^[92] It seems that evolution has tuned the rigidity of the protein surface to become softer and more flexible than the hot-adapted one, while the active site residues appear to be identical. Mutations at the protein surface may provide a means for shifting the activation enthalpy-entropy balance as response to the altered working temperature. It seems that all naturally occurring enzymes that have been optimized to function at low temperatures catalyze their reactions with reduced activation enthalpies (ΔH^\ddagger) and more negative activation entropies (ΔS^\ddagger) compared with their warm-active ancestors. Decreasing the enthalpy activation barrier results in a reduction of temperature dependence.^[93] The

challenging task is to optimize the balance among minimizing the energetic cost of the catalytic conformational changes at ambient temperatures while keeping or enhancing simultaneously thermostability properties (thermoadaptation), which is assessed in **Chapter 4.2**.

(iii) Allosteric properties: Allostery is a process by which two distinct functional sites within a macromolecule are dynamically connected. The development of allosteric communication pathways in proteins has been essential for evolution. It conferred living organisms the capability for cell signaling and enzyme regulation. As expected, allosteric communication in multimeric enzyme complexes makes the enzyme subunits less active when isolated.^[94] This is a handicap in protein engineering because the use of isolated subunits is advantageous for biosynthetic applications. It decreases the metabolic load on the host cell and makes engineering other enzyme properties more feasible, such as stereoselectivity and stability.^[95] Arnold and co-workers addressed this problem in the Tryptophan Synthase (TrpS) enzyme and applied DE to the β -subunit of the allosteric enzyme complex TrpS, thereby successfully optimizing the β -subunit for stand-alone function (i.e. loss of catalytic activity dependency exerted by the allosteric protein partner).^[95-96] The allosteric and stand-alone function properties are assessed in **Chapter 5**.

From an engineering perspective, allostery has been a hassle. In essence allostery is an intrinsic behavior of proteins.^[82, 97] Practically it means that any signal that is propagated from a given protein region towards a distinct region allowing for function operates under the same rules. In this context, the signal or allosteric effector can differ significantly in their nature. It can be the binding of a ligand or substrate, a distal mutation introduced in the protein sequence, protein-protein interactions or a covalent attachment of a molecule (e.g. phosphorylation).^[98] In all cases, the allosteric effect induces a redistribution of the conformational ensemble that results in positive modulation (allosteric activation) or negative modulation (allosteric inhibition). Optimistically, one may think that controlling allosteric properties by means of protein engineering should be rather straightforward. However, the control of allosteric properties is extremely complicated.

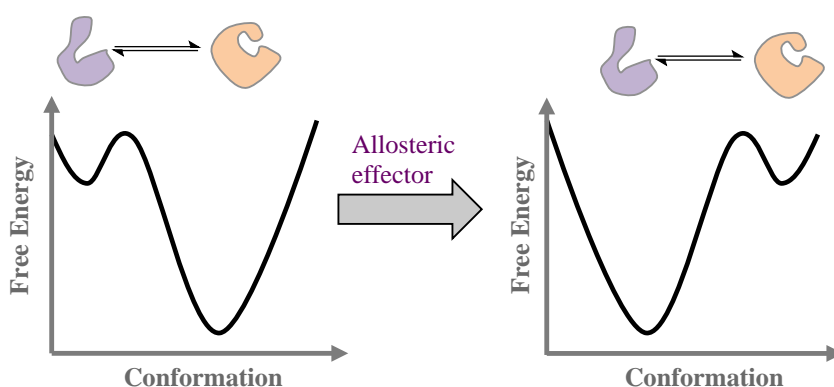
To approach the allosteric mechanisms, it is crucial to address first the thermodynamics of the enzyme-ligand binding process. When an allosteric effector (e.g. ligand or substrate) binds an allosteric site, both enthalpic and entropic contributions are involved. An increase in enthalpy by tighter binding results in a decrease in entropy through the restriction of the mobility of the interacting partners. This phenomenon is referred as entropy-enthalpy

compensation.^[99] Although such a compensation is widely observed, it is not a requirement. If both energy components were always compensated the binding process would never be favored.^[100] Another interesting phenomenon occurs in some proteins that tend to compensate the unfavorable entropic contributions in the allosteric site by increasing the protein dynamics in distant regions.^[100-101] These entropic effects are very difficult to predict in enzyme design. Besides, the non-covalent interactions formed between the effector and the binding pocket residues (i.e. binding energy) can induce structural tightening resulting in conformational changes via long-range interaction networks involving distant regions of the protein. In other words, the binding energy pays the energetic cost associated with distant conformational changes. Not least is the free energy of solvation in the interaction interface between the ligand and the allosteric site, which may have a dramatic effect on the binding process. In some cases, the large favorable solvation entropy that accompanies the binding makes the process entropically driven, as for instance in hydrophobic substrates. Hence, although it seems obvious that tighter interactions between the allosteric site and the ligand would favor the binding, the thermodynamic signature of a “good” binder does not need to be determined by the enthalpic term.^[100]

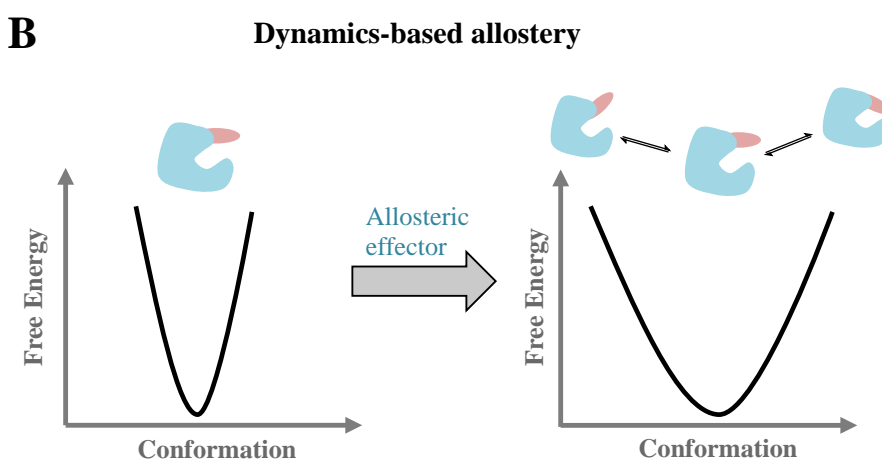
As seen above, the net free energy of the binding process accounts for many energetic contributions as the entropic (i.e. degrees of freedom) and enthalpic (i.e. non-covalent interactions) components among the substrate, the protein and the solvation shell water molecules and any of these partners can be the driving force or the major penalty, depending on the case. Note that in essence allosteric effects (i.e. distal effects induced by the allosteric effector) are the consequence of the thermodynamic signature of the binding process. In general, allosteric mechanisms can be classified in conformationally-based and dynamically-based allostery (see **Fig. 1.9**).

- **Conformation-based allostery:** This definition of allostery is much more intuitive than the dynamics-based because it is purely structural. In this context, the allosteric effector induces a population shift involved in long-range conformational changes that affects the function in a distal region of the protein, as for instance a conformational change from an inactive conformation towards an active conformation in the active site pocket. This redistribution of the conformational ensemble affects the relative stabilities of the preexisting structural equilibrium between inactive and active states and their rates of interconversion. It means that allostery allows for selectively

stabilizing a given conformational state (e.g. active) over the others. Conformation-based allostery has been extensively described for instance in the activation process of kinases upon phosphorylation, substrate binding and the introduction of mutations.^[102]



Conformational ensemble redistribution involving long-range conformational changes. After the allosteric effect the mean conformation is shifted from the orange conformations towards the purple ones (i.e. a population shift occurs)



Conformational ensemble redistribution involving conformational fluctuations within the same energy basin. After the allosteric effect the mean conformation is not altered and the red colored region increases the fluctuations (i.e. an activation of motions occurs)

Figure 1.9 Representation of the two main allosteric mechanisms including the conformation-based (**A**) and the dynamics-based (**B**) allostery.

- **Dynamics-based allostery:** This concept was popularized by Cooper and Dryden and relies on the changes of fluctuations in the protein conformational state that are limited to the basin of the free energy landscape of the protein (i.e. absence of population shifts associated with long-range conformational changes).^[103] Thus the nature of such allostery is based on entropic effects in distant regions of the protein. It is noteworthy

to mention that the conformational entropy contributions can be attributed to the micro- to millisecond time scale motions of the protein backbone and also to the fast backbone and side chain dynamics on the sub nanosecond time scale. Both of them have to be considered as contributors to the allosteric free energy transduction of proteins.^[104] In this regard, no clear correlation between side chain and backbone dynamics has been found. One of the first examples reported of dynamic allostery is the case of the negatively cooperative binding of the cyclic AMP (cAMP) molecule to the dimeric catabolite activator protein (CAP), which exists as a homodimer in solution and each subunit comprises a cAMP binding site. Interestingly, the binding of the first cAMP does not induce long-range conformational changes into the non-ligated subunit. Instead, it activates slow motions at the μ s-ms time scale resulting in an enhancement of dynamic fluctuations distributed in protein regions that are linked by cooperative interactions. Thus, providing a means of propagating the allosteric signal in absence of structural changes. The binding of the second cAMP extensively suppresses the motions (fast and slow) of almost all residues throughout both subunits, which drastically decreases the total entropy of the system. This large entropic penalty accompanying the binding of the second cAMP results in a weaker, and thus, anti-cooperative binding.^[105] Following this reasoning, in the cases where the allosteric effector reduces the flexibility of the active site residues, the active site preorganization primes the substrate to bind better and favors the binding process due to a lower entropic cost for the reaction process.^[103a, 104, 106] However, it has been postulated that bulky substrates, which are often the precursors of compounds of pharmacological interests have a higher dependency on the conformational dynamics of the binding site, in contrast to small substrates that are better recognized in more conformationally restricted active site cages.^[107] Thus the active site flexibility needs to be tuned accordingly, which in some cases is difficult to estimate.

In a recent study Taylor and co-workers compared the allosteric dynamic-based propagation with the mode when a violin is played (the “violin” model).^[103a] In this way, playing a particular note induces a redistribution of the vibration pattern on the body of the violin like the allosteric effector induces a redistribution of the dynamic fluctuations throughout the protein. I guess to learn how to play a violin may be complicated and take some years of training, but to learn how to play a protein is exponentially harder. Moreover, both allosteric mechanisms proposed (i.e. conformational-based and dynamic-based) may certainly act simultaneously.

The main techniques to explore allosteric properties are the isothermal titration calorimetry (ITC), Nuclear Magnetic Resonance (NMR) spectroscopy and theoretical approaches such as MD simulations. ITC allows to quantify the free energy of the binding process together with the enthalpic and entropic components as global parameters, thus containing a mixture of different contributions. NMR contributes to a deeper insight in the process. On one hand the different fast and slow motions among the protein structure are estimated revealing their role in presence and absence of the ligand (i.e. estimation of activated/suppressed motions after ligand binding). These dynamic data can be related to entropy by the order parameter S^2 though the relationship developed by Yang and Kay in 1996.^[108] Another key parameter extracted from NMR experiment is the chemical shift change upon ligand binding ($\Delta\delta$), which allows to distinguish between conformational-based and dynamic-based allostery. $\Delta\delta = 0$ means that the binding process does not alter the mean conformation of the protein (i.e. dynamic-based) except the flexibility and rates of interconversion between conformational states. In contrast, positive $\Delta\delta$ values indicate a substantial population shift towards a given conformational state (conformational-based).

Finally, theoretical approaches provide advantageous additional information. MD methods have been widely employed to assess the individual contributions of the binding free energy and their deconvolution. In particular, the conformational entropy can be for instance obtained from the MD trajectories through diagonalization of the covariance matrix of displacements of atomic Cartesian coordinates (Schlitter's approach)^[109] or quantified NMR-like via generalized order parameters (Yang Kay's relationship).^[108] MD-based approaches used to calculate the thermodynamics of the binding process include free energy perturbation (FEP), thermodynamic integration (TI), lambda-dynamics simulations, Molecular Mechanics-Poisson Boltzman surface area (MM-PBSA), linear interaction energy and hybrid quantum chemical/molecular mechanics (QM/MM).^[100] Moreover, the analysis of the free energy landscape after ligand binding (i.e. comparison of non-ligated with ligand-bound protein systems) provides a meaningful information about the nature of the allosteric effects since the conformational changes, the intermediates states and the energy barriers that connect them can be estimated, as well as the width of the energy basins, which is related with entropy. Recently, dynamical cross-correlation analysis of the MD trajectories has been found to be an interesting strategy to uncover allosteric signal transmissions and the allosteric residues involved (i.e. subset or residues that participate in the allosteric mechanism).^[110]

1.5.4 Ancestral enzyme properties

In recent years, scientists have been intrigued to assess the biotechnological potential of ancestral enzymes. Ancestral reconstruction consists of the process by which phylogenetic and statistical analysis using simple models of sequence evolution allow to resurrect ancestral proteins. Thus, it allows the reconstruction of how the modern descendants were generated from the Last Common Bacterial Ancestor (LCBA), that was supposed to exist a few billions of years ago. Their recent implementation in protein engineering arises from their exceptional properties:

- **High expression levels:** Ancient proteins likely had to fold in absence of the assistance of chaperones. Hence, their efficient folding process may have contributed to the enhanced expression levels reported in some of them.^[111]
- **Enhanced stability:** The high stability observed in ancestral enzymes reflects the high-temperature environment of ancient life (see above). The enhancement in denaturation temperatures are often on the order of a few tens of degrees, which is larger than the increments obtained in laboratory-evolved enzymes, but also when compared with the extant thermophilic enzymes.^[112] From an engineering perspective, this enhanced stability may be an essential factor for evolvability. In this context, the price of destabilizing mutations may be rewarded with enhancement of function.
- **Enhanced promiscuity:** Enzymes are not as specialized as we thought, being able to catalyze side reactions in addition to its native catalytic activity. In some cases, there seems to be no fundamental constraint in the number of tasks an enzyme can perform (e.g. enzymes involved in detoxification). Note that even in enzymes that are highly specialized (i.e. evolved to only perform one physiological function) low-level activity with no known physiological relevance is often observed. A promiscuous low-level activity provides an exceptional starting point for engineering a useful activity. However, given the fact that in most of the modern enzyme's promiscuity is an accident, finding promiscuous activities is a difficult task. Interestingly, many studies support that ancestral enzymes were more generalists with broad substrate scope.^[113] Hence, it is advantageous to search promiscuous activities in ancestral enzymes. In the near future ancestral reconstruction may become the common strategy to obtain initial promiscuous activities for biotechnological applications.

It is widely accepted that the origin of the enhanced promiscuity is due to an increase in flexibility near active site regions, which is compensated by increased rigidity in distal sites, thus maintaining the stability of ancestral proteins.^[112] The observation of flexibility changes at specific positions suggests a fine-tuning of the conformation ensemble along evolution. It is worth mentioning that analyses of protein families indicate that proteins evolve for different functions through sequence changes while conserving their 3D structure; this finding is shown when analyzing the 3D structures of ancestral enzymes. In addition, functionally critical positions (i.e. catalytic residues) are sequentially conserved suggesting that the evolvability towards novel function uses substitutions of distal positions that are dynamically coupled to catalytic sites rather than substitutions at critical active site positions,^[114] which reinforces the needs for computational protocols accounting for distal positions predictions in the enzyme design field, and especially when using ancestral scaffolds. The study of the conformational ensemble of a reconstructed LBCA and the distal sites prediction for stand-alone function is assessed in **Chapter 5.2**.

1.5.5 Computational design outlook

As seen above there are many properties that have to be fit towards the generation of an efficient enzyme. In order to climb on the fitness landscape, future approaches may involve enzyme engineering cycles including computational design and experimental tests. The challenge relies on the development of computational design protocols that take into consideration the conformational dynamics as well as the transition state stabilization. The introduction of mutations in the protein sequence should be assessed by determining how competent is the conformational ensemble, and for optimizing them towards the chemical steps. Existing computational protocols can properly estimate active site mutations for stabilizing the transition states of the desired reactions based in EVB and hybrid QM/MM approaches (i.e. active site optimal preorganization).^[115] In this context, the free energy landscape construction based on MM methods determines how the pre-existing conformational states are redistributed, estimating to which extent the competent states towards a target reaction are populated. However, the massive data obtained from the MD simulations makes the prediction of mutations to induce a population redistribution on the conformational ensemble towards the targeted active conformations very challenging. These engineering cycles mentioned could be reinforced through machine learning algorithms, by which the information that arises from the engineering cycles is used to climb the fitness landscape

efficiently. It can be done for instance by optimizing the computational descriptors (e.g. dihedrals, RMSF, map contacts, PCA, TICA, active site volumes and water displacements among others) that better predict enzyme improvements (computational scores). These protocols are in development and have been successfully applied for instance to enhance binding affinities.^[116]

Another aspect to consider is the selection of the initial scaffold used as starting point for enzyme evolution. Ancestral enzymes seem to be a potential option because of their large conformational heterogeneity and their ability to accelerate a wide range of promiscuous reactions. As explained above, ancestral enzymes have great properties to exploit evolvability. Regarding the rational selection of mutations, the high similarity between allosteric and distal mutations effects suggest that the tools developed for studying allostery, e.g. based on correlation measures taken from the MD simulations,^[110a] could be useful for predicting active site and distal positions that by mutation can induce a population shift.^[21a] Therefore, dynamical cross-correlation tools are promising for the generation of “small but smart” libraries for the rational design of enzymes, which combined with computational design strategies and bioinformatics tools could be of great relevance on the enzyme design field.

A boost in the computational power available together with the development of MM, EVB and hybrid QM/MM predictions will be also of great advantage since it would allow complete reconstruction of the free energy landscapes of proteins and more accurate transition state stabilization estimations in time-scales compatible with industrial demands.

Chapter 2. Methodologies

***In silico* approaches for enzyme studies**

As stated by Dirac in 1929, “the underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble”.^[117] The complication arises from the large number of particles (the so called many-body problem). A protein is a many-body system with tremendous degrees of freedom; so, physics and chemist have developed different ways to simplify their study. In this regard, proteins can be computationally studied at different levels of complexity. Generally, they can be described at atomistic level of resolution or using coarse-grained models. The latter are lower resolution models that are extremely efficient from a computational point of view since atomic details are lost (i.e. several atoms are grouped into single beads).^[32a]

Regarding the atomistic level of complexity, in principle, it can be split into three different layers of accuracy (**Fig. 2.1**). Quantum Mechanics (QM) is the most accurate methodology; QM calculations recover the energy of a molecule considering the nuclear and electronic configurations. To that end, the time-dependent Schrödinger equation has to be solved. Although many approximated methods have been developed to simplify the QM calculations, they still have an enormous computational cost associated, so their employment when operating with large systems is difficult. The strength of QM techniques in enzymatic studies is the underlying of enzyme reaction mechanisms, which is feasible by only treating with high QM accuracy a small and well-chosen part of the enzyme (e.g. cluster model approach^[118], QM/MM^[71a] or theozyme^[42a]). In this way, the active site reactivity (i.e. bond breaking and forming events) and properties that depend on the electronic configuration can be successfully computed. On the other hand, molecular mechanics (MM) techniques recover the energy of molecules using classical force fields (see next section), which ignore quantum effects. MM methods cannot compute enzyme reactivity, but in exchange, the decrease in accuracy guarantees a much better computational efficiency, making appropriate the study of enzyme dynamic properties encompassing the whole enzyme and explicit solvent.

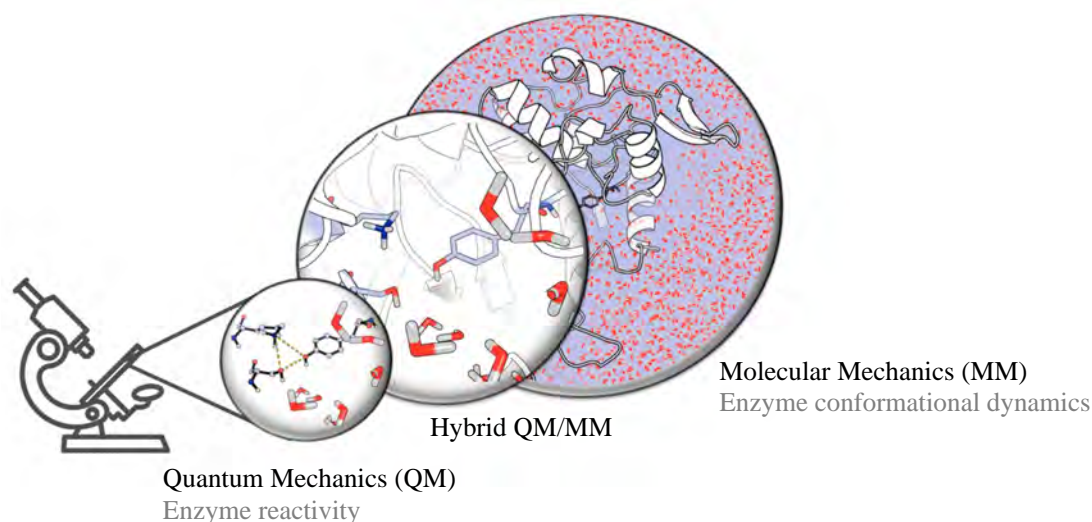


Figure 2.1 Representation of the three layers of accuracy to study enzyme reactivity and conformational dynamics computationally.

The balance between accuracy and system size has to be considered together with the nature of the process in order to obtain the most realistic data possible. In this thesis, we have been targeting the characterization of enzyme conformational dynamics, thus we have mostly used MM methodologies. However, QM calculations have also been performed to obtain force-field parameters for metalloproteins, substrates and cofactors (see **section 2.1.2**).

2.1 Classical Force Fields

Molecules are conservative systems, which are subjected to forces, and as a consequence work is performed. In this section the energy terms from a MM approach are described.

2.1.1 Potential energy function and parameters

A Force Field (FF) is a potential energy function, by which any protein conformation yields a potential energy value; $U(\vec{r}^N)$. Any conformation is then associated with a specific set of Cartesian coordinates of all atoms as follows:

$$\vec{r}^N = (r_1, r_2, \dots, r_N) = (x_1, y_1, z_1, x_2, y_2, z_2, \dots, x_N, y_N, z_N) \quad (2.1)$$

Where \vec{r}_i denotes the location of atom i and similarly for x_i, y_i, z_i . As mentioned above, MM force fields are based on classical physical models to predict the energy of the configurations

neglecting quantum effects. In this regard, only the positions of the nuclei (atom) are considered for calculating the energy. This approach assumes the Born-Oppenheimer approximation (i.e. the nuclear and electron motions in a system can be separated). In MM, the nuclei (atoms) are treated as “balls-on-spring” to represent molecules, where each ball corresponds to an atom and each spring to a covalent bond. Such balls or atoms oscillate around equilibrium distances.

The total potential energy (U_{FF}) is formulated as the sum of several terms allowing for all bond stretching (U_{str}), all angle bending (U_{bend}), all dihedral torsions ($U_{torsion}$), all out-of-plane distortions (U_{oop}), all van der Waals (U_{vdw}) and all electrostatic interactions (U_{el}) among the protein atoms. The FF is indeed a combination of the potential energy function and a set of parameters (e.g. force constants).

$$U(\vec{r}^N) = U_{FF} = U_{str} + U_{bend} + U_{oop} + U_{torsion} + U_{vdw} + U_{el} \quad (2.2)$$

It is worth mentioning that these terms are used in the FF to treat large systems such as proteins. These FFs are the so-called class I (e.g. AMBER, CHARMM and GROMOS), and recover the energy considering the potential energy function as simplest as possible. The FF terms can be grouped as bonded terms (U_{str} , U_{bend} , U_{oop} and $U_{torsion}$) and nonbonded terms (U_{vdw} and U_{el}). The U_{str} is the energy function for stretching a bond between two atoms (A-B). It is mostly used expression is a quadratic displacement of the minimum as harmonic oscillator (**Fig. 2.2**).

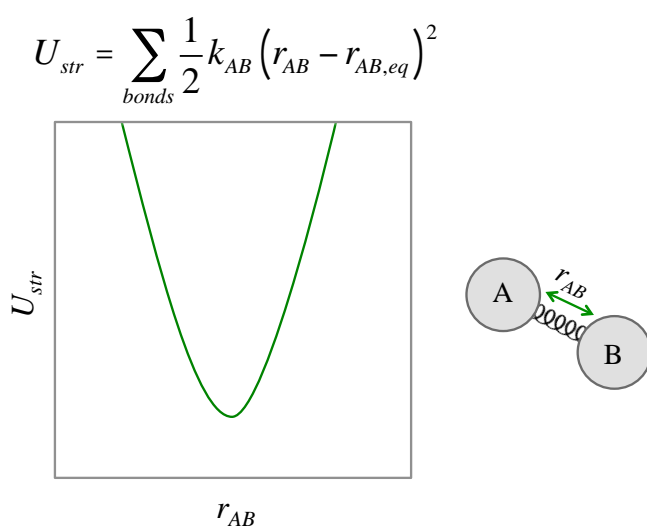


Figure 2.2 Representation of the potential energy function of the bond stretching term between A and B atoms, with k_{AB} the force constant of the bond, r_{AB} the bond length and $r_{AB,eq}$ the equilibrium distance.

The U_{bend} is the energy associated with an angle bend formed by three atoms connected by bonds (A-B-C), which is also described by a harmonic approximation (**Fig. 2.3**).

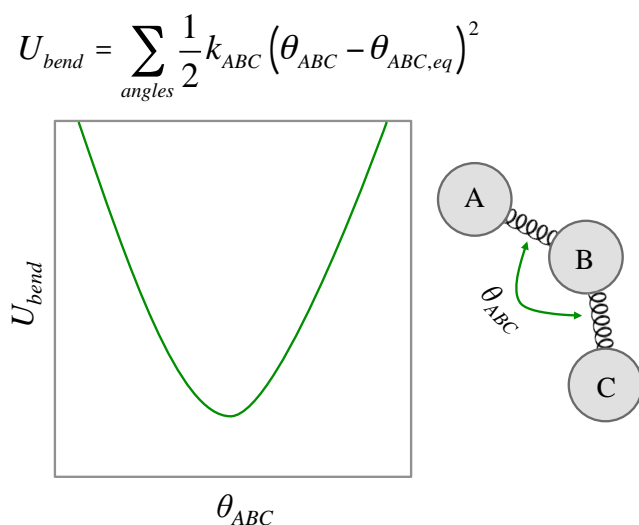


Figure 2.3 Representation of the potential energy function of the angle bending term between A, B and C atoms, with k_{ABC} the force constant of the bond, θ_{ABC} the angle and $\theta_{ABC,eq}$ the equilibrium angle.

In some FF, the out-of-plane distortions energy term (U_{oop}) is also included. This term is considered as the improper angle where the central atom is sp^2 -hybridized (ABC). The latest bonded term is $U_{torsion}$, which is the dihedral torsion energy associated with the rotation of the four atoms connected by bonds (A-B-C-D). This term is described as a (number of) Fourier series to describe the periodicity (**Fig. 2.4**).

$$U_{torsion} = \sum_{dihedrals} \frac{1}{2} V_{ABCD} (1 + n \cos(\omega_{ABCD} - \gamma_{ABCD}))$$

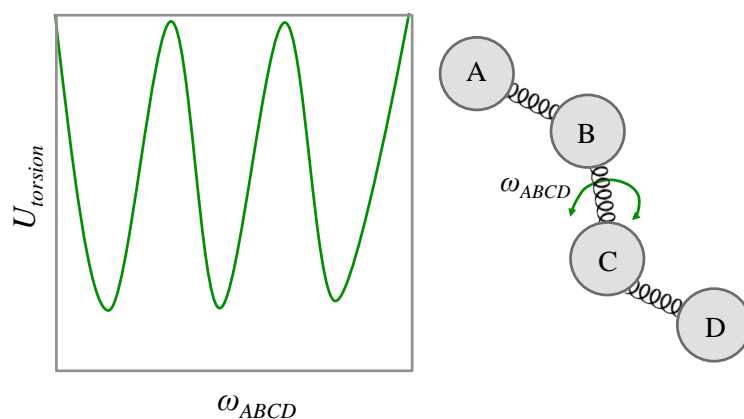


Figure 2.4 Representation of the potential energy function of the bond twisting term between A, B, C and D atoms, with V_{ABCD} the torsional force constant, n the multiplicity of the \cos function, ω_{ABCD} the dihedral angle and γ_{ABCD} the phase angle.

Regarding the non-bonded terms, U_{vdw} is the van der Waals energy, which is the energy associated with the van der Waals forces (also referred as London dispersion forces) between atoms. A common function that fits with the van der Waals energy behavior is the Lennard-Jones (LJ) potential set at 6-12 exponents (**Fig. 2.5**).

$$U_{vdw} = \sum_A \sum_{B=A+1}^N \left(4\epsilon_{AB} \left[\left(\frac{\sigma_{AB}}{r_{AB}} \right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}} \right)^6 \right] \right)$$

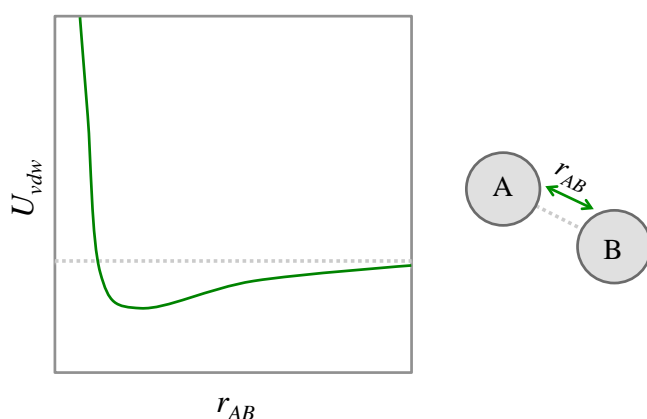


Figure 2.5 Representation of the potential energy function of the van der Waals term between A and B atoms, with ϵ_{AB} the well-depth, r_{AB} the distance between A and B atoms and σ_{AB} the interatomic distance at which repulsive and attractive forces exactly balance.

Finally, U_{el} corresponds to the energy associated with the electrostatic interaction between point charges given by a Coulomb potential (**Fig. 2.6**).

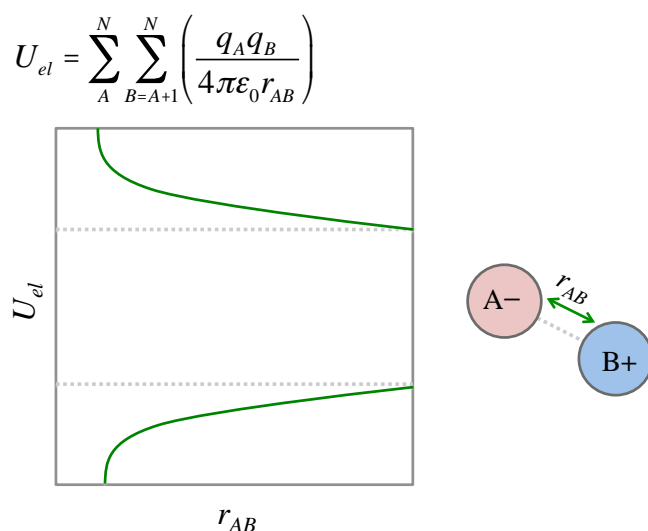


Figure 2.6 Representation of the potential energy function of the electrostatic term between A and B atoms, with q_A and q_B the atomic charges of atoms A and B, respectively, r_{AB} the distance between A and B and ϵ_0 the dielectric constant (usually set to 1).

Besides, class I FF have also been developed to consider explicit water molecules in the MD simulations, such as the rigid water models TIP3P^[119] and TIP4P-D.^[120] These are the simplest water models and rely only on non-bonded interactions (i.e. bonding interactions are constrained). The selection of more complex functional forms compromises between accuracy and computational efficiency. In this context, other FFs with higher accuracy have been developed to study smaller systems, such as organic compounds. This is the case of class II and class III FFs. class II (e.g. MMFF94) incorporate higher order terms (Taylor series) to treat bond, angle and dihedral terms and use 14-7 exponents on the van der Waals term. It also includes cross terms (U_{cross}) for describing the coupling between stretching, bending, and torsion. Cross terms are required to account for some terms affecting others. For example, a strongly bent H₂O molecule tends to stretch its O–H bonds.^[121] Class III FFs (e.g. polarizable CHARMM and AMBER) are considered the next generation of FFs and are optimized for hybrid QM/MM methods. They include quantum effects such as electronic polarization, where the charges of the atoms are not fixed during the simulation and so atom charge distributions can change induced by the environment. Polarizable FFs has been developed using various classical models, such us induced dipoles, fluctuating charges or charge-on-spring models (Drude oscillators).^[71a] FFs are in continuous development by the biomolecular simulation community in order to better reproduce natural processes. FFs development efforts are focused

on the integration of dynamic information derived from NMR experiments in the parameterization, and in the extension of the quantum terms like polarization in a more realistic manner.^[32a]

As stated before, FF consists of a combination of potential energy functions and parameters. Therefore, the quality of these parameters is a key factor to ensure realistic energies and conformational inter-conversions. FFs developed for proteins (e.g. AMBER) allow for all parameters needed for any protein. These parameters (if possible) come from experimental and computational data. Many of them allowing for bonded-terms were adjusted to reproduce experimental mode frequencies by fitting to structural and vibrational frequency data on small molecular fragments (peptides). However other parameters (e.g. dihedrals) were complicated to abstract from experimental data. For those cases the use of QM calculations is very common. In particular electrostatic non-bonded parameters are commonly obtained through QM calculations using the restrained electrostatic potential (RESP).^[122] The van der Waals terms are very difficult to obtain from QM calculations due to its limited accuracy when dealing with dispersion interactions. As a consequence, their values are usually fitted to experimental data in solid and liquid phases.^[123] All these procedures are settled in an iterative approach where several rounds of optimization are performed until the final model is obtained. Another comment worth mentioning is the importance of the atom types. In FFs, each atom has to be assigned an atom type. The assignment is based on the element, its hybridization and its local environment (e.g. the CT atom type is assigned to any sp^3 carbon). This approach decreases the dimensionality of the large number of atoms of the system. A relevant utility of empirical FFs is to include substrate/drug molecules in the calculations. However, to encompass the extent of structural and chemical diversity of substrates/drugs, parameters that show high transferability across a wide range of compounds is required. In this context, the Generalized AMBER Force Field (GAFF) was constructed to provide the parameters for many compounds allowing adding those to perform MD simulations. For consistency with the previous parameterization of amino acids, the general protocol to obtain ligands parameters consist of a single point energy (SPE) calculation at HF/6-31G* level of theory to the ligand geometry previously optimized. At this point, the optimized geometry and the electrostatic potential are used to estimate the atom types and its connectivity together with the atomic RESP charges, respectively (e.g. using antechamber). This information is used to search in GAFF (using the parmchk tool) for those force constants that match with the atom type and connectivity

provided. Nevertheless, for some compounds there are no parameters available, and in some cases, they are very difficult to extract, which complicates the setup of the MD simulations.

2.1.2 Missing parameters in metalloenzymes

Classical FFs often lack parameters for the metal atoms and metal-based cofactors present in metalloproteins. The description of coordination sphere metals of enzyme active sites is not straightforward and is a complicated task. Different approaches have been developed to consider the metals in class I FFs. The most important methods are the non-bonded model, the bonded model and the cationic dummy model (CDM).^[124]

The non-bonded model only considers the non-bonded energy terms to model the interactions of the metal in the center of coordinated atoms. The values of the U_{vdw} term parameters have been extensively adjusted for different metals and oxidation states to reproduce free energy of solvation (cations) and metal oxygen distances.^[124c] Moreover, more terms have been included in the LJ equation to consider the dipole-induced dipole interactions.^[125] The main advantage of this method relies on the low motion restrictions between the metal and the coordination atoms due to the lack of bonded-forces between them. Nevertheless, the main drawback is that the charge transfer is not considered treating the metal as an integer charged ball whose value depends on the oxidation state of the model (e.g. 1+, 2+, 3+...).

The bonded model considers all FFs terms described previously (i.e. bonded and non-bonded terms). Hence covalent bonds between the metal and coordination residues have to be created. To that end, the Seminario algorithm^[124a] presents a common strategy to obtain the parameters associated with the bonded terms although other protocols have been tested.^[124b, 126] Seminario's approach consists of a QM geometry optimization of the metal and coordinated residues followed by a frequency calculation. The algorithm suggested by Seminario obtains the force constants for bond, angle, dihedral and improper torsions involving the metal atom directly from the Hessian matrix (a matrix of the second derivatives of the energy with respect to atomic coordinates, i.e. the frequencies). The covalent bonds created between the metal and the coordination residues permits to account for charge transfer, whose values are calculated using e.g. RESP algorithm.^[127] The QM calculations for molecules involving metals are calculated at the B3LYP/6-31G* level because B3LYP includes electron correlation and the

computational cost is affordable. It is worth to say that apart from the coordination residues, reactant intermediates can also be attached to the metal. Nevertheless, the main limitation of this model is that the coordination number cannot change during the simulation, hence any oxidation state of interest and coordination spheres have to be parameterized independently and studied in a different MD simulation.

Finally, the CDM approach describes the metal by a set of cationic dummy atoms connected around a central atom in the specific coordination geometry to be attained. [124d, 128] The central atom is usually negatively charged while the dummies are positively charged. The number of dummy sites depends on the coordination number of the model. [124d, 129] This method provides a powerful non-bonded description for a range of alkaline-earth and transition-metal centers. It captures both structural and electrostatic effects since the charge can be distributed in different regions of the metal mimicking the coordination geometry.

In this thesis class I FF have been successfully used to study protein conformational dynamics associated with two main properties: allostery and enantioselectivity. The latest also required the employment non-bonded method using Seminario's algorithm in order to obtain metal coordination parameters of a zinc-dependent Alcohol Dehydrogenase (ADH) enzyme.

2.2 Molecular Dynamics

2.2.1 Newtonian dynamics

Molecular dynamics permits the study of complex, dynamic processes that occur in nature, such as protein allosteric transitions or enzyme-substrate binding. In a molecular dynamic simulation, a series of structural changes over time are obtained using Newtonian dynamics. First, the forces on any atomic coordinates are calculated from the FF. The forces on any atomic coordinates, for instance, atom i , x_i , y_i , z_i (\vec{r}_i), are given by the partial derivative of the potential energy as a function of atom i position ($-\partial U/\partial \vec{r}_i$). Once the forces are calculated, Newton's second law (Equation 2.3) can be used to follow the motion of all atoms (N) in a molecular assembly.

$$\vec{F}_i(t) = m_i \vec{a}_i(t) = m_i \frac{\partial^2 \vec{r}_i}{\partial t^2}; i = 1, 2, \dots, N \quad (2.3)$$

Where atom i at position r_i , is treated as a point with a mass m_i and a fixed charge q_i . This equation describes the motion of a particle of mass m_i along the coordinate r_i with F_i being the force on the particle in that direction. A MD simulation reenacts the simple life of atom: an atom will move at its current speed and direction unless it experiences a force that will accelerate or decelerate it and/or perturb its direction. In a molecular assembly, the force on each particle depends on its position relative to the other particles. As a consequence, a many-body problem arises which cannot be solved analytically.

Finite difference methods are used to generate molecular dynamics with continuous potential models (force fields), which are assumed to be pairwise additive. It means that the total force on a particle in the configuration at a given time is the sum of its interactions with other particles. The essential idea consists of breaking the integration of Equation 2.3 into many small steps, each separated in time by a fixed time (δt), under the assumption that speed and acceleration are constant over this small-time interval. Thus, from the forces calculated from the FFs the accelerations are determined and the positions and velocities at time $t+\delta t$ are calculated. Then the force on the particles in their new positions are determined, leading to new positions and velocities at a time $t+2\delta t$, and so on.

Many algorithms exist for numerically integrating the equations of motion (Equation 2.3) assuming that the positions and dynamic properties can be approximated using a Taylor series expansion.

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2} \delta t^2 a(t) + \frac{1}{6} \delta t^3 b(t) + \dots \quad (2.4)$$

Where $v(t)$, $a(t)$ and $b(t)$ are the first, the second and the third derivative with respect to $r(t)$. The Verlet algorithm^[130] is commonly used to further operate from equation 2.4 in order to solve the positions at $t+\delta t$ using positions and accelerations at time t , and the positions from the previous step, $r(t-\delta t)$.

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2} \delta t^2 a(t) + \dots \quad (2.5)$$

$$r(t-\delta t) = r(t) - \delta t v(t) + \frac{1}{2} \delta t^2 a(t) - \dots \quad (2.6)$$

By summing Equations 2.5 and 2.6:

$$r(t-\partial t) = 2r(t) - r(t-\partial t) + \partial t^2 a(t) \quad (2.7)$$

Note that the term involving change in acceleration (b) disappears and that velocities do not explicitly appear, thus having to be obtained with additional operations. The velocity Verlet algorithm improves the original version and accounts explicitly for the velocity from the beginning.

$$r(t+\partial t) = r(t) + \partial t v(t) + \frac{1}{2} \partial t^2 a(t) \quad (2.8)$$

$$r(t-\partial t) = v(t) + \frac{1}{2} \partial t [a(t) + a(t+\partial t)] \quad (2.9)$$

It is actually a three-stage algorithm. First the positions at time $t+\partial t$ are calculated using Equation 2.8. Second the velocities at *time* $t+\partial t/2$ are computed.

$$v\left(t-\frac{1}{2}\partial t\right) = v(t) + \frac{1}{2} \partial t a(t) \quad (2.10)$$

At this point new forces are computed from the current positions, thus giving $a(t+\partial t)$. Finally, the velocities at $t+\partial t$ are determined using:

$$v(t+\partial t) = v\left(t+\frac{1}{2}\partial t\right) + \frac{1}{2} \partial t a(t+\partial t) \quad (2.11)$$

Therefore, the accelerations, velocities and positions needed to start a new integration step, $t+2\partial t$ are obtained. Other algorithms have been developed from Verlet approach as leap-frog algorithm^[131], which also explicitly includes the velocities or the Beeman's algorithm,^[132] and uses a more accurate expression for the velocity.

2.2.2 Initial velocities

In the setup of a MD simulation, initial positions and velocities have to be assigned. The initial positions (configuration) ideally come from experimental data (X-Ray and/or NMR). Initial velocities can be randomly selected from a Maxwell-Boltzmann distribution at the temperature of interest (Equation 2.12).

$$p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T} \right)^2 \exp \left[-\frac{1}{2} \frac{m_i v_{ix}^2}{k_B T} \right] \quad (2.12)$$

The Maxwell-Boltzmann equation provides the probability of an atom i of mass m_i of having a velocity v_{ix} in the x direction at a temperature T . It is indeed a Gaussian distribution, which can be obtained using a random number generator. Most random number generators are designed to produce random numbers that are uniform in the range of 0 to 1. In this context, multiple separate and independent MD simulations starting from the same conformation (i.e. replicas) lead to different initial velocities, which means that the conformational exploration over time of any MD replica would be dramatically different, speeding up the FEL sampling by collecting better statistics. Note that even a small difference in the initial configuration, such as the coordinates of one single atom, lead to an exponential divergence of the time evolution of the system.^[133] The decision among running a large number of replicas or only a few long MD simulations is compromised by the size of the system and time-scale of the process under study. However, in many cases, it is not straightforward.

2.2.3 Time step

The selection of the time step (δt) is a critical parameter for a MD simulation. If it is too big instabilities may arise due to high energy overlaps between atoms. On the other hand, if it is too small the trajectory will cover only a limited portion of the phase space, which for a system of N particles means each combination of $3N$ coordinates (r_i) and $3N$ momenta (v_i), i.e. in $6N$ dimensional space; with a high computational cost associated. The smaller the time step the larger the number of integration steps have to be performed for a given simulation time. Thus, the appropriate time step should cover the phase space efficiently with collisions occurring smoothly. A useful guide to determine it is that the time step should be approximately one-tenth the time of the shortest period of motion. The highest-frequency vibrations are due to bond stretches, especially those of bonds to hydrogen atoms (e.g. O-H stretch vibrates with a repeat period of approximately 10 fs).^[134] So for treating accurately the molecule motions a time step of 1fs is required. However, the computational cost is dramatically decreased when constraining hydrogen bond to fixed lengths. This is done by using different methods like SHAKE^[135] and LINCS^[136] algorithms. These approaches allow for larger time steps, often double (2 fs). Another approach consists of classifying forces within a system in groups

according to how rapidly the force varies over time. Each group has its own time step while maintaining accuracy and numerical stability (multiple time step dynamics).

2.2.4 Periodic boundary conditions and cutoff distance

The selection of the periodic boundary conditions is also crucial in a MD simulation. It allows more accurate estimation of macroscopic properties from simulations using a relatively small number of particles. In this context, the simulated system within a periodic box interacts with periodic images of the same system alleviating many of the issues with finite size effects (i.e. the periodic approach attempts to reproduce infinite lab-scale systems). In this way, the particles experience forces as if they were in bulk fluid. If a particle leaves the box during the simulation, then it is replaced by an image particle within the central box. Thus, the total number of particles remains constant. It is worth mentioning that multiple interactions between periodic systems are indeed undesirable. To prevent this, a cutoff distance regarding the non-bonded terms is employed, in such a way that the non-bonded interactions between all pairs of atoms that are further apart than the cutoff are set to 0. The cutoff value should be chosen such that it is less than half the length of the simulation box in any dimension.^[133] A value of 8-12 Å is generally recommended. These cutoffs impose a natural lower limit to the size of a periodic simulation box, as the box must be large enough to capture all of the most significant non-bonded interactions. In principle, any cell shape can be used if it fills all of space by translation operations of the central box in three dimensions. It is often sensible to choose a periodic cell that reflects the underlying geometry of the system.

In principle, the proper selection of the cutoff value will assure that proteins do not directly interact with each other while they may interact through the perturbation of nearby solvent. However, if the solvent does not reach a bulk-like state between proteins, the simulation will undergo finite-size effects.^[133] The employment of the cutoff clearly decreases the computational cost of the simulation since the non-bonded terms calculations are more time-consuming than the bonded terms. Moreover, the minimum image convention is applied where each atom 'sees' at most just one image of every other atom in the system. The energy or force is calculated with the closest atom or image.

2.3 Running a MD simulation

Setting up and running a MD simulation includes several steps (system preparation, minimization, heating, equilibration and production run). Here, we aim to highlight some relevant aspects on every step allowing for a general understanding instead of encompassing the technical aspects (i.e. input files or specific issues regarding the nature of the system). The system preparation consists of many steps allowing for the selection of the initial configuration of the protein and ligand (if added), the protonation of the protein constituents, the addition of counter ions and solvent box, and the selection of the FF. All of these steps are critical and deserve as much care as possible. If any of the steps turns out to be wrong according to what you intended to describe, then all the accumulated data and post analysis would lead to a misguided outcome. Once one has checked that the system is correct, the minimization step proceeds to find a local energy minimum of the starting structure to avoid instabilities when running the MD. Afterwards, the heating process consist of a MD run that increases the temperature in several steps up to the targeted temperature and then a short MD run is performed to equilibrate the system. A MD run is performed in a particular thermodynamic ensemble, i.e. it is a collection of points in phase space satisfying the conditions of a particular thermodynamic state (e.g. energy). Traditionally MDs are performed in the constant NVE ensemble (microcanonical), which is characterized by fixed number of atoms (N), volume (V) and energy (E). The two most common alternative ensembles are the constant NVT (canonical) and the constant NPT (isobaric-isothermal ensemble). Regarding the equilibration step, the main goal is to monitor the macroscopic properties (e.g. E and T) to ensure they reach a steady state on average. Although not rigorous, the equilibration step is over when the macroscopic properties, as for instance the energy, fluctuate around constant values with minimal drift. A more difficult aspect is to ensure other properties of the system do not oscillate that much over time (i.e. protein interactions, protein conformations...). A common strategy is to monitor the root mean squared deviation (RMSD) of the system. Once the RMSD values are not systematically changing, the equilibration step has been successfully performed. After proper equilibration, the production run can be started. At difference with the equilibration step, the data obtained from the MD production run can be collected for analysis. However, when altering the previous conditions for some reason (e.g. box dimensions, temperature or pressure), data should not be collected immediately, instead an additional equilibration step is needed.

During MD simulations, a thermostat adds and removes heat from the system. The temperature of a molecular dynamic simulation is related to the time-averaged kinetic energy using the equipartition theorem^[133]

$$\frac{3}{2}Nk_{\text{B}}T = \left\langle \sum_{i=1}^N \frac{1}{2}m_i v_i^2 \right\rangle \quad (2.13)$$

The simplest way to alter the temperature is to scale the velocities (simple velocity rescaling). There are many other thermostats algorithms, and all of them work by altering the Newtonian equations.

2.4 Free energy landscape construction

2.4.1 Dimensionality reduction of the MD data set

The accumulated data obtained from the production runs is used to recover the Free Energy Landscape (FEL) associated with the protein conformational population distribution. Unfortunately, as a result of the large number of atoms present in the simulations (*ca.* 100,000 atoms for a protein of regular size in an explicit solvent environment), the atomic population distributions are defined by an extremely high dimensional space. A potential solution to this drawback is to focus on a reduced set of global or collective degrees of freedom (DOFs). These DOFs can be defined as a simplification of the enzyme coordinates that describes any explicit function, relevant to the process of interest. For instance, distances between catalytic residues, backbone dihedral angles or the RMSD of a loop. High dimensional data obtained from the MD simulations can be projected onto these collective DOFs obtaining the probability distributions and reconstructing the associated free energy landscape (**Fig. 2.7**).^[68]

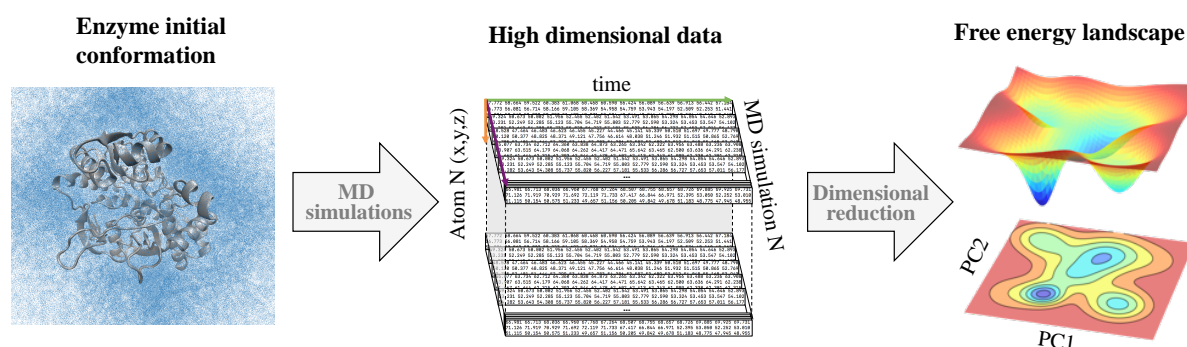


Figure 2.7. Schematic view of the dimensional reduction process of multiple MDs accumulated data set.

Reducing the dimensionality of the MD data to only a few DOFs can omit essential kinetic or thermodynamic information relevant to the process under study. Besides, choosing an appropriate set of DOFs requires a detailed knowledge of our system and for many systems is not easy to identify. A minimum criterion for the low-dimensional projection of the conformational space is that the set of DOFs clearly distinguish among the different metastable states under study and the transition states that interconnect them. In this thesis, we apply dimensionality reduction techniques to the study of two different biological processes: enantioselectivity and allosteric transitions. For the enantioselectivity study (**Chapter 4**), we evaluate the competent pro-(S) and pro-(R) conformational landscape by computing the angle between two substrate atoms and an atom from a rigid active site residue, together with the measure of the hydride transfer distance between the substrate and the cofactor. The angle selected successfully captures the substrate rotations in the active site and the angle values properly discriminates between the pro-(S) and pro-(R) states. Besides the hydride transfer distance successfully estimates catalytically competent conformations (at short distances). For the allosteric transition study (**Chapter 5**), more complex DOFs were chosen in order to describe an open-to-closed transition of a rigid domain that acts an active site lid. In this case, a path of conformations from open to closed conformations is generated by linear interpolation of available X-ray data. The reconstruction of the conformational energy landscape associated with the open-to-closed path generated together with a restricted degree of deviation (i.e. distance from the reference path) successfully captures the relevant catalytic states, which allows us to decipher allosteric effects.

Apart from the examples mentioned, other approaches to automatically reduce the dimensionality of the data while preserving as much information as possible have been developed, such as the Principal Component Analysis (PCA).^[137] PCA performs a dimensionality reduction accounting for as much variance in the data set as possible. In a nutshell, if we define variance as the deviation of an atom from its mean position along the MD, then each principal component will be a linear combination of strongly correlated atomic motions with large oscillations. The resulting low dimensional PCA space can then be used to reconstruct the associated free energy landscape (**Fig. 2.7**). For example, PCA has been applied in several studies of protein folding and allostery.^[80b, 138] However, transitions with the highest variance do not strictly correlate with the slowest (i.e. kinetically relevant) processes. Contrary to PCA, the time-structure independent component analysis (tICA) seeks to lower the dimensionality of our data while minimizing the loss of kinetic information.^[139] This is done

by considering the time correlation of the data instead of the variance. Alternative approaches to reduce dimensionality include Diffusion Maps,^[140] the variational approach,^[141] and the Sketch-Map^[142] among others. Once the MD data has been collected and the DOFs have been properly selected, the accumulated probability density as a function of the selected DOFs can be obtained (i.e. the histograms). There are several methods to calculate the histograms, for instance the kernel density estimation as implemented in PLUMED module analysis.^[143] Finally, as stated in **section 1.3.2**, the free energy is related with its probability via the Boltzmann factor ($G(x) \approx -k_B T \ln p(x)$) and the FEL plot can be constructed.

2.4.2 The sampling problem

The time-dependent thermodynamic properties extracted from MD simulations can only be connected with experimental observables if all relevant states or conformations of the system are visited (i.e. ergodic principle).^[17] The ergodic principle states that if the system evolves in time indefinitely, the system will pass through all accessible microstates in statistical equilibrium. So that the time-averaged conformational sampling and the average over the statistical ensemble (i.e. probability distribution of the microstates at thermodynamic equilibrium) are the same. In practical situations, this is not normally the case. To properly integrate the equations of motion, atomistic MD calculations using empirical force fields typically use time steps of the order of femtoseconds (i.e. 10^{-15} seconds), being able to compute few nanoseconds with a personal computer, but far from the millisecond to second time scales of domain motions and allosteric transition occurring in some enzymes.^[144] This time scale gap frustrates direct comparison with experimental data, encouraging for alternative approaches, which can be broadly classified in unbiased and biased methods.^[68]

2.5 Enhanced sampling techniques

After more than 40 years since the first MD simulation of a protein was performed,^[145] the basic MD algorithm remains unaltered. Then, the question is, how can we increase the accessible time scales to make reliable connections with experiments? Here we detail some of the most commonly used strategies.

2.5.1 Unbiased MD methods

(i) CPU parallelization leads to an enormous increase in the accessible simulation time scales. This strategy simulates extremely large systems during moderately long simulation time thanks to a *divide and conquer* approach (i.e. the enzyme system is broken down to smaller entities, each one being computed on the different connected CPU). This approach was used in a MD simulation of a complete solvated tobacco mosaic virus capsid with up to 1 million atoms.^[146]

(ii) The Anton supercomputer, which was specifically developed as a special purpose computer by D.E. Shaw and co-workers to perform long single MD simulations of biological systems. The first atomistic millisecond MD simulation of a WW protein domain was performed with Anton.^[147] This computer has also been used to study the fold of a series of small proteins,^[148] allosteric transitions in G-protein membrane receptors,^[149] and ligand binding kinetics^[150] among others.

(iii) GPU based clusters offer an affordable alternative to increase MD accessible times by running either single long and/or multiple short simulations of the same system. Some MD codes have been specifically designed to run on GPU's, such as AceMD,^[151] and OpenMM,^[152] whereas others have been ported to GPU's (Amber,^[153] Gromacs,^[154] and NAMD^[155]). The idea behind multiple MD runs is to promote infrequent transitions or *rare* events by running several MD simulations from different initial structures and combine them to recover the associated conformational free energy landscape. However, dealing with the resulting flood of data, comprising hundreds or even thousands of simulations, becomes a challenge. Markov State Models (MSMs) arise as an approach to analyze large MD data sets in an objective methodological way to recover thermodynamic and kinetic parameters between conformational states. MSMs are also based on a dimensional reduction (e.g. tICA) to recover the free energy landscape associated with slow collective DOFs and the kinetics of the process. Quantitative predictions from MSMs can be compared with available experimental data.^[156] In particular, this approach has recently been used to study serine protease trypsin^[157] and Bruton tyrosine kinase conformational plasticity.^[158] Besides, MD simulations together with MSMs were also used to guide a regioselective switch in nitrating P450 from *Streptomyces scabies*.^[69b]

(iv) Replica exchange or parallel tempering^[159] is an alternative strategy based on running several copies of the same system at different temperatures and exchanging conformations at certain time intervals. Probability distributions are only meaningful at room temperatures and can be recovered by projecting atomic coordinates onto some selected DOFs, as explained before, whereas high temperatures facilitate barrier crossing. This approach has been widely used for protein folding,^[160] although the number of replicas required to ensure temperature exchanges is proportional to the number of atoms, thus making it unaffordable for large systems.

2.5.2 Biased MD methods

It is possible to increase the frequency with which barriers separating stable states are crossed by introducing external energy potentials into our MD simulations. The selection of the proper bias method can be guided by the amount of structural information that we have about our system. For instance, to study the transition of a protein domain from open (A) to a closed (B) conformation, two main questions can be formulated: do we have enough structural information of A and B to define some DOFs, e.g. dihedral angles, describing the transition? Do we have intermediate structures between the two states? Based on the answer to both questions a proper bias method can be chosen:

(i) When detailed structural knowledge is available of states A, B and also the transition path, independent MD simulations at states A and B together with a spectrum of intermediate conformations can be performed. In umbrella sampling (US),^[161] for example, several MD simulations are computed with restraining bias potentials added at small increments along one or a few preselected DOFs, also referred as collective variables (CVs), forcing the system to sample all the desired conformational states, therefore cancelling the effect of energy barriers and exploring low probability regions. Overlapping umbrella sampling simulations can be analyzed together to recover probability distributions and the free energy within the A to B transition.^[162] This method provides good estimates of the free energy, since each point on the transition is equally sampled, but detailed structural knowledge is required to define a suitable set of starting conformations describing a continuous pathway between A and B. Several methods have been also developed to construct the transition paths between known states and in most cases do not require the definition of CVs, as for instance finite-temperature string method^[163] and transition path sampling.^[164]

(ii) When no clear information about the transition path between A and B states is available, methods that explore all possible transitions between A and B along a small set of CVs is a proper choice. These methods can explore unexpected intermediates and identify novel metastable states. Metadynamics^[165] is a powerful technique to accelerate conformational transitions between metastable states.^[166] As other enhanced sampling techniques, it requires the introduction of low-dimensional descriptors (CVs), whose selection is a critical step. Ideally, they should clearly distinguish between A, B and the intermediates, describe all the slow events that are relevant to the process of interest and their number should not be too large, otherwise a very long simulation time is required.^[167] As mentioned in **section 2.4.1**, in this thesis we relied on a path of collective variables^[168] to explore an allosteric conformational transition. In this regard two CVs are required. The progression along the reference path of conformations generated (thanks to the available X-ray structures) represents one CV, while the distance from the reference path the other CV. This approach alleviates the burden of the high dimensionality. Notice that although the progress along a high-dimensional path is computed, the position of the system along the path is an intrinsically two-dimensional quantity defined by only two CVs values.^[167]

Metadynamics^[165] is based on the addition of small repulsive potentials (Gaussians) to a selected set of CVs at a regular number of MD steps (**Fig. 2.8**), such that the external potential (V_G) acting on the system at time t is given by:

$$V_G(S(x),t) = \omega \sum_{\substack{t' = \tau_G, 2\tau_G, \dots \\ t' < t}} \exp\left(-\frac{(S(x)-s(t'))^2}{2\delta s^2}\right) \quad (2.14)$$

Where $S(x)$ refers to the CVs values as a function of the coordinates and s denotes the value of the CVs. Thus $s(t) = S(x(t))$, which is the value taken by the CV at time t . Three parameters are pivotal to describe external energy, V_G : the Gaussian height (ω), the Gaussian width (δ) and the frequency τ_G with which the Gaussians are added. These parameters have to be carefully selected since they influence the accuracy of the free energy reconstruction. If the Gaussians are too large, the free energy landscape would be explored faster, but the reconstructed profile will be affected by large errors. On the other hand, if the Gaussians are too small or are placed infrequently the reconstruction would be accurate but will take a longer time.^[167]

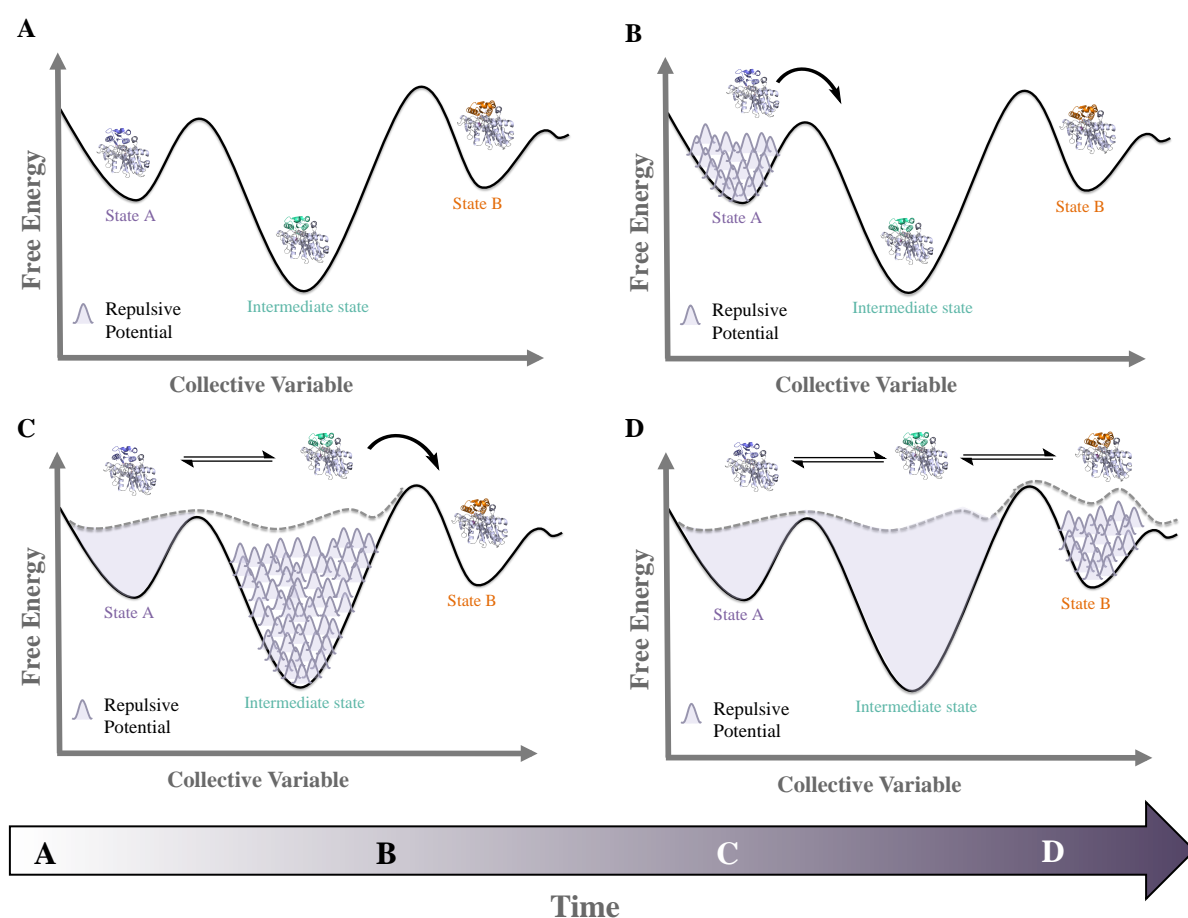


Figure 2.8 Representation of a metadynamics simulation. The repulsive potentials are deposited to the collective variable over time (from **A** to **D**) until the full free energy landscape is covered (**D**).

As shown in **Fig 2.8**, these external potentials discourage the system from visiting prior configurations, forcing it to escape from energy minimum A to explore B through the lowest energy path. In addition to accelerate transitions between states, metadynamics allows to recover the free energy associated with the A to B transition by the sum of all the repulsive potentials added along the MD as a function of the CV values (Equation 2.14). Theoretically, after sufficiently long time metadynamics simulations provide a reliable estimate of the underlying free energy (**Fig. 2.8 D**).

$$\lim_{t \rightarrow \infty} V_G(s, t) \sim F(s) \quad (2.15)$$

Since the initial implementation of metadynamics, many derivatives have been developed. In this thesis we use a well-tempered version^[169] in combination with multiple-walkers approach.^[170] In the well-tempered version the Gaussian height is gradually decreased with

time proportional to a decaying exponential function of the potential deposited in the currently visited point of the CV space.^[166] With this rescaling on the Gaussian height, the bias potential smoothly converges in the long time limit. The so-called bias factor parameter can be selected in order to control how quickly the Gaussian height is decreased. Thus, it has to be carefully chosen in order to efficiently cross the relevant energy barriers in the course of the simulation. The multiple-walkers approach was the first strategy taking advantage of running metadynamics on multiple replicas of the system simultaneously. It was originally developed with the purpose of speeding up free energy calculations using coupled parallel machines.^[167] It is based on running in parallel interacting replicas (walkers) where each walker biases the identical CVs and reads the Gaussian potentials deposited by the others during the simulation. Since all walkers contribute to construct the same metadynamics bias, the free energy landscape is estimated by summing the Gaussian potentials deposited by all walker replicas as a function of the CVs values.

In general, this method usually provides higher accuracy, but can also experience convergence issues since it is not easy to decide when to stop a simulation, avoiding the addition of useless repulsive terms. An intuitive way to decide when the metadynamics simulation is over is the observation of the diffusion of the CV in the entire relevant region, which is an indicator of convergence. Metadynamics has the advantage that not much structural information is required to set up the simulation, although choosing a proper set of CVs can sometimes be tricky. Metadynamics has been widely used to study the conformational landscapes of proteins,³⁴ and the effect of pathogenic mutations in cancer related kinases.^[171]

(iii) Only one conformational state is known (e.g. A) and, therefore, no clear information about the transition is available. In this situation, methods to explore biomolecular conformations without a priori structural knowledge, such as accelerated MD (aMD),^[172] are advantageous. In aMD, a constant bias potential (i.e. boost potential) is added to raise the energy minima while keeping transition states almost unaffected, therefore, smoothing the free energy landscape and enhancing conformational exchanges. aMD has the advantage that it is not necessary the preselection of CVs, becoming really useful when little structural information is available. Nevertheless, a non-trivial post-processing is needed to recover unbiased free energy values. This method has been applied to fold a set of small proteins^[173] and to study the conformational dynamics of biomolecules, such as the maltose binding protein.^[174]

2.6 Residue-by-residue correlation and proximity tools

Residue-by-residue correlation methodologies appear to be an interesting strategy to uncover the connections of the different dynamic regions of the proteins. In this context, these analyses provide a fingerprint of the enzyme motions along the MDs. Some studies have shown that these tools are promising for the underlying characteristics of allosteric pathways between subunits, as for instance in the imidazole glycerol phosphate synthase (IGPS) enzyme,^[110a] but also to characterize the intrinsic enzyme allosteric properties, as in the case of kinases^[110b] and tRNA protein complexes.^[110c] Our group has recently shown that these methods can also be applied to identify mutations found by means of directed evolution (DE) techniques in retroaldolase enzymes.^[21a] The same strategy has been applied in this thesis to the tryptophan synthase enzyme. The python code developed by our group provides a Shortest Path Map (SPM) as an output, which operates as follows:

(i) Generation of the correlation and distance matrix values from MD data; the covariance between the two measures how random these variables change together. In our case it measures how the C_α of each residue deviates from its average position along the MD simulation run. The averaged positions can be obtained by clustering the MD data to obtain the most populated cluster as a representative structure. Thus, the covariance between $C_{\alpha j}$ and $C_{\alpha k}$ for N observations along the MD simulations is estimated as:

$$q_{jk} = \frac{1}{N-1} \sum_{i=1}^N (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k), \quad (2.16)$$

Covariance is sometimes called a measure of “linear dependence” between two random variables. When covariance is normalized the correlation coefficient (Pearson coefficient) is estimated. Correlation can be defined as any class of statistical relationship involving dependence. The Pearson coefficient is expressed as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (2.17)$$

Where X and Y are two random variables (e.g. $C_{\alpha j}$ and $C_{\alpha k}$), cov is the covariance associated with these variables and σ_X and σ_Y the standard deviation of X and Y , respectively. The Pearson

correlation is +1 in the case of a perfect direct linear relationship (correlation), -1 in the case of inverse linear relationship (anticorrelation). As it approaches to 0 there is hardly any relationship between the two (uncorrelated). The correlation matrix is indeed a $n \times m$ matrix whose i and j entry is the correlation between X_i and X_j obtained along the MD simulations. By analogy, the distance matrix entries correspond to the mean distance between X_i and X_j .

(ii) Graph construction. To create a network for our system we first need to specify how the nodes and edges will be created. Typically, one node represents some set of atoms, e.g. an amino acid within a protein. We assign one node to each C_α residue. The edges are defined between pairs of nodes if the mean distance matrix values, $\text{dist}(C_{\alpha i}-C_{\alpha j})$, are shorter than a selected threshold (e.g. $< 6 \text{ \AA}$). Then the length of the edges drawn is weighted according to the correlation matrix values. So that the length of the edge that connects node $_i$ and node $_j$ is estimated as:

$$d_{ij} = -\log|C_{ij}| \quad (2.18)$$

Where C_{ij} is the Pearson correlation coefficient for residues i and j . Larger (absolute) correlation values (closer to 1 or -1) will have shorter edge distances, whereas less correlated residue pairs (values closer to 0) will have edges with long distances (see original network system in **Fig 2.9**).

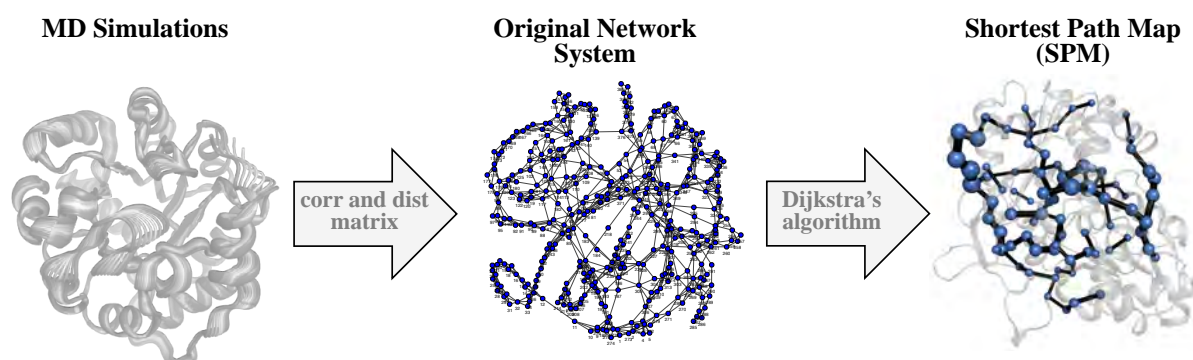


Figure 2.9 Schematic view of the Shortest Path Map (SPM) construction from the MD simulations data set. The blue spheres represent the graph nodes while the lines that connect them the edges.

Once the original graph is generated, we use the Dijkstra algorithm as implemented in igraph module.^[175] The algorithm operates by exploring all possible paths to go from the node of origin (e.g. residue 1) to the rest of nodes through the shortest path in terms of d node distances

obtained from Equation 2.18. When all nodes (i.e. all residues) have been targeted as nodes of origin, the exploration is over. Thus, the width of each edge and the size of each node are proportional to the number of shortest paths passing through that edge or node during the calculation (**Fig 2.9**). Note that not all nodes are included in the SPM. This is due to a threshold selected manually that discards those edges that are explored less frequently by the Dijkstra algorithm. SPM provides a view of the enzyme pathways that most contribute to the protein dynamics in terms of correlated motions, which were found to coincide with many DE mutation positions (see **Chapter 5.1**).

It is worth mentioning that according to the SPM workflow, the selection of the conformational ensemble subjected to the analysis is a relevant step. In this thesis, we performed the SPM analysis on the conformational ensemble obtained through the metadynamics simulations. In contrast with conventional MDs, that in some cases may not sample the complete conformational exchange under study due to the associated time-scale of the transition, metadynamics simulations has the advantage to statistically sample the targeted conformational exchange. Thus, the proper selection of the conformational ensemble provides a more accurate view of the protein dynamics in the SPM graph. No further adjustments were made to the original formulation of the SPM method due to the constraints imposed in the metadynamics conformational ensemble.

Chapter 3. Objectives

The **major goal** of this thesis is the characterization of the protein conformational energy landscape by means of computational methods and investigating its connection with enzyme properties such as enantioselectivity, catalytic activity, thermoadaptation and allostery. In general, we aim to rationalize the novel function achieved in laboratory-evolved enzymes through the exploration of the protein conformational dynamics and exploit this information to rationally design promising enzyme variants. The objectives of this thesis are divided into two main blocks encompassing **specific objectives** regarding the nature and the properties of the system studied:

I. Alcohol dehydrogenases (ADH) studies (Chapter 4):

- Explore the conformational dynamics of a zinc dependent ADH enzyme using the bonded model protocol for metalloproteins to analyze the conformational population distribution as a function of the enantioselectivity and catalytic activity.
- Perform an in-depth structural analysis of the major conformational states found in order to investigate the molecular basis of the enantioselectivity control.
- Rationalize the reversion of enantioselectivity and thermoadaptation at ambient temperatures towards non-natural induced by DE mutations in the laboratory-evolved ADH enzyme variants.

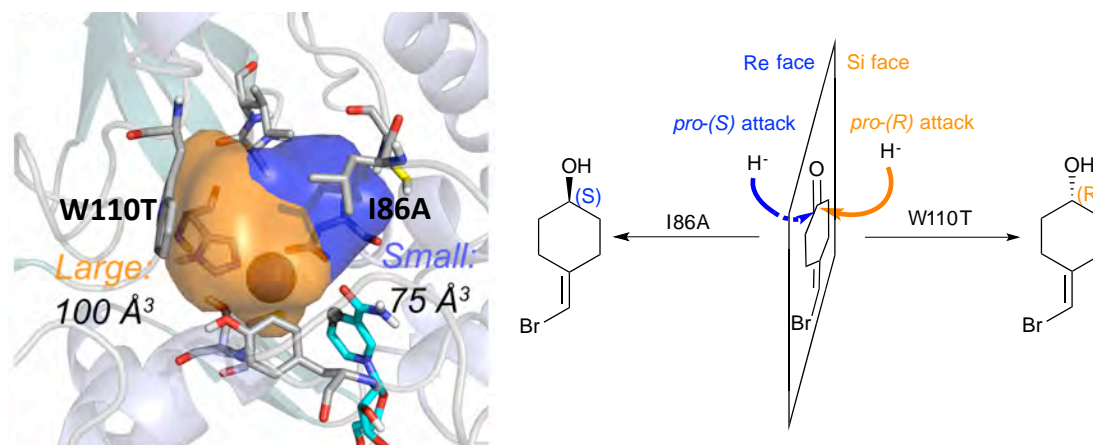
II. Tryptophan synthase (TrpS) studies (Chapter 5):

- Reconstruct the free energy landscape (FEL) of the TrpS enzyme (formed by the two subunits: TrpA, TrpB) associated with an allosteric transition using enhanced sampling techniques (metadynamics) in order to decipher the conformational ensemble of TrpS, TrpB in absence of the TrpA protein partner and laboratory-evolved stand-alone TrpB enzyme variants.
- Rationalize the loss of activity in the absence of the TrpB protein partner (i.e. TrpA) and the recovery of activity induced by the DE mutations achieved in laboratory-evolved stand-alone enzyme variants.

- Test the predictive power of correlation-based tools (Shortest Path Map, SPM) for the identification of DE mutations introduced in TrpB for stand-alone enzyme variants.
- Develop new computational strategies for the rational design of TrpB stand-alone enzyme variants using SPM correlation-based tools in combination with bioinformatic computational tools.

Chapter 4. Enantioselectivity and thermoadaptation properties of alcohol dehydrogenase (ADH) enzymes

4.1 Exploring the reversal of enantioselectivity on a zinc-dependent alcohol dehydrogenase



Maria-Solano, M. A.; Romero-Rivera, A.; Osuna, S.* Exploring the reversal of enantioselectivity on a Zinc-dependent Alcohol Dehydrogenase, *Org. Biomol. Chem.* **2017**, *15*, 4122-4129. [Chemistry, Organic, 3.564, Q1]. <https://doi.org/10.1039/C7OB00482F>

Abstract

Alcohol Dehydrogenase (ADH) enzymes catalyse the reversible reduction of prochiral ketones to the corresponding alcohols. These enzymes present two differently shaped active site pockets, which dictate their substrate scope and selectivity. In this study, we computationally evaluate the effect of two commonly reported active site mutations (I86A, and W110T) on a secondary alcohol dehydrogenase from *Thermoanaerobacter Brockii* (TbSADH) through Molecular Dynamics simulations. Our results indicate that the introduced mutations induce dramatic changes on the shape of the active site, but most importantly they impact the substrate-enzyme interactions. We demonstrate that the combination of Molecular Dynamics simulations with the tools POVME and NCIplot correspond to a powerful strategy for rationalising and engineering the stereoselectivity of ADH variants.



Cite this: *Org. Biomol. Chem.*, 2017, **15**, 4122

Exploring the reversal of enantioselectivity on a zinc-dependent alcohol dehydrogenase†

Miguel A. Maria-Solano, Adrian Romero-Rivera and Sílvia Osuna *

Alcohol Dehydrogenase (ADH) enzymes catalyse the reversible reduction of prochiral ketones to the corresponding alcohols. These enzymes present two differently shaped active site pockets, which dictate their substrate scope and selectivity. In this study, we computationally evaluate the effect of two commonly reported active site mutations (I86A, and W110T) on a secondary alcohol dehydrogenase from *Thermoanaerobacter brockii* (TbSADH) through Molecular Dynamics simulations. Our results indicate that the introduced mutations induce dramatic changes in the shape of the active site, but most importantly they impact the substrate–enzyme interactions. We demonstrate that the combination of Molecular Dynamics simulations with the tools POVME and NCIplot corresponds to a powerful strategy for rationalising and engineering the stereoselectivity of ADH variants.

Received 26th February 2017,
Accepted 9th April 2017

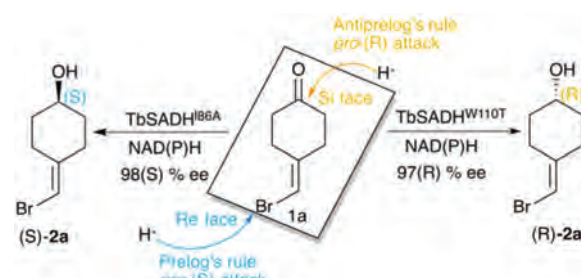
DOI: 10.1039/c7ob00482f

rsc.li/obc

1. Introduction

Biocatalysis is based on the application of natural catalysts for new purposes, for which the enzymes were not designed. The advantages of biocatalysts with respect to traditional catalysts make enzyme-based routes a preferable alternative for the synthesis of optically active compounds.¹ The asymmetric reduction of prochiral ketones to yield optically pure alcohols can be achieved with metal-based catalysts,² but also with enzymes such as alcohol dehydrogenases (ADH). Many studies have been reported in the literature showing the importance of ADH in asymmetric synthesis,^{3–5} of relevance is their usually high thermostability,^{6,7} and the ability to operate in non-aqueous media with high activity and selectivity.^{8,9}

ADH enzymes catalyse the reversible reduction of prochiral ketones to their corresponding alcohols. They require the presence of NAD(P)H as a cofactor, which delivers its *pro*-(R) hydride to the usually *Re* face of the ketone yielding the corresponding (S)-alcohols (see Scheme 1). The stereoselectivity of ADHs towards the formation of (S)-alcohols mainly arises from the shape of the active site of the enzymes that usually present a small and a large binding pocket (see Fig. 1).¹⁰ As most ADH follow Prelog's rule (Scheme 1), the engineering of their active sites for the formation of the (R)-enantiomer, *i.e.* anti-Prelog ADHs, is of great interest. In addition to that, the expansion of the substrate scope of ADH is also highly appealing for broad-



Scheme 1 Representation of Prelog and anti-Prelog rules for the studied substrate **1a**, together with the stereoselectivity of the engineered variants TbSADH^{I86A}, and TbSADH^{W110T} by Reetz *et al.*²⁰

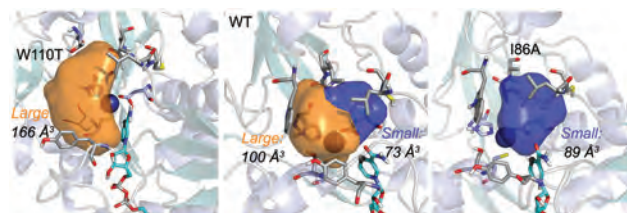


Fig. 1 Volume representation of the small and large TbSADH binding pockets for the WT enzyme (middle), W110T (left), and I86A (right) variants. These calculations have been performed with POVME 2.0.³¹

ening their applicability in asymmetric synthesis. To that end, Directed Evolution (DE)^{1,11–14} and rational site-specific mutagenesis¹⁵ have been applied in some ADH enzymes. Reetz *et al.* developed a powerful strategy for reducing the number of variants to screen by generating a collection of small but 'smart' enzyme libraries.¹⁶ This was applied on the zinc-depend-

Institut de Química Computacional i Catalàlisi (IQCC) and Departament de Química, Universitat de Girona, Carrer Maria Aurèlia Capmany 6, 17003 Girona, Spain.
E-mail: silvia.osuna@udg.edu

† Electronic supplementary information (ESI) available. See DOI: 10.1039/C7OB00482F



dent secondary ADH from *Thermoanaerobacter brockii* (TbSADH) for the asymmetric reduction of tetrahydrofuran-3-one towards the (S)-alcohol, which is of importance for the synthesis of the HIV inhibitors amprenavir and fosamprenavir.^{16,17} Engineered variants of *Lactobacillus kefir* short-chain alcohol dehydrogenase were also developed for the asymmetric reduction of the same tetrahydrofuran-3-one, but also for the related thiolan-3-one.¹⁸ Phillips and coworkers engineered TbSADH for accepting several structurally diverse ketones.¹⁹ Similarly, Reetz evolved the same ADH for accepting a set of non-cyclic ketones.⁶ They also engineered TbSADH for the catalytic asymmetric reduction of prochiral ketones of type 4-alkylidene cyclohexanone with formation of the corresponding axially chiral (R) or (S)-alcohols.²⁰ Interestingly, the singly mutated variants TbSADH^{W110T} and TbSADH^{I86A} were found to yield respectively either the unusual (R)-alcohol or the (S)-alcohol with high conversion rates and selectivity. The same W110 and I86 positions were found to be important in determining the enantioselectivity of the highly homologous secondary ADH from *Thermoanaerobacter ethanolicus* (TeSADH) enzyme.^{8,21–23}

The previously mentioned examples highlight the outstanding performance of laboratory-evolution for enhancing activity, and reversing the enantioselectivity of ADHs. Complementary to experimental evolution, computational methods can be used for rationalizing the activity and selectivity of natural and laboratory-engineered enzymes.²⁴ Bocola and coworkers elucidated through Quantum Mechanics and Molecular Mechanics (*i.e.* QM/MM) calculations the mechanism of hydride and proton transfer of the oxidoreductase from *Candida Parapsilosis*.²⁵ Electronic structure calculations and Molecular Dynamics (MD) simulations were performed to investigate the mechanism of liver alcohol dehydrogenase (LADH).²⁶ The calculations revealed a lower activation barrier for the hydride transfer step if alcohol deprotonation occurs first. Many computational studies have been devoted to elucidate the fundamental nature of hydrogen tunnelling that occurs in these NAD(P)H-dependent enzymes.^{27–29} Some of us explored through MD simulations of the Michaelis–Menten and transition state-bound complexes the stereoselectivity of some *Lactobacillus kefir* short-chain alcohol dehydrogenases.¹⁸ These simulations allow rationalising the effect of active site mutations on the selectivity of this Zn(II) free ADH enzyme.

In this study we computationally evaluate the effect of W110 and I86A active site mutations on a series of zinc-dependent TbSADH variants²⁰ through MD simulations. We demonstrate that the introduced mutations induce dramatic changes in the shape of the enzyme active site, which affect the substrate–enzyme interactions thus determining the stereoselectivity of the TbSADH variants.

2. Results and discussion

ADH enzymes present two differently shaped active site pockets, which are responsible for their substrate scope and

selectivities (see Fig. 1). By introducing mutations to the ADH active site, both *pro*-(R) and *pro*-(S) selectivities can be obtained. In most experimental studies based on TbSADH and the homologous TeSADH published so far two positions, namely I86 and W110, have been found to be key for either enhancing the enzyme activity towards bulky substrates and/or reverting the stereoselectivity of ADHs.^{9,20,21,23,30} In order to shed some light into the role of the latter mutations in ADH catalytic activity and selectivity, we performed MD simulations on the Wild-Type (WT) TbSADH enzyme, and the variants TbSADH^{W110T} and TbSADH^{I86A}. We restricted our study to the analysis of the prochiral ketone of type 4-alkylidene cyclohexanone (**1a**, see Scheme 1) studied by Reetz and coworkers.²⁰ This ketone is especially challenging as the steric preferences of the carbon atoms surrounding the carbonyl group are identical. Of importance is the fact that positions I86 and W110 are key to revert the enzyme enantioselectivity even with this non-conventional substrate.

We carried out five independent 200 ns MD simulations (*i.e.* accumulated simulation time of 1 microsecond) in both *pro*-(R) and *pro*-(S) conformations of **1a** in the WT TbSADH, TbSADH^{W110T}, and TbSADH^{I86A} enzyme variants. In order to maintain the substrate **1a** bound to the Zn(II) metal ion, a force constant was applied. This approach allows us to analyse the positioning of **1a** for efficient hydride transfer, and thus explain the activity and origin of enantioselectivity observed experimentally.

As shown in Scheme 1, **1a** has a bromide atom that can be differently oriented in the small and large binding pockets depending on the variant and the starting pose (*pro*-(R) and *pro*-(S) conformation, see Fig. 2–5). The positioning of both the

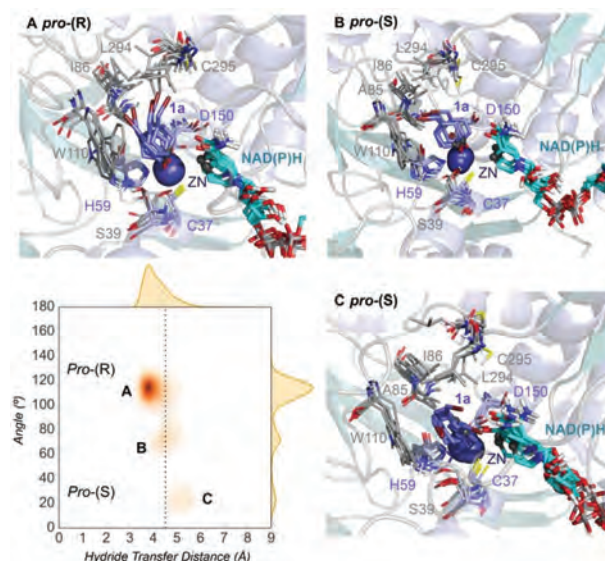


Fig. 2 Representation of some representative snapshots of the different conformational states sampled along the MD simulations for TbSADH starting from the *pro*-(R) orientation of **1a**. The histogram of the hydride transfer distance together with the *pro*-(R)/*pro*-(S) angle (detailed in Fig. S1†) is displayed.



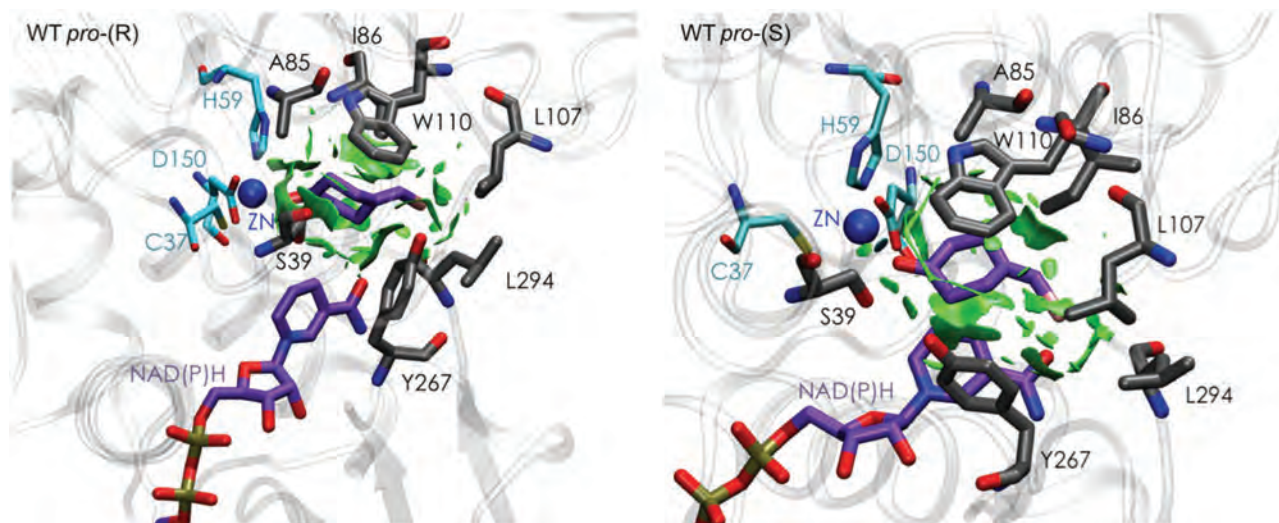


Fig. 3 Representation of the non-covalent interactions for the *pro*-(S) and *pro*-(R) conformations of **1a** in the active site pocket of TbSADH enzyme, computed with the computational tool NCIplot.^{34,35}

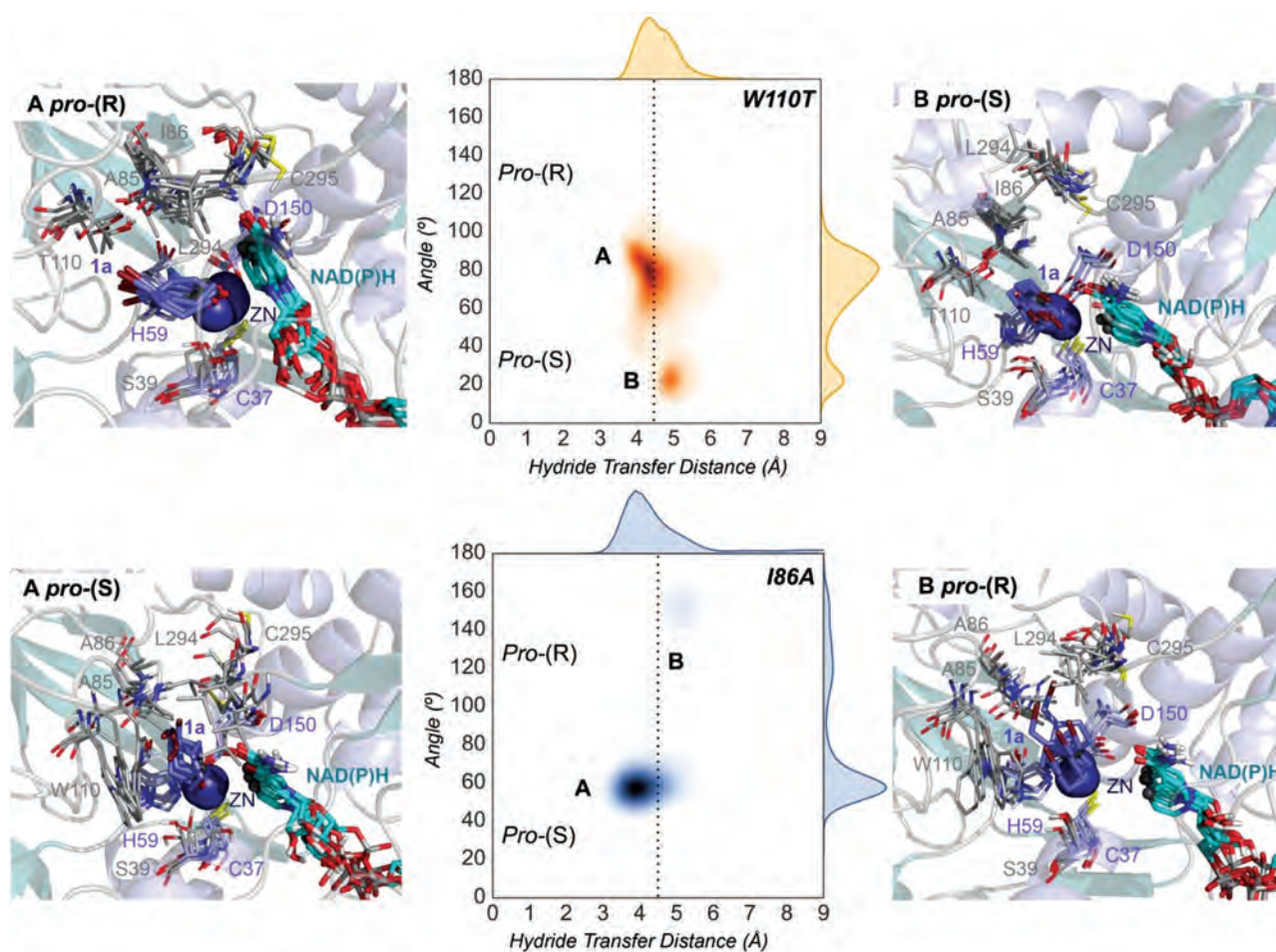


Fig. 4 Representation of some representative snapshots of the different conformational states sampled along the MD simulations for the TbSADH^{W110T} and TbSADH^{I86A} starting from the *pro*-(R) (in orange) and *pro*-(S) (in blue) orientations of **1a**, respectively. The histogram of the hydride transfer distance together with the *pro*-(R)/*pro*-(S) angle (detailed in Fig. S1†) is displayed for both variants.



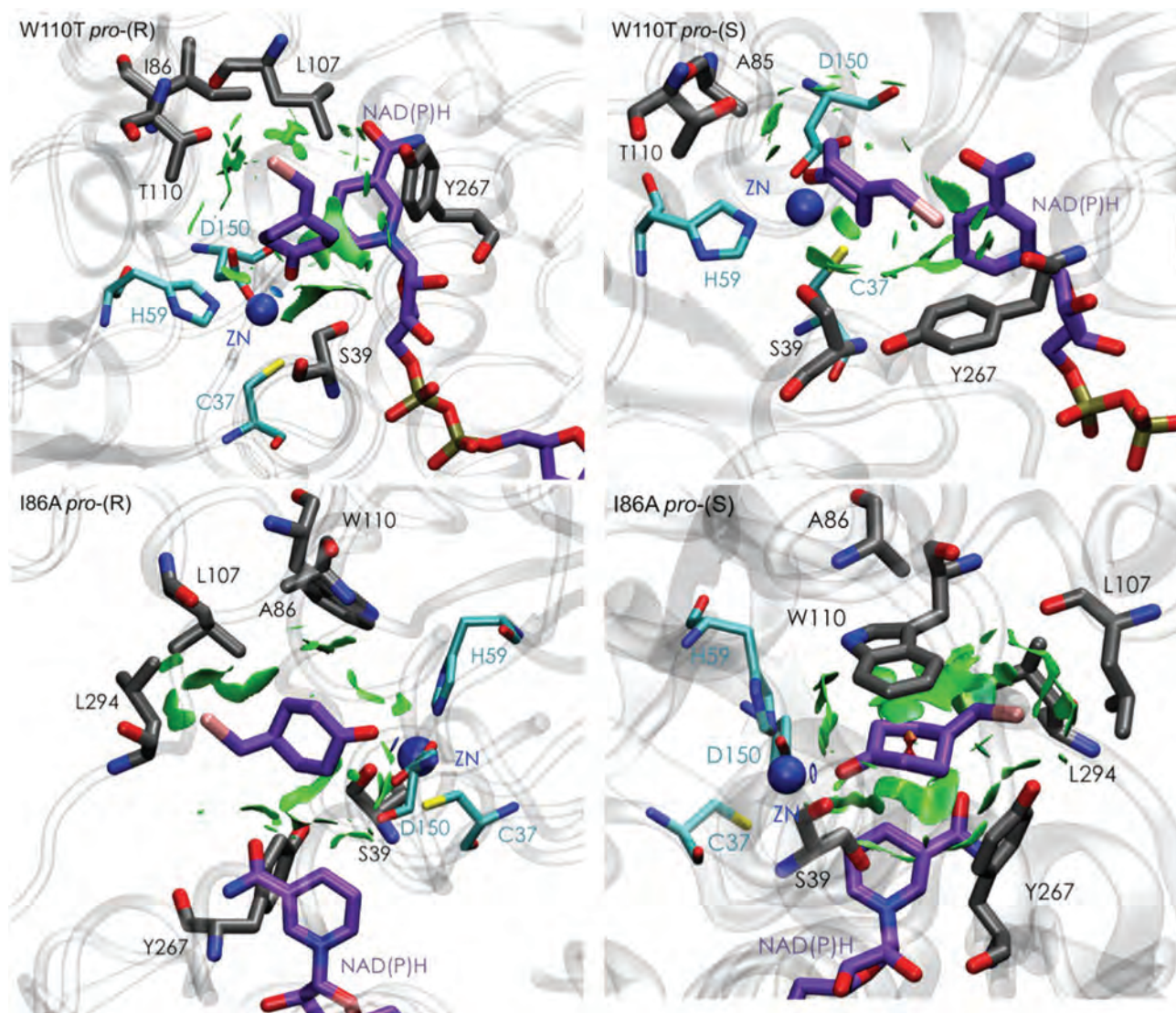
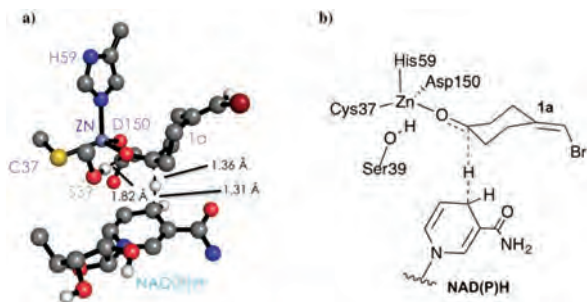


Fig. 5 Representation of the non-covalent interactions for *pro*-(S) and *pro*-(R) conformations of **1a** in the active site pocket of TbSADH^{W110T} (top) and TbSADH^{I86A} (down) enzymes, computed with the computational tool NCIPLOT.^{34,35}

bromide and the cyclohexane ring of **1a** and its interactions with the enzyme active site pocket will dictate ADH selectivity. The difference in activity can be predicted by analysing the distance between the carbonyl group of the substrate and the NAD(P)H carbon atom involved in the hydride transfer (see Scheme 2). Our computed hydride transfer Transition State (TS) using DFT and a small subset of the enzyme active site residues, *i.e.* following the *theozyme* approach,³² indicates that at the TS it is *ca.* 2.7 Å for both axial and equatorial attacks (see Scheme 2, and Fig. S8†). This is in line with previous calculations for the hydride transfer. As with classical MD simulations we cannot model the bond-breaking/forming hydride transfer step, we instead evaluate the active site preorganization towards the *pro*-(S)/*pro*-(R) enzyme–substrate complexes to shed some light into the enzyme enantioselectivities. We define as catalytically competent poses those MD conformations that present hydride transfer distances shorter than 4.5 Å, whereas those orientations with longer hydride distances were defined as non-catalytic. This allows us to indirectly quantify the number of reactive events along the simulation time, *i.e.* it provides an estimate of the enzyme catalytic activity.

We evaluate ADH enantioselectivity preferences by comparing the angle formed between the carbonyl group of the active site residue T38 (situated next to C37, one of the Zn(II)-coordinating residues), the C5 and C3 carbon of **1a** cyclohexane ring (see Fig. S1†) in all variants. As the NAD(P)H cofactor is in some cases displaced from the active site, the angle provided by the rigid T38 residue together with the measure of the hydride transfer distance allow us to better evaluate the catalytically competent *pro*-(S) and *pro*-(R) conformations. As done in previous studies,¹⁸ by computing the relative populations of





Scheme 2 DFT optimized TS structure for the hydride transfer step. For visualization purposes, non-polar hydrogen atoms are hidden.

the reactive *pro*-(S) and *pro*-(R) poses observed along the MD simulations, the experimental enantiomeric excess ratios can be estimated (see Table S2†).

Evaluation of the TbSADH WT enzyme stereoselectivity

Our analysis starts with the evaluation of the WT TbSADH enzyme activity and selectivity towards **1a**. Reetz *et al.* reported that TbSADH is able to produce the corresponding (R)-alcohol in a 95% conversion, but only with 66 (R) % ee.²⁰ We evaluated the WT enzyme active site pockets in the most populated conformational states (*i.e.* most visited along the MD simulations) using the computational tool POVME,³¹ indicating that the small and large active site pockets have an approximated volume of *ca.* 73 Å³ and 100 Å³ (see Fig. 1 and Table S1†), which evidence their drastic difference in size as observed with X-ray structures.³³

In our TbSADH MD simulations starting from the *pro*-(R) orientation of **1a**, the bromide atom is forced to fit in the small pocket because the bulky W110 residue does not allow the rotation of the substrate towards the large binding pocket (see Fig. 2, *pro*-(R) A). This corresponds to the most populated conformation, where **1a** remains properly positioned for the hydride transfer to occur towards its Si-face and thus allowing the (R)-alcohol formation. The average hydride distance is *ca.* 3.9 Å, which coincides with the computed hydride transfer distance at the reactant complex (*i.e.* 3.8 Å).²⁷ This rather short distance is in agreement with the high conversion rate observed experimentally. The analysis of non-covalent interactions with the NCI plot of **1a** in the *pro*-(R) conformation reveals stabilizing C–H... π interactions between H59, Y267, and W110 with the cyclohexane ring of the substrate (see Fig. 3). In contrast, the latter stabilizing interactions are much weaker in the *pro*-(S) conformations (in particular non-covalent interactions with the residue W110), which evidence how the TbSADH pocket is more complementary to the *pro*-(R) conformation of **1a** to produce the corresponding (R)-alcohol.

In the MD simulations starting from the *pro*-(S) conformation of **1a**, short catalytic distances of *ca.* 3.9 Å are also observed (see Fig. S2 A†), where **1a** is properly positioned for the formation of the (S)-alcohol. However, this *pro*-(S) catalytically active conformation has a quite low population. This rather low stability of the *pro*-(S) conformation is also evi-

denced by analysing the non-covalent interactions of **1a** and the active site pocket of TbSADH. The enzyme also adopts some intermediate conformations that present substantially longer unproductive hydride transfer distances. Overall, our MD simulations on TbSADH starting from both *pro*-(R) and *pro*-(S) orientations of **1a** indicate that the formation of the (R)-alcohol is substantially preferred, although some catalytically competent *pro*-(S) conformations are also explored. This is in line with the 66% (R) ee observed in the experimental assays.

Evaluation of the TbSADH^{W110T} and TbSADH^{I86A} enzyme stereoselectivity

The substitution of W110 by threonine makes the enzyme large binding pocket even wider. The computed volume is *ca.* 166 Å³, whereas for the TbSADH it was 100 Å³ (as discussed previously). This mutation therefore gives extra space to **1a** for a better accommodation of the bromide substituent in the enzyme active site pocket, and thus allows the substrate to rotate towards the large binding site. Experimentally, it was found that TbSADH^{W110T} was able to convert **1a** into the corresponding (R)-alcohol with high conversion rates and high enantioselectivities (99% conversion, and 97 (R) % ee).²⁰ In this enzyme variant, angles of *ca.* 70° are observed for the *pro*-(R) conformation, whereas *ca.* 20° for the *pro*-(S) attack (see Fig. 4, W110T A and B).

In our MD simulations starting from the *pro*-(R) orientation of **1a**, the substrate rapidly rotates to position the bromide into the large binding pocket, and remains in this *pro*-(R) orientation most of the simulation time (see Fig. 4, W110T). The NAD(P)H cofactor is perfectly positioned to deliver the hydride and allow the (R)-alcohol formation (see Fig. 4, W110T *pro*-(R) A) displaying catalytically competent hydride distances and angles. Moreover, starting from *pro*-(S) orientations (Fig. S3†) **1a** rapidly rotates towards *pro*-(R) conformations.

The analysis of non-covalent interactions in the *pro*-(R) conformations of **1a** reveals stabilizing C–H... π interactions between the cyclohexane ring of the substrate and residues H59, and Y267, but also with the nicotinamide ring of the NAD(P)H cofactor (see Fig. 5, W110T *pro*-(R)). The W110T mutation enlarges the active site pocket, but also allows the formation of stabilizing interactions between the bromide and the side-chains of L107 and the newly introduced T110 residue. We also observe during the MD simulations that the substrate can rotate to explore *pro*-(S) conformations (see Fig. 4, W110T *pro*-(S) B), however long hydride distances are observed due to the displacement of the NAD(P)H cofactor, which interacts with the bromide atom of the substrate (see Fig. 5, W110T *pro*-(S)).

We finally evaluated the TbSADH^{I86A} enzyme variant, which was found to allow the formation of the opposite (S)-alcohol in high enantiomeric excess (98 (S) % ee), and conversion (95%). Our volume calculations on the most populated conformational states indicate that the small enzyme active site pocket is enlarged from *ca.* 73 to 89 Å³. In contrast to what we observe in the TbSADH and TbSADH^{W110T} variants, MD simulations



starting from the *pro*(S) poses of **1a** reveal that the substrate stays in the *pro*(S) conformations with an angle of *ca.* 60° most of the simulation time (see Fig. 4, I86A *pro*(S) A). In this most populated state, catalytically competent hydride transfer distances are sampled (*ca.* 4 Å), which fits with the high activity of the variant observed experimentally. This favourable *pro*(S) conformations are mainly stabilized by C–H... π interactions between the cyclohexane ring and residues W110, H59, and Y267 (see Fig. 5, I86A *pro*(S)). As observed in the case of TbsADH^{W110T}, C–H... π interactions are also observed within the cyclohexane ring and the nicotinamide ring of NAD(P)H. The mutation introduced at position 86 (*i.e.* I86A) creates additional space in the small binding pocket, which is occupied by the indole ring of W110. This new conformation of W110 maximizes the C–H... π interactions with the cyclohexane ring of **1a**, and thus favors the *pro*(S) attack (see Fig. 5, I86A *pro*(S)).

In the MD simulations, when **1a** rotates to explore *pro*(R) conformations, long hydride transfer distances are observed due to the displacement of the NAD(P)H cofactor (see Fig. 4, I86A *pro*(R) B). MD simulations starting from the *pro*(R) conformation (Fig. S4†) show that the substrate stays most of the time in the *pro*(R) orientation, but again leads to the displacement of the NAD(P)H cofactor and thus results in a non-catalytic configuration. The analysis of non-covalent interactions in this *pro*(R) conformation reveals that most of the above mentioned interactions with W110, H59, and Y267 are lost (see Fig. 5, I86A *pro*(R)). These results point out that although **1a** can adopt both *pro*(R) and *pro*(S) orientations, *pro*(S) is the catalytically competent pose as only with this orientation both **1a** and NAD(P)H are properly positioned for the catalysis.

3. Conclusions

Our MD simulations indicate that the poor selectivity of the WT TbsADH enzyme is due to the possible positioning of the substrate in both *pro*(R) and *pro*(S) orientations. The *pro*(R) conformation is, however, substantially favoured due to stronger non-covalent interactions between the substrate and the enzyme active site. TbsADH^{W110T} presents a substantially wider active site, especially the large binding pocket, which allows the substrate to explore *pro*(R) conformations with catalytically active hydride transfer distances. In the *pro*(R) conformation, C–H... π interactions are observed between the cyclohexane ring and active site residues H59 and Y267. The introduced threonine residue at position 110 also allows the formation of stabilizing interactions between its side-chain and the bromide group of **1a**. TbsADH^{I86A} enzyme variant shows a significantly different behaviour revealing a highly pre-organized active site for the *pro*(S) conformation with catalytically efficient distances. The introduced I86A mutation enlarges the small binding pocket, and induces a conformational change in W110 that optimally positions the indole group for enhanced C–H... π interactions with the cyclohexane ring of the substrate. The combination of MD simulations, *theozyme* calculations,

and in-depth analysis of the active site pocket through the computational tools POVME and NCIPLOT allows us to rationalise the effect of these two key active site mutations in the enantioselectivity of the zinc-dependent TbsADH enzyme. Given that many studies based on TbsADH and TeSADH target the same active site mutations, we believe that the obtained results are rather general. Our results also highlight the feasibility of MD simulations, coupled with POVME and NCIPLOT calculations for the engineering of natural enzyme active sites for enhanced activity and selectivity.

Computational methods

MD simulations in explicit water were performed using AMBER 16 package⁴ and starting from the PDB structure: 1YKF.³³ The W110T and I86A variants were generated using the mutagenesis tool included in PyMOL (<http://www.pymol.org>). Parameters for substrate **1a** for the MD simulations were generated within the *antechamber* module of AMBER 16 using the general AMBER force field (GAFF),³⁶ with partial charges set to fit the electrostatic potential generated at the B3LYP/6-31G(d) level by the restrained electrostatic potential (RESP) model.³⁷ The charges were calculated according to the Merz–Singh–Kollman scheme^{38,39} using Gaussian 09.⁴⁰ Amino acid protonation states were predicted using the H++ server (<http://biophysics.cs.vt.edu/H++>).⁴¹ We have used the bonded model for Zn and the residues of the first coordination sphere, in particular we used the Seminario approach⁴² to obtain the metal parameters needed for the simulation as implemented in Prof. Ryde program.⁴³ The optimization, frequencies and charge calculations to obtain the parameters were done at the B3LYP/6-31G(d) level using Gaussian 09.⁴⁰ The parameters for NAD(P)H were extracted from previous studies by Prof. Ryde.^{44,45} The WT enzyme (PDB: 1YKF) and variant were solvated in a pre-equilibrated truncated cuboid box with a 10 Å buffer of TIP3P⁴⁶ water molecules using the AMBER16 *leap* module, resulting in the addition of *ca.* 11 000 solvent molecules. The system was neutralized by the addition of explicit counterions (Na⁺ and Cl⁻). All calculations were done using the *ff14SB* Amber force field.⁴⁷ A two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on the solute by a harmonic potential with a force constant of 500 kcal mol⁻¹ Å⁻², and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems are gently heated using six 50 ps steps, incrementing the temperature 50 K each step (0–300 K) under constant volume and periodic boundary conditions. Water molecules were treated with the SHAKE algorithm such that the angle between the hydrogen atoms is kept fixed. Long-range electrostatic effects were modeled using the particle-mesh-Ewald method.⁴⁸ An 8 Å cutoff was applied to Lennard-Jones and electrostatic interactions. Harmonic restraints of 10 kcal mol⁻¹ were applied to the solute, and the Langevin equilibration scheme was used to control and equalize the temperature. The time step was maintained at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each



system was then equilibrated without restraints for 2 ns with a 2 fs time step at a constant pressure of 1 atm and a temperature of 300 K. After the systems were equilibrated in the NPT ensemble, 3 independent five hundred nanosecond MD simulations were performed under the NVT ensemble and periodic-boundary conditions.

The *theozyme* calculations for the hydride transfer step were performed at the B3LYP/6-31G(d) level of theory using Gaussian 09.⁴⁰ Active site volume calculations were performed with the computational tool POVME 2.0.³¹

Acknowledgements

A. R. R. thanks the Generalitat de Catalunya for a PhD fellowship (2015-FI-B-00165), M. A. M. S. is grateful to the Spanish MINECO for a PhD fellowship (BES-2015-074964). S. O. thanks the Spanish MINECO for project CTQ2014-59212-P, Ramón y Cajal contract (RYC-2014-16846), the European Community for CIG project (PCIG14-GA-2013-630978), and the funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC-2015-StG-679001). We are grateful for the computer resources, technical expertise, and assistance provided by the Barcelona Supercomputing Center – Centro Nacional de Supercomputación.

Notes and references

- U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore and K. Robins, *Nature*, 2012, **485**, 185–194.
- R. H. Morris, *Chem. Soc. Rev.*, 2009, **38**, 2282–2291.
- E. García-urdiales, I. Alfonso and V. Gotor, *Chem. Rev.*, 2005, **105**, 313–354.
- W. Kroutil, H. Mang, K. Edegger and K. Faber, *Curr. Opin. Chem. Biol.*, 2004, **8**, 120–126.
- Y.-G. Zheng, H.-H. Yin, D.-F. Yu, X. Chen, X.-L. Tang, X.-J. Zhang, Y.-P. Xue, Y.-J. Wang and Z.-Q. Liu, *Appl. Microbiol. Biotechnol.*, 2017, **101**, 987–1001.
- Z. Sun, G. Li, A. Ilie and M. T. Reetz, *Tetrahedron Lett.*, 2016, **57**, 3648–3651.
- D. S. Burdette, V. Tchernajenko and J. G. Zeikus, *Enzyme Microb. Technol.*, 2000, **27**, 11–18.
- M. M. Musa, K. I. Ziegelmann-fjeld, C. Vieille, J. G. Zeikus and R. S. Phillips, *Angew. Chem., Int. Ed.*, 2007, **46**, 3091–3094.
- M. M. Musa, K. I. Ziegelmann-fjeld, C. Vieille and R. S. Phillips, *Org. Biomol. Chem.*, 2008, **6**, 887–892.
- V. Prelog, in *Pure Appl. Chem*, 1964, vol. 9, p. 119.
- A. S. Bommarius, *Annu. Rev. Chem. Biomol. Eng.*, 2015, **6**, 319–345.
- M. T. Reetz, *Directed Evolution of Selective enzymes: Catalysts for Organic Chemistry and Biotechnology*, Wiley-VCH, Weinheim, 2016.
- A. Currin, N. Swainston, P. J. Day and D. B. Kell, *Chem. Soc. Rev.*, 2015, **44**, 1172–1239.
- S. Lutz and U. T. Bornscheuer, *Protein Engineering Handbook*, Wiley-VCH Verlag GmbH & Co. KGaA, 2008.
- J. Pleiss, in *Enzyme Catalysis in Organic Synthesis*, Wiley-VCH Verlag GmbH & Co. KGaA, 2012, pp. 89–117.
- Z. Sun, R. Lonsdale, A. Ilie, G. Li, J. Zhou and M. T. Reetz, *ACS Catal.*, 2016, **6**, 1598–1605.
- A. Nobili, M. G. Gall, I. V. Pavlidis, M. L. Thompson, M. Schmidt and U. T. Bornscheuer, *FEBS J.*, 2013, **280**, 3084–3093.
- E. L. Noey, N. Tibrewal, G. Jiménez-osés, S. Osuna, J. Park, C. M. Bond, D. Cascio, J. Liang, X. Zhang, G. W. Huisman, Y. Tang and K. N. Houk, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, E7065–E7072.
- C. M. Nealon, M. M. Musa, J. M. Patel and R. S. Phillips, *ACS Catal.*, 2015, **5**, 2100–2114.
- R. Agudo, G.-D. Roiban and M. T. Reetz, *J. Am. Chem. Soc.*, 2012, **135**, 1665–1668.
- M. M. Musa, N. Lott, M. Laivenieks, L. Watanabe, C. Vieille and R. S. Phillips, *ChemCatChem*, 2009, **1**, 89–93.
- K. I. Ziegelmann-fjeld, M. M. Musa, R. S. Phillips, J. G. Zeikus and C. Vieille, *Protein Eng., Des. Sel.*, 2007, **20**, 47–55.
- M. M. Musa, K. I. Ziegelmann-fjeld, C. Vieille, J. G. Zeikus and R. S. Phillips, *J. Org. Chem.*, 2007, **72**, 30–34.
- A. Romero-rivera, M. Garcia-borras and S. Osuna, *Chem. Commun.*, 2017, **53**, 284–297.
- G. V. Dhoke, M. D. Davari, U. Schwaneberg and M. Bocola, *ACS Catal.*, 2015, **5**, 3207–3215.
- P. K. Agarwal, S. P. Webb and S. Hammes-schiffer, *J. Am. Chem. Soc.*, 2000, **122**, 4803–4812.
- D. Roston and A. Kohen, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 9572–9577.
- S. R. Billeter, S. P. Webb, P. K. Agarwal, T. Iordanov and S. Hammes-schiffer, *J. Am. Chem. Soc.*, 2001, **123**, 11262–11272.
- D. Roston and A. Kohen, *J. Am. Chem. Soc.*, 2013, **135**, 13624–13627.
- M. M. Musa, J. M. Patel, C. M. Nealon, C. S. Kim, R. S. Phillips and I. Karume, *J. Mol. Catal. B: Enzym.*, 2015, **115**, 155–159.
- J. D. Durrant, L. Votapka, J. Sørensen and R. E. Amaro, *J. Chem. Theory Comput.*, 2014, **10**, 5047–5056.
- D. J. Tantillo, C. Jiangang and K. N. Houk, *Curr. Opin. Chem. Biol.*, 1998, **2**, 743–750.
- Y. Korkhin, A. J. Kalb, M. Peretz, O. Bogin, Y. Burstein and F. Frolow, *J. Mol. Biol.*, 1998, **278**, 967–981.
- J. Contreras-García, E. R. Johnson, S. Keinan, R. Chaudret, J.-P. Piquemal, D. N. Beratan and W. Yang, *J. Chem. Theory Comput.*, 2011, **7**, 625–632.
- E. R. Johnson, S. Keinan, P. Mori-Sánchez, J. Contreras-garcía, A. J. Cohen and W. Yang, *J. Am. Chem. Soc.*, 2010, **132**, 6498–6506.
- J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.



- 37 C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10269–10280.
- 38 U. C. Singh and P. A. Kollman, *J. Comput. Chem.*, 1984, **5**, 129–145.
- 39 B. H. Besler, K. M. Merz and P. A. Kollman, *J. Comput. Chem.*, 1990, **11**, 431–439.
- 40 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 09, Revision D.01*, Gaussian Inc., Wallingford, CT, 2009.
- 41 R. Anandakrishnan, B. Aguilar and A. V. Onufriev, *Nucleic Acids Res.*, 2012, **40**, W537–W541.
- 42 J. M. Seminario, *Int. J. Quantum Chem.*, 1996, **60**, 1271–1277.
- 43 L. Hu and U. Ryde, *J. Chem. Theory Comput.*, 2011, **7**, 2452–2463.
- 44 U. Ryde, *Proteins*, 1995, **21**, 40–56.
- 45 U. Ryde, *Protein Sci.*, 1995, **4**, 1124–1132.
- 46 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.
- 47 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712–725.
- 48 T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.



Exploring the reversal of enantioselectivity on a Zinc-dependent Alcohol Dehydrogenase

Miguel A. Maria-Solano,^a Adrian Romero-Rivera^a, Sílvia Osuna^{*a}

SUPPORTING INFORMATION

COMPUTATIONAL METHODS WITH FULL REFERENCES

Molecular Dynamics Simulations. MD simulations in explicit water were performed using AMBER 16 package¹ and starting from the PDB structure: 1YKF.² The I86A and W110T variants were generated using the mutagenesis tool included in PyMOL (<http://www.pymol.org>). Parameters for substrate **1a** for the MD simulations were generated within the *antechamber* module of AMBER 16 using the general AMBER force field (GAFF),³ with partial charges set to fit the electrostatic potential generated at the B3LYP/6-31G(d) level by the restrained electrostatic potential (RESP) model.⁴ The charges were calculated according to the Merz-Singh-Kollman scheme^{5, 6} using Gaussian 09.⁷ Amino acid protonation states were predicted using the H++ server (<http://biophysics.cs.vt.edu/H++>).⁸ We have used the bonded model for Zn and the residues of the first coordination sphere, in particular we used the Seminario approach⁹ to obtain the metal parameters needed for the simulation as implemented in Prof. Ryde program.¹⁰ The optimization, frequencies and charge calculations to obtain the parameters was done at the B3LYP/6-31G(d) level using Gaussian 09.⁷ The parameters for NAD(P)H were extracted from previous studies by Prof. Ryde.^{11, 12} The Wild-Type (WT) enzyme (PDB: 1YKF) and variants were solvated in a pre-equilibrated truncated cuboid box with a 10-Å buffer of TIP3P¹³ water molecules using the AMBER16 *leap* module, resulting in the addition of *ca.* 9,000 solvent molecules. The system was neutralized by addition of explicit counterions (Na⁺ and Cl⁻). All calculations were done using a modification of the *ff99SB* force field (*ff14SB*).¹⁴ A two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on solute by a harmonic potential with a force constant of 500 kcal mol⁻¹ Å⁻², and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems are gently heated using six 50-ps steps, incrementing the temperature 50 K each step (0–300 K) under constant volume and periodic boundary conditions. Water molecules were treated with the SHAKE algorithm such that the angle between the hydrogen atoms is kept fixed. Long-range electrostatic effects were modeled using the particle-mesh-Ewald method.¹⁵ An 8-Å cutoff was applied to Lennard-Jones and electrostatic interactions. Harmonic restraints of 10 kcal/mol were applied to the solute, and the Langevin equilibration scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was then equilibrated without restraints for 2 ns with a 2-fs timestep at a constant pressure of 1 atm and temperature of 300 K. After the systems were equilibrated in the NPT ensemble, 3 independent five hundred nanosecond MD simulations were performed under the NVT ensemble and periodic-boundary conditions.

The *theozyme* calculations for the hydride transfer step were performed at the B3LYP/6-31G(d) level of theory using Gaussian 09.⁷ Active site volume calculations were performed with the computational tool POVME 2.0.¹⁶

Table S1. Volume calculated (\AA^3) on the different variants on the small and big pocket without **1a** in the 3 most populated clusters using POVME¹⁶.

Pockets	WT TbSADH						TbSADH ^{I86A}						TbSADH ^{W110T}					
	Prelog			Anti-Prelog			Prelog			Anti-Prelog			Prelog			Anti-Prelog		
	C ₀	C ₁	C ₂	C ₀	C ₁	C ₂	C ₀	C ₁	C ₂	C ₀	C ₁	C ₂	C ₀	C ₁	C ₂	C ₀	C ₁	C ₂
Small	78	62	68	77	79	72	91	86	84	93	98	83	-	-	-	-	-	-
Large	97	99	83	104	104	115	-	-	-	-	-	-	197	145	154	174	168	155

Table S2. Calculation of the %ee of *pro*-(R) and *pro*-(S) conformations. *Calculation of the conformations taking into account distances lower than 4.5 \AA (closer the catalytic distance) and their corresponding angles are used to classify the *pro*-(R) and *pro*-(S) conformations. $C_{R/S}$ is the productive number of *pro*-(R) and *pro*-(S) conformations, N is the total number of frames in the MD simulation.

$$\%ee = \left(\frac{P_R - P_S}{P_R + P_S} \right) \times 100 \quad \begin{aligned} P_R &= \frac{C_R}{N} \\ P_S &= \frac{C_S}{N} \end{aligned} \quad \text{eq. 1}$$

Variants		R*	S*	% ee (R)	% ee (S)
WT TbSADH	<i>Pro</i> -(S)	0.36	0.20	29	-
	<i>Pro</i> -(R)	0.68	0.08	79	-
TbSADH ^{I86A}	<i>Pro</i> -(S)	0.02	0.51	-	92
	<i>Pro</i> -(R)	0.69	0.16	62	-
TbSADH ^{W110T}	<i>Pro</i> -(S)	0.33	0.12	47	-
	<i>Pro</i> -(R)	0.45	0.03	88	-

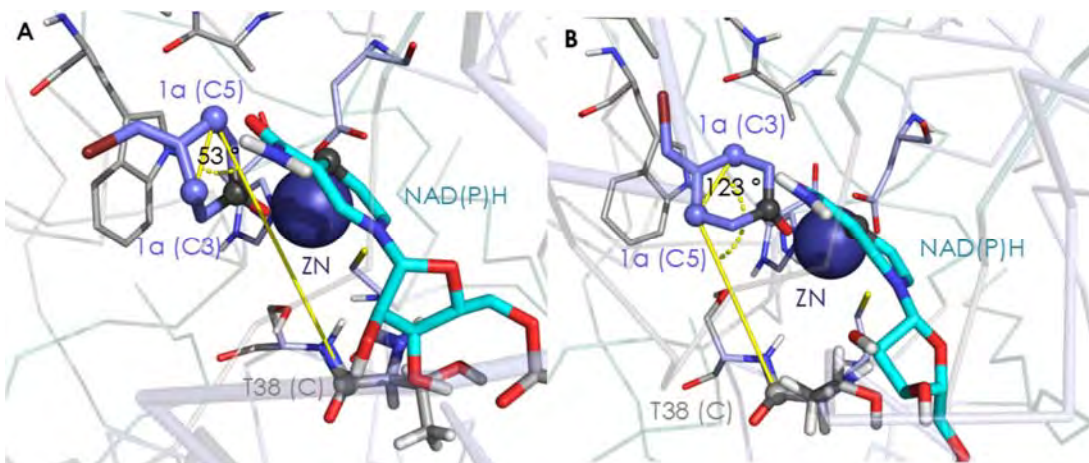


Figure S1. Representation of the selected angle between T38 (C), 1a (C5) and 1a (C3) for the determination of *pro*-(S) (A) and *pro*-(R) (B) orientations. The atoms involved in the angle and in the hydride transfer are shown in spheres.

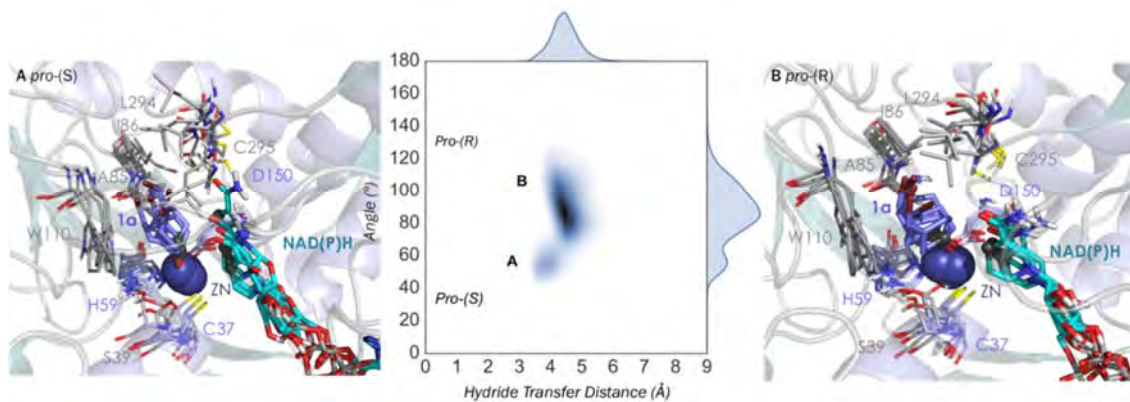


Figure S2. Representation of some representative snapshots of the different conformational states sampled along the MD simulations for TbSADH starting from the *pro-S* orientation of **1a**. The histogram of the hydride transfer distance together with the *pro-R*/*pro-S* angle (detailed in Figure S1) is displayed.

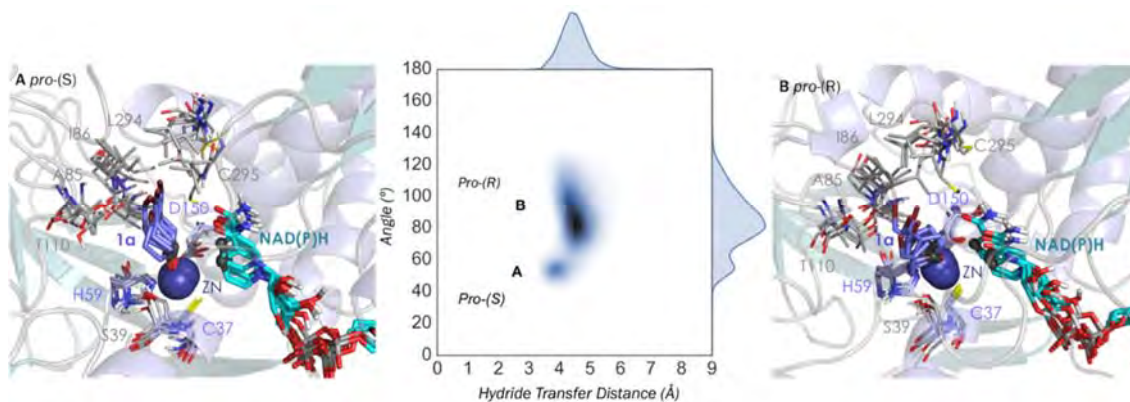


Figure S3. Representation of some representative snapshots of the different conformational states sampled along the MD simulations for the TbSADH^{W110T} starting from the *pro-S* (in blue) orientations of **1a**. The histogram of the hydride transfer distance together with the *pro-R*/*pro-S* angle (detailed in Figure S1) is displayed.

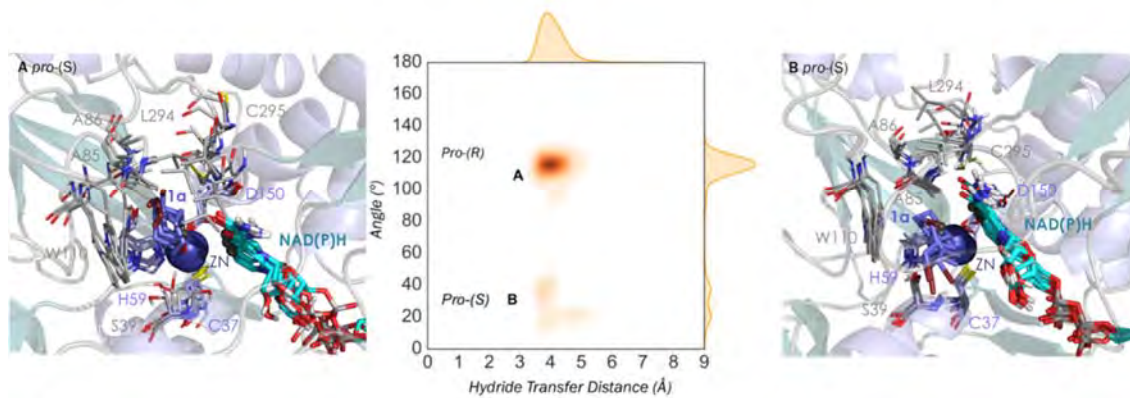


Figure S4. Representation of some representative snapshots of the different conformational states sampled along the MD simulations for TbSADH^{I86A} starting from the *pro*-(R) (in orange) orientations of **1a**. The histogram of the hydride transfer distance together with the *pro*-(R)/*pro*-(S) angle (detailed in Figure S1) is displayed.

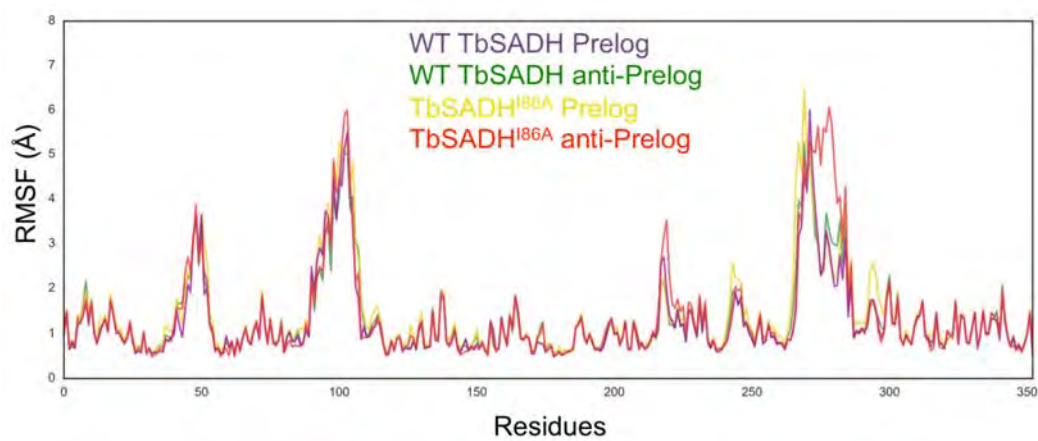
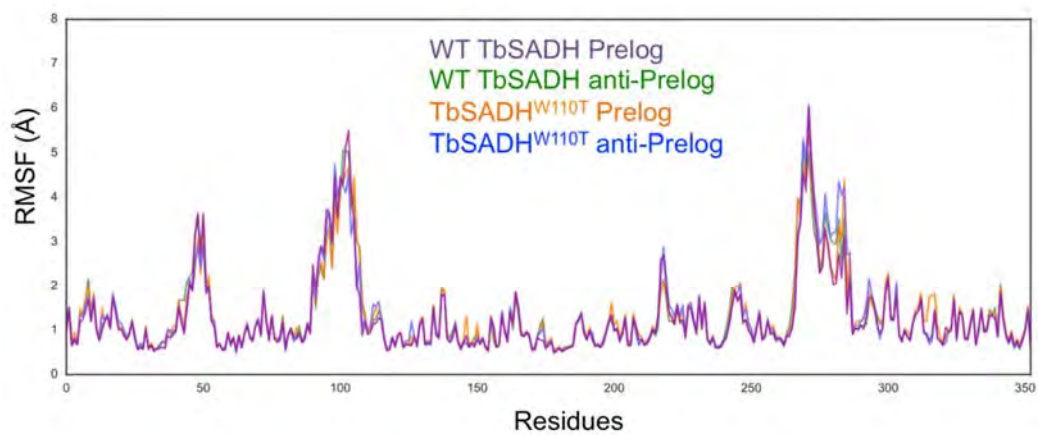


Figure S5. Root Mean Square Fluctuation (RMSF, in Å) along the microsecond timescale MD simulations.

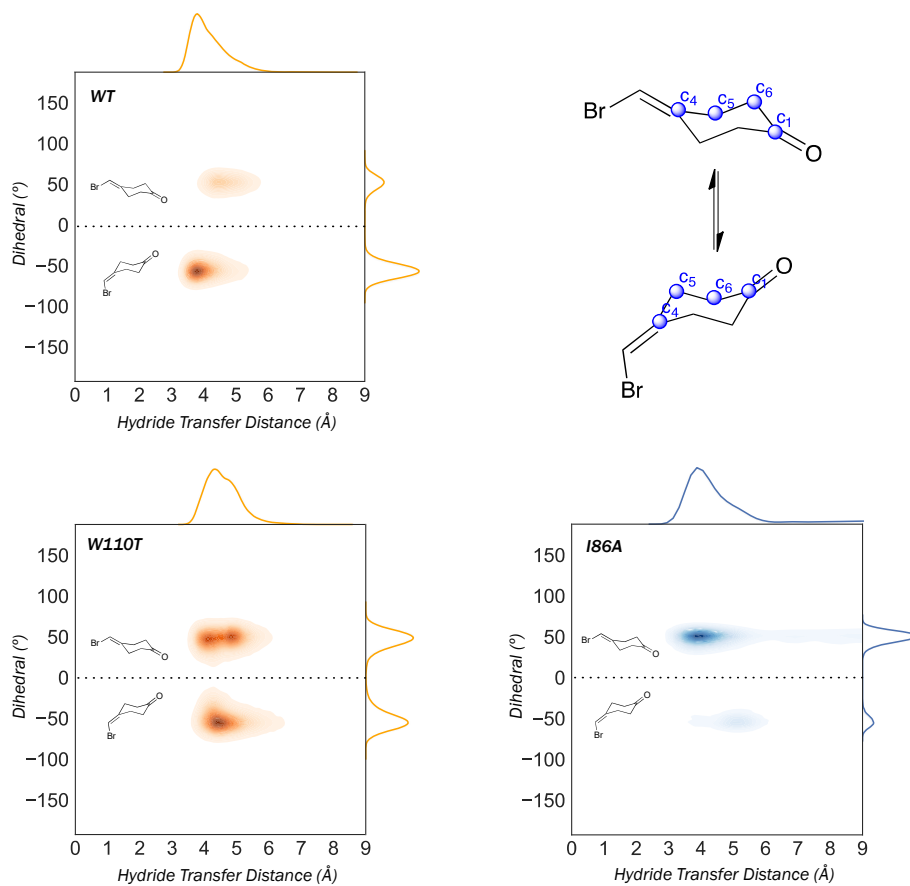


Figure S6. Representation of **1a** conformations sampled along the MD simulations for the TbSADH^{W110T}, TbSADH and TbSADH^{I86A}. The histogram of the hydride transfer distance together with the dihedral of the chair of **1a** (C1, C6, C5, C4) is displayed.

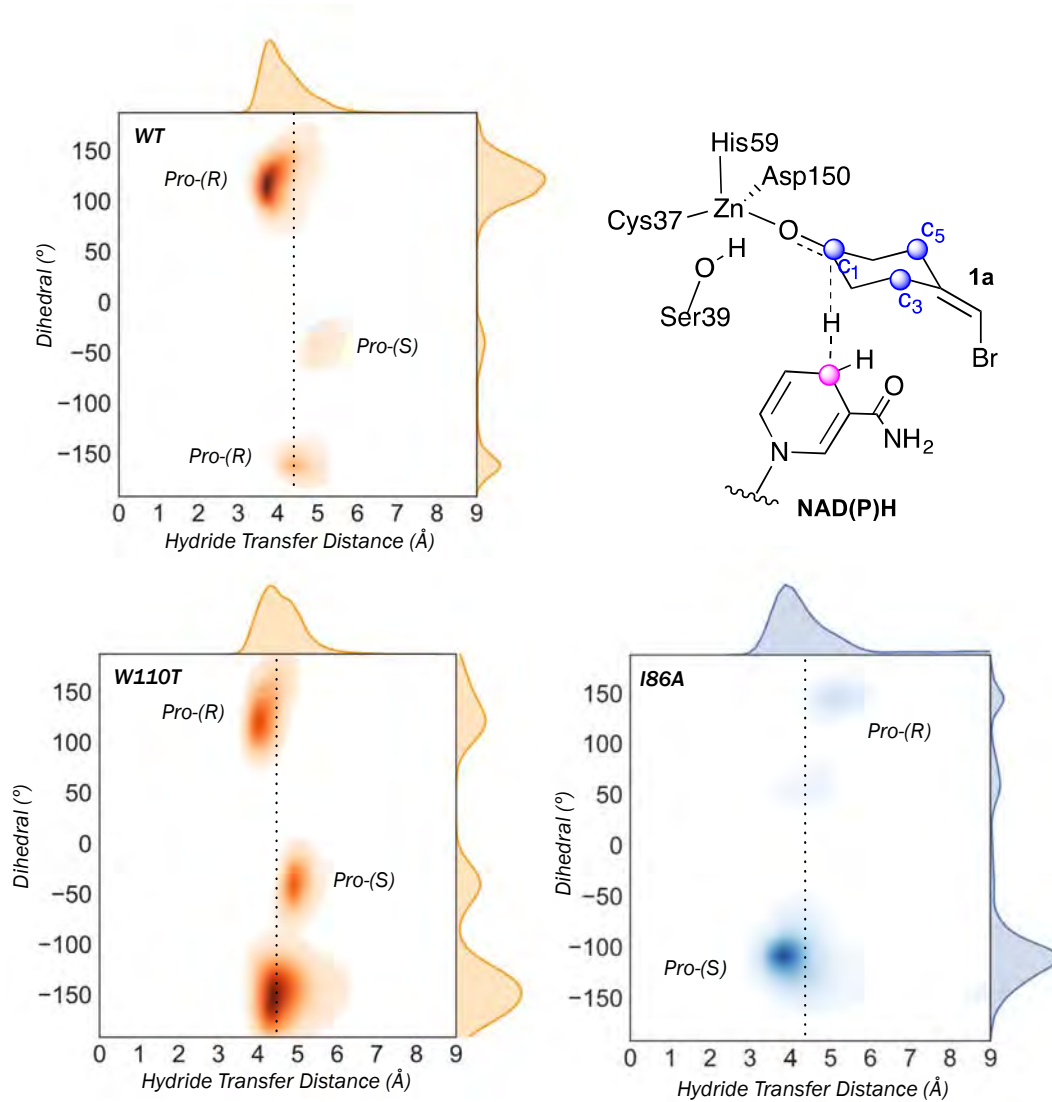


Figure S7. Representation of the *pro-(R)/pro-(S)* conformations sampled along the MD simulations for WT TbSADH, TbSADH^{W110T}, and TbSADH^{I86A}. The histogram of the hydride transfer distance together with the dihedral of **1a** (C3, C5, C1) and **C** (NAD(P)H, *i.e.* hydride transfer carbon) is displayed. Atoms used to calculate the dihedral angle are shown in spheres.

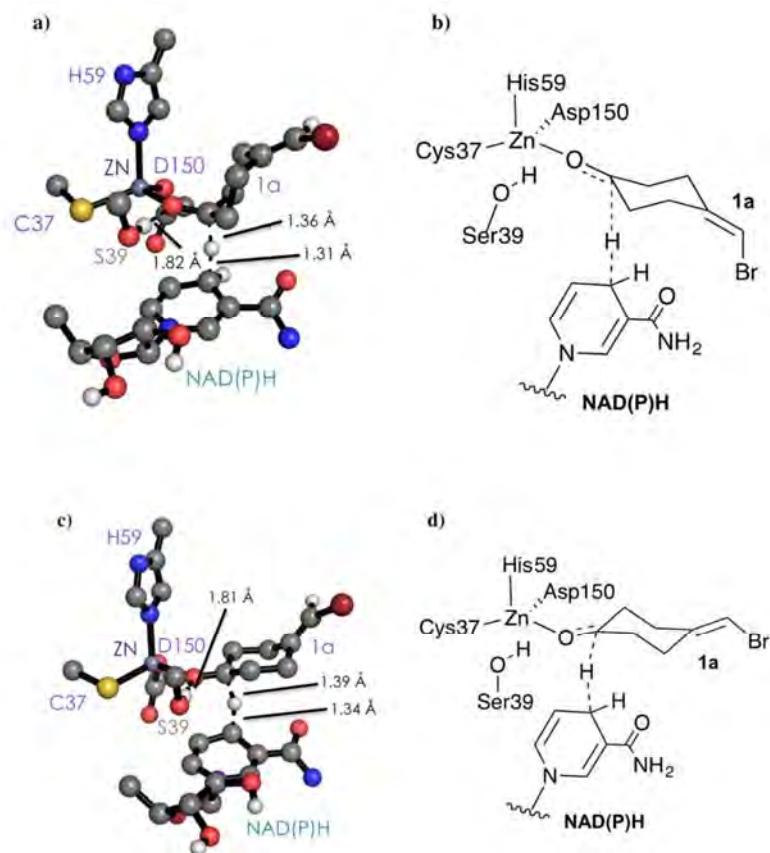


Figure S8. TS calculations for equatorial (a) and axial (c) attacks, followed by their respective chemdraw representations.

Dataset S1. Optimized Cartesian coordinates for the TS corresponding to the equatorial attack.

```

C 1.787437 1.125009 1.433691
C 0.015505 1.967248 -0.183447
C 1.122809 2.568934 -1.088241
C 2.894615 1.747238 0.542756
H -0.838897 1.657653 -0.792277
H 1.476635 1.853329 2.190716
H 2.170144 0.234133 1.938962
H 1.387609 1.825490 -1.853071
H 0.742288 3.452516 -1.606558
H 3.275205 0.970779 -0.135035
H 3.733844 2.068696 1.168569
H -0.318991 2.726421 0.531164
C 2.352647 2.898518 -0.280793
C 0.578201 0.763950 0.572228
C 2.952933 4.087893 -0.219666
O 0.581346 -0.371383 -0.087695
Br 2.388460 5.642809 -1.200481
H 3.827454 4.296858 0.383700
C 1.337269 -5.247790 -0.621626
S 0.210596 -3.785782 -0.722782
C -0.872518 -0.465519 -3.256821
O -1.473963 -0.641159 -1.980598

```


C	6.946485	-1.435586	-1.363926
C	5.477721	-1.701110	-1.352834
C	4.546963	-1.690375	-0.345452
N	4.759984	-2.037201	-2.489192
C	3.464276	-2.213680	-2.150074
N	3.298887	-2.010788	-0.851535
C	2.755433	-2.574335	4.270317
C	1.821436	-2.543972	3.063093
O	0.624535	-2.838857	3.192478
O	2.399601	-2.181645	1.958585
N	-3.750303	-0.095630	0.998582
C	-3.489546	1.110464	1.579755
C	-2.357536	1.332747	2.316805
C	-1.324152	0.299188	2.379866
C	-1.764437	-1.019127	1.946210
C	-2.909803	-1.162566	1.230498
C	-2.075935	2.637565	2.974554
N	-3.141798	3.443769	3.253770
O	-0.913187	2.979713	3.216944
C	-4.924850	-0.200903	0.097054
C	-4.641117	0.374531	-1.314182
C	-5.601567	-0.463551	-2.170026
C	-5.553070	-1.835896	-1.482626
C	-4.496909	-2.784499	-2.036125
O	-4.849593	1.764386	-1.407051
O	-6.874407	0.171331	-2.029336
O	-5.278064	-1.545139	-0.073702
H	-3.508582	-2.313094	-2.066781
H	-4.436724	-3.681205	-1.411138
H	-5.732816	0.363004	0.581066
H	-3.601113	0.166014	-1.583104
H	-5.299072	-0.512115	-3.221237
H	-6.541811	-2.304256	-1.509488
H	-5.810234	1.869444	-1.548005
H	-7.456974	-0.130036	-2.743517
H	-0.376227	0.630674	1.533465
H	-0.679058	0.333259	3.259464
H	-4.236209	1.877120	1.407530
H	-4.066920	3.055710	3.375455
H	-2.943270	4.286888	3.777629
H	-1.150990	-1.883066	2.177741
H	-3.259836	-2.109878	0.845674
H	-1.664506	-0.537294	-4.008084
H	-0.391866	0.518453	-3.359174
H	-0.123479	-1.241105	-3.471628
H	-4.768892	-3.096070	-3.050938
H	7.198306	-0.588643	-2.013107
H	7.287507	-1.200280	-0.352668
H	7.510860	-2.305773	-1.719306
H	5.138852	-2.136542	-3.422390
H	-0.764386	-0.580118	-1.299600
H	2.308655	-5.029591	-1.074908
H	1.498531	-5.555106	0.416099
H	0.883700	-6.083743	-1.161826
H	4.683076	-1.487739	0.705664
H	2.689781	-2.489611	-2.849811
H	3.144690	-1.567976	4.463005

H 2.232927 -2.934338 5.159132
H 3.617603 -3.218485 4.066375
Zn 1.534876 -2.112381 0.196976

Dataset S2. Optimized Cartesian coordinates for the TS corresponding to the axial attack

N 4.653884 -2.347811 -2.466331
N 3.213009 -2.195351 -0.817006
C 1.164852 -5.461163 -0.321457
C -0.895516 -0.885553 -3.303573
C 6.822327 -1.515092 -1.457850
C 5.368599 -1.846952 -1.390258
C 4.450716 -1.760820 -0.374846
C 3.372545 -2.542822 -2.085272
C 2.635038 -2.661780 4.318216
C 1.711581 -2.629025 3.103192
O -1.522512 -0.912624 -2.028384
O 0.506219 -2.892778 3.227441
O 2.305988 -2.301919 1.997152
S 0.119748 -3.969492 -0.640993
Zn 1.447057 -2.282020 0.236115
H -1.677177 -1.008361 -4.059227
H -0.381688 0.068198 -3.494031
H -0.169108 -1.701707 -3.423772
H 7.025552 -0.746396 -2.212786
H 7.160452 -1.136878 -0.489893
H 7.426262 -2.395091 -1.708774
H 5.025580 -2.542566 -3.387368
H -0.823913 -0.797690 -1.342589
H 2.171799 -5.331520 -0.729308
H 1.246968 -5.668546 0.749693
H 0.704799 -6.327347 -0.805607
H 4.594064 -1.428554 0.641848
H 2.602668 -2.935619 -2.732325
H 3.066512 -1.667896 4.483028
H 2.092199 -2.972660 5.213314
H 3.469658 -3.348595 4.139973
N -3.553408 0.137370 0.962291
N -2.988358 3.827792 2.918280
C -3.291436 1.392947 1.421827
C -2.136381 1.693675 2.095243
C -1.078713 0.689182 2.194281
C -1.521746 -0.670154 1.919383
C -2.694875 -0.893682 1.269072
C -1.899649 3.030968 2.703476
C -4.769569 -0.070285 0.137630
C -4.551635 0.305425 -1.349258
C -5.578360 -0.609532 -2.032185
C -5.526970 -1.879920 -1.169943
C -4.546185 -2.939353 -1.656119
O -0.754143 3.410146 2.973010
O -4.734843 1.676344 -1.615120
O -6.825868 0.079193 -1.912396
O -5.140817 -1.419966 0.166374
H -3.546495 -2.519409 -1.810929
H -4.475929 -3.747565 -0.921047
H -5.547672 0.566018 0.578945

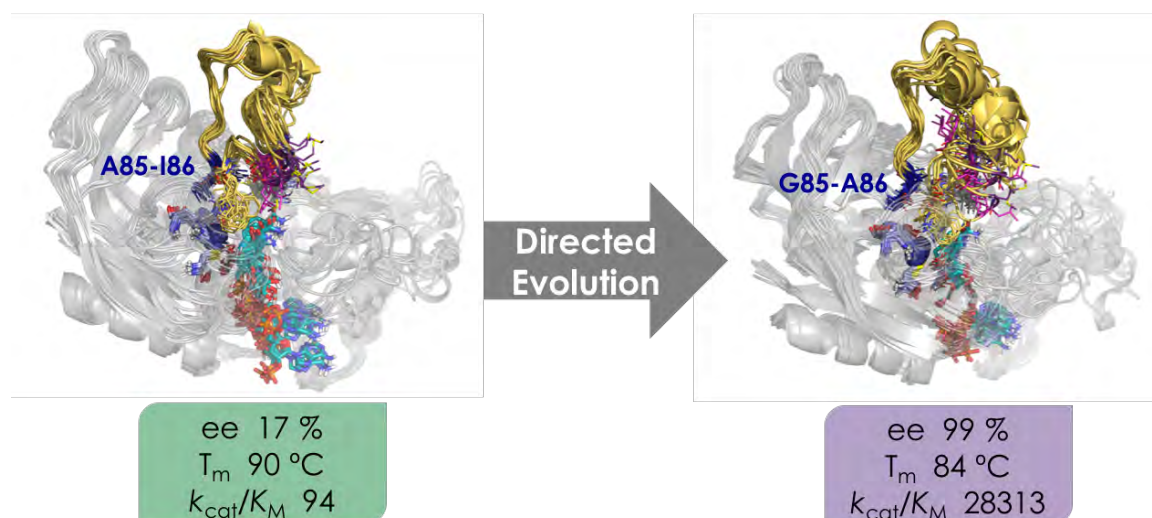
H	-3.533369	0.032050	-1.642442
H	-5.336981	-0.807986	-3.081435
H	-6.531319	-2.303309	-1.069745
H	-5.699347	1.789478	-1.718690
H	-7.451777	-0.288966	-2.554956
H	-0.157006	0.921432	1.254308
H	-0.393335	0.820223	3.032631
H	-4.061618	2.128246	1.220565
H	-3.908570	3.427880	3.039049
H	-2.815819	4.689607	3.420673
H	-0.908606	-1.511739	2.226072
H	-3.051803	-1.877960	1.001003
H	-4.898042	-3.369468	-2.600653
C	0.314843	1.763893	-0.766831
C	0.749232	0.687008	0.231558
C	2.069186	0.963273	0.951861
C	2.221863	2.420667	1.426356
C	1.908935	3.390576	0.305343
C	0.535154	3.212038	-0.287643
C	2.821251	4.286471	-0.071039
O	0.508604	-0.544446	-0.145952
Br	2.558589	5.580799	-1.471522
H	2.186986	0.264718	1.786633
H	0.912719	1.586914	-1.673253
H	-0.731659	1.600059	-1.045813
H	1.515102	2.611597	2.244311
H	3.232496	2.576803	1.817614
H	-0.196429	3.454396	0.492475
H	0.370451	3.902922	-1.118263
H	2.867069	0.728376	0.231653
H	3.800580	4.385869	0.380189

References:

1. D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin and P. A. Kollman, *AMBER 16, University of California, San Francisco, 2016*.
2. Y. Korkhin, A. J. Kalb, M. Peretz, O. Bogin, Y. Burstein and F. Frolow, *J. Mol. Biol.*, 1998, **278**, 967-981.
3. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157-1174.
4. C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10269-10280.
5. U. C. Singh and P. A. Kollman, *J. Comput. Chem.*, 1984, **5**, 129-145.
6. B. H. Besler, K. M. Merz and P. A. Kollman, *J. Comput. Chem.*, 1990, **11**, 431-439.
7. G. W. T. M. J. Frisch, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa,

- M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, *Inc.: Wallingford, CT*, 2009.
8. R. Anandakrishnan, B. Aguilar and A. V. Onufriev, *Nucleic Acids Res.*, 2012, **40**, W537-W541.
 9. J. M. Seminario, *Int. J. Quantum Chem.*, 1996, **60**, 1271-1277.
 10. L. Hu and U. Ryde, *J. Chem. Theory Comput.*, 2011, **7**, 2452-2463.
 11. U. Ryde, *Proteins*, 1995, **21**, 40-56.
 12. U. Ryde, *Protein Sci.*, 1995, **4**, 1124-1132.
 13. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926-935.
 14. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712-725.
 15. T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089-10092.
 16. J. D. Durrant, L. Votapka, J. Sørensen and R. E. Amaro, *J. Chem. Theory Comput.*, 2014, **10**, 5047-5056.

4.2 Inducing high activity of a thermophilic enzyme at ambient temperature by directed evolution



Li, G.; Maria-Solano, M. A., Romero-Rivera, A., Osuna, S.*, Reetz, M.* Inducing High Activity of a Thermophilic Enzyme At Ambient Temperature by Directed Evolution, *Chem. Commun.* **2017**, 53, 9454-9457. [Chemistry, Multidisciplinary, 6.319, Q1]. <https://doi.org/10.1039/C7CC05377K>

The work included in this chapter has been carried out in collaboration with an experimental group led by Manfred Reetz. The directed evolution strategy for the generation of the evolved variant was performed by the Reetz group, while the computational exploration and its subsequent analysis for the rationalization of the enhanced enzyme properties by our group.

Abstract



The long-standing problem of achieving high activity of a thermophilic enzyme at low temperatures and short reaction times with little tradeoff in thermostability has been solved by directed evolution, an alcohol dehydrogenase found in hot springs serving as the catalyst in enantioselective ketone reductions.

Cite this: *Chem. Commun.*, 2017, 53, 9454Received 13th July 2017,
Accepted 1st August 2017

DOI: 10.1039/c7cc05377k

rsc.li/chemcomm

Inducing high activity of a thermophilic enzyme at ambient temperatures by directed evolution†

Guangyue Li,^{ab} Miguel A. Maria-Solano,^c Adrian Romero-Rivera,^c Silvia Osuna *^c and Manfred T. Reetz *^{ab}

The long-standing problem of achieving high activity of a thermophilic enzyme at low temperatures and short reaction times with little tradeoff in thermostability has been solved by directed evolution, an alcohol dehydrogenase found in hot springs serving as the catalyst in enantioselective ketone reductions.

Robust enzymes derived from thermophilic organisms that thrive under extreme conditions as in hot springs are valuable catalysts in such processes as paper production, baking, laundry detergents and waste-treatment which operate at elevated temperatures.^{1,2} At room temperature these enzymes generally show no activity or low turnover, which is unacceptable for other types of applications, as in the production of chiral pharmaceuticals or other fine chemicals.³ For practical (industrial) applications, maximal stability and activity are needed, yet these appear to be opposing properties. Combining the virtues of pronounced enzyme robustness with high activity at ambient temperatures would lower energy expenditure and enable shorter reaction times under operating conditions, enabling high space-time yields.^{3,4} A limited number of protein engineering studies of such thermostable enzymes using rational design or directed evolution based on mutator strains, epPCR and/or DNA shuffling have appeared.⁵ The improvements proved to be moderate, generally with a tradeoff in thermostability.

The present study likewise focuses on increasing activity of a (hyper)thermally stable enzyme, but this time utilizing a different directed evolution technique. Our goal is opposite to that of conventional thermostabilization of mesophilic enzymes by directed evolution, the usual alternative that is generally accompanied by a tradeoff in activity. For example, Arnold *et al.* applied six cycles of random mutagenesis and DNA shuffling to the

p-nitrobenzyl esterase from *Bacillus subtilis* in order to enhance thermostability, the melting temperature (T_m) increasing from 57 °C to 71 °C, and the k_{cat} -value decreasing from 720 s⁻¹ to 470 s⁻¹ at 30 °C.⁶ This kind of approach has been reviewed.⁷ The mutational effects have been traced to protein rigidification due to newly introduced intramolecular H-bonds and salt bridges as well as disulfide bond formations. In the present approach the opposite effect can be anticipated, namely increased flexibility especially around the active site. Thus, a strategy complementary to the traditional approach would be of theoretical and practical interest. As will be seen, our results are also relevant to the current debate in evolutionary biology regarding changes of enzyme activity and stability starting from a hot environment to a cooled earth over a period of three billion years.⁸

As the model thermophilic enzyme we chose the NAD(P)H- and Zn-dependent alcohol dehydrogenase TbSADH^{9,10} from *Thermoanaerobacter brockii*, first discovered in the hot springs of Yellowstone Park.^{9a} It is identical to *Thermoanaerobacter ethanolicus* (TeSADH) from a different source, a previously used designation. In the purified form this ADH displays high thermostability as demonstrated by a half-life of 1.7 hours at 90 °C and 1.2 days at 80 °C.^{9f} Thermostability as measured by differential scanning microcalorimetry is $T_m = 98.5$ °C.^{9e} Using circular dichroism (CD), we determined T_m to be 90 °C, which is in the range of many hyperthermally stable enzymes.⁵ It has been noted that in the reduction of a variety of structurally different ketones at ambient temperatures using this enzyme long reaction times of several days are needed,^{9b} and keto-esters require 72 °C for reasonable conversion,^{11a} as also reported for other thermophilic ADHs.^{11b} In other cases, overnight reactions had to be performed.⁹ In previous studies, protein engineering of TbSADH was applied for various purposes, including the increase and reversal of stereoselectivity for different substrates,¹² but a significant tradeoff in stability was often noted^{12f,g} or thermostability was not measured.^{12e}

The purpose of the present study was to evolve high activity of a thermophilic enzyme at low temperatures, enabling short reaction times for complete conversion while maintaining robustness.

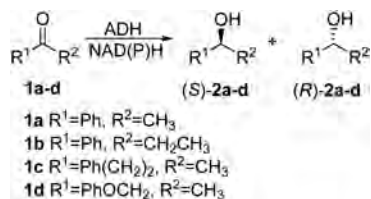
^a Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470, Mülheim an der Ruhr, Germany

^b Fachbereich Chemie der Philipps-Universität Marburg, Hans-Meerwein-Strasse, 35032, Marburg, Germany. E-mail: reetz@mpi-muelheim.mpg.de

^c Institut de Química Computacional i Catàlisi and Department de Química, Universitat de Girona, Carrer Maria Aurèlia Capmany 6, Girona 17003, Catalonia, Spain. E-mail: silvia.osuna@udg.edu

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c7cc05377k





Scheme 1 Asymmetric reduction of ketones **1a–d** catalyzed by TbSADH mutants.

The TbSADH-catalyzed asymmetric reduction of acetophenone (**1a**) was chosen as the model system with enantioselectivity playing a secondary role (Scheme 1). Like many other thermostable ADHs,¹³ TbSADH shows very low activity towards **1a** (and similar substrates)^{9,11a} at 30 °C, and requires extended reaction times and more forcing conditions (*e.g.*, overnight at 50–60 °C).⁹ The wildtype (WT) is slightly (*S*)-selective (17–18% ee).^{12g}

Two crystal structures of TbSADH have been reported,¹⁰ one containing NAD(P)H which is the catalytically active form displaying an open binding pocket and a wide “entrance channel”.^{10a} We employed this structure (1YKF) in order to build a model for docking substrate **1a** into the binding pocket. In this way 13 residues were identified for potential saturation mutagenesis, 10 surrounding the substrate [C37, S39, A85, I86, L107, W110, T154, Y267, L294 and C295], and the rest occupying positions in the entrance channel [I49, C283 and M285] (Fig. 1). These positions were then subjected individually to NNK-based saturation mutagenesis in which all 20 canonical amino acids are used as combinatorial building blocks, requiring in each case the screening of ~96 transformants for 95% library coverage.¹⁴

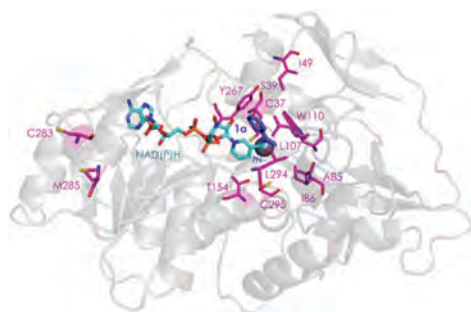


Fig. 1 TbSADH structure model showing docked acetophenone (**1a**) as substrate (in purple) based on the crystal structure of wildtype (1YKF),^{8a} which served as a guide for choosing amino acid positions for saturation mutagenesis (in pink).

In the mini-libraries generated by randomization at positions 85, 86, 110, 283, 285 and 294, several mutants were discovered showing more than a 2-fold activity improvement, namely A85G, I86C, I86E, I86A, W110I, W110L, W110E, C283V, M285L, M285V, L294T and L294V. The libraries created at the other seven positions failed to harbor significantly improved variants (Table S1, ESI†). This information was then used as a basis for performing saturation mutagenesis at a relatively large 6-residue randomization site defined by the above hot spots. The use of NNK codon degeneracy would require for 95% library coverage the screening of >10⁹ transformants.¹⁴ As a practical alternative requiring only 1728 transformants, an appropriate reduced amino acid alphabet^{14,15} was designed individually for each one of the six residues (Table S2, ESI†). The choice of the respective building blocks was guided by the amino acid substitutions that had shown positive effects in the initial NNK-based single libraries. This library harbored several distinctly improved variants (Table S3, ESI†), the best ones being TbSADH-1 (A85G/I86A) and TbSADH-2 (A85G/I86C) as shown by kinetic experiments using purified proteins (Table 1). At 30 °C the two variants show, relative to WT, 58- and 52-fold increases in k_{cat} and 301- and 61-fold improvements in catalytic efficiency ($k_{\text{cat}}/K_{\text{m}}$), respectively. At 45 °C, variants TbSADH-1 and TbSADH-2 also show notably better catalytic performance than WT, namely 51- and 36-fold increases in k_{cat} and improvements in $k_{\text{cat}}/K_{\text{m}}$ by factors of 216 and 52-fold, respectively.

The best mutants TbSADH-1 and TbSADH-2 were tested in upscaled reactions at different temperatures using 50 mM of substrate **1a** in 1 mL of reaction volume (Table S4, ESI†). Excellent results were achieved, *e.g.*, at 30 °C both variants ensured 96% conversion within 1.5 hour with complete enantioselectivity (>99% ee (*R*)). In contrast, at the same temperature WT TbSADH required 20 hours for a mere 4% conversion and 17% ee (*S*). In further experiments, whole cell catalysis at 30 °C using TbSADH-1 and TbSADH-2 was successfully performed using substrate **1a** at concentrations ranging between 200 mM to 2 M (Table S5, ESI†).

We also measured the kinetics of WT TbSADH and variants TbSADH-1 and TbSADH-2 using ketones **1b–d**, revealing similar activity increases (Table S6, ESI†). Synthetically useful results were achieved once more, *e.g.*, in the case of **1b** both variants reaching 96% conversion within one hour with 98% ee (*R*) (Table S7, ESI†). At the same temperature WT TbSADH led to less than 5% conversion after 20 hours with poor (*S*)-selectivity (27% ee).

The thermostability of both variants was measured by determining the melting temperature (T_{m}) using circular dichroism. Relative to WT TbSADH ($T_{\text{m}} = 90$ °C), the robustness of the two

Table 1 Kinetic results using acetophenone (**1a**) as substrate

	Enzyme	Mutations	K_{m} (mM)	k_{cat} (min ⁻¹)	$k_{\text{cat}}/K_{\text{m}}$ (min ⁻¹ M ⁻¹)
30 °C	WT TbSADH		19.01 ± 1.68	1.80 ± 0.07	94
	TbSADH-1	A85G/I86A	3.70 ± 0.17	104.76 ± 1.31	28 313
	TbSADH-2	A85G/I86C	16.20 ± 2.46	93.15 ± 2.35	5750
45 °C	WT TbSADH		20.78 ± 1.25	3.79 ± 0.15	182
	TbSADH-1	A85G/I86A	4.95 ± 0.25	194.54 ± 3.14	39 301
	TbSADH-2	A85G/I86C	14.47 ± 2.05	136.04 ± 9.32	9402



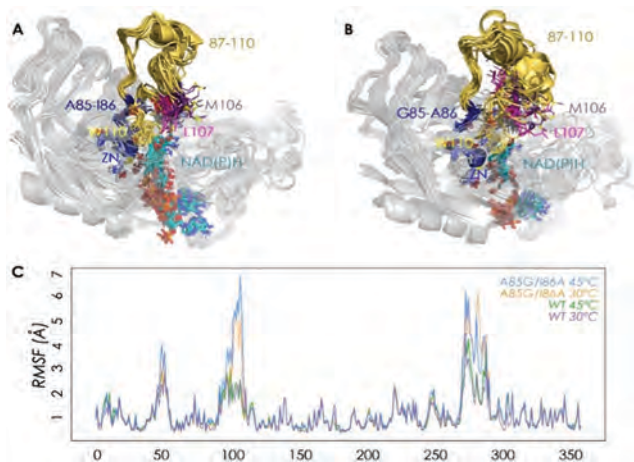


Fig. 2 Overlay of representative snapshots for WT (A) and A85G/I86A variant (B) in the *apo* state at 30 °C. Average values of Root Mean Square Fluctuation (RMSF) of all residues computed from the MD simulations in the *apo* state (C).

best variants TbSADH-1 ($T_m = 84$ °C) and TbSADH-2 ($T_m = 87.5$ °C) is lowered by only 6 °C and 2.5 °C, respectively (Fig. S1 and Table S9, ESI[†]).

We then performed Molecular Dynamics (MD) simulations on the WT enzyme and the A85G/I86A variant, firstly for explaining the origin of dramatically enhanced activity at ambient temperature, and secondly to understand the reversed enantioselectivity (see ESI[†] for computational details). In Fig. 2 and Fig. S7 (ESI[†]), an overlay of representative snapshots from the MD simulations performed in the *apo* state is represented, together with the Root Mean Square Fluctuations (RMSF) of all residues at 30 and 45 °C. The analysis of RMSF allows us to identify the most flexible regions of the enzyme structure, and rationalize the effect of the A85G/I86A mutations on the TbSADH conformational dynamics. The loop composed of residues 87–110 that partially covers the active site of the enzyme (represented in yellow in Fig. 2A and B) is quite rigid in the case of the WT enzyme. The introduction of A85G/I86A induces a higher flexibility on the active site 87–110 loop, which is mainly due to a change in the backbone conformation of residues 106–107 as observed in the Ramachandran plot (see Fig. S8, and ESI[†] Movies). Residues 106–107 are located close to the active site, and make hydrophobic interactions with the substrate (see below). The change in the backbone conformation of 106–107 increases the volume of the active site (from *ca.* 96 Å³ for WT to *ca.* 117 Å³ for A85G/I86A), (Fig. S9 and Table S12, ESI[†]).

The higher flexibility of the A85G/I86A variant, especially in the active site loop, confers the enzyme the ability to change the shape of the active site easily and to adapt to the new non-natural substrate, thus leading to higher activity at low temperatures.

We have also performed MD simulations in the presence of acetophenone (**1a**) to elucidate the origin of reversed enantioselectivity as done in a previous study (see ESI[†] for details).¹⁶ The higher flexibility of the active site loop 87–110 in the A85G/I86A variant plays a key role in dictating the enantioselectivity

of the process. In WT, L107 occupies the small binding pocket of the enzyme forcing **1a** to position the phenyl group in the large binding pocket. This orientation maximizes the CH \cdots π and CH \cdots CH interactions of **1a** and W110, L107, A85, and favors the *pro*-*S* pose. The higher flexibility of the active site loop in the A85G/I86A variant allows **1a** to position the phenyl ring in the small binding pocket, thus favoring the formation of the (*R*)-product. This *pro*-*R* orientation is stabilized by CH \cdots π and CH \cdots CH interactions between **1a** and residues W110, A86, and L294 (see Fig. S10 and S11, ESI[†]).

In conclusion, we have applied an efficient directed evolution strategy to evolve high activity of the thermophilic alcohol dehydrogenase TbSADH at ambient temperatures with little tradeoff in thermostability. Ketones such as acetophenone are rapidly reduced with pronounced enantioselectivity (99% ee) at ambient temperatures within short reaction times. The high thermostability of the mutant(s) suggests that further mutational changes if needed for other purposes can be tolerated.¹⁷ The respective molecular phenomenon, uncovered by MD simulations, points to notably enhanced flexibility of an active site loop. A comparison of the movies of the wildtype and one of the mutants at the respective binding pockets nicely visualizes the underlying effect. This confers the active site pocket higher plasticity and the ability to adapt to new non-natural substrates at lower temperatures. Higher flexibility of the active site loop also has implications in the enantioselectivity of the process, as it changes the preferred orientation of the substrate in the active site pocket.

Our findings have bearing on a recent study in which the putative evolutionary drivers of thermoadaptation in enzyme catalysis were identified.^{8a} On the basis of the hot-start hypothesis of ancestral proteins,⁸ the authors note that “the challenge of evolving efficient enzymatic turnover at lower temperatures has not been addressed”, emphasizing that the traditional concept of stability/activity tradeoff needs to be questioned in Darwinian evolution. While care must be taken when comparing natural with laboratory evolution, our results demonstrate the physical feasibility of evolving such mutational effects. On the practical side, the present mutagenesis approach needs to be generalized by including other (hyper)thermostable enzymes. It will be interesting to see if flexibilization around the binding pocket is a general phenomenon characteristic of such enzyme mutants.

We thank the Max-Planck-Society and the LOEWE cluster SynChemBio for generous support. A. R. R. thanks the Generalitat de Catalunya for PhD fellowship (2015-FI-B-00165), M. A. M. S. is grateful to the Spanish MINECO for PhD fellowship (BES-2015-074964). S. O. thanks the Spanish MINECO for project CTQ2014-59212-P, Ramón y Cajal contract (RYC-2014-16846), the European Community for CIG project (PCIG14-GA-2013-630978), and the funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (ERC-2015-StG-679001). We are grateful for the computer resources, technical expertise, and assistance provided by the Barcelona Supercomputing Center – Centro Nacional de Supercomputación. Open Access funding provided by the Max Planck Society.



Notes and references

- Recent reviews of (hyper)thermophilic enzymes: (a) S. Elleuche, C. Schröder, K. Sahn and G. Antranikian, *Curr. Opin. Biotechnol.*, 2014, **29**, 116–123; (b) K. S. Siddiqui, *Biotechnol. Adv.*, 2015, **33**, 1912–1922; (c) P. A. Fields, Y. Dong, X. Meng and G. N. Somero, *J. Exp. Biol.*, 2015, **218**, 1801–1811; (d) N. Raddadi, A. Cherif, D. Daffonchio, M. Neifar and F. Fava, *Appl. Microbiol. Biotechnol.*, 2015, **99**, 7907–7913.
- Reviews of applications of extremophiles: (a) J. G. Zeikus, C. Vieille and A. Savchenko, *Extremophiles*, 1998, **2**, 179–183; (b) M. E. Bruins, A. E. M. Janssen and R. M. Boom, *Appl. Biochem. Biotechnol.*, 2001, **90**, 155–186; (c) P. Falcicchio, M. Levisson, S. W. Kengen, S. Koutsopoulos and J. van der Oost, *Methods Mol. Biol.*, 2014, **1129**, 487–496; (d) F. Sarmiento, R. Peralta and J. M. Blamey, *Front. Bioeng. Biotechnol.*, 2015, **3**, 148; (e) N. Raddadi, A. Cherif, D. Daffonchio, M. Neifar and F. Fava, *Appl. Microbiol. Biotechnol.*, 2015, **99**, 7907–7913.
- (a) A. S. Bommarius and B. Riebel, *Biocatalysis: Fundamentals and Applications*, Wiley-VCH, Weinheim, 2006; (b) *Industrial Biotransformations*, ed. A. Liese, K. Seelbach and C. Wandrey, Wiley-VCH, Weinheim, 2006; (c) K. Faber, *Biotransformations in Organic Chemistry*, Springer, Heidelberg, 6th edn, 2011; (d) *Enzyme Catalysis in Organic Synthesis*, ed. K. Drauz, H. Gröger and O. May, Wiley-VCH, Weinheim, 3rd edn, 2012; (e) *Organic Synthesis Using Biocatalysis*, ed. A. Goswami and J. Stewart, Elsevier, Amsterdam, 2015.
- Reviews of process engineering and ecological aspects of biocatalysis:³ (a) P. Tufvesson, J. Lima-Ramos, M. Nordblad and J. M. Woodley, *Org. Process Res. Dev.*, 2011, **15**, 266–274; (b) R. Sheldon, *Green Chem.*, 2017, **19**, 18–43; (c) M. Schrewe, M. K. Julsing, B. Bühler and A. Schmid, *Chem. Soc. Rev.*, 2013, **42**, 6346–6377; (d) Y. Ni, D. H. Holtmann and F. Hollmann, *ChemCatChem*, 2014, **6**, 930–943.
- Examples of protein engineering of (hyper)thermophilic enzymes: (a) A. Merz, M.-C. Yee, H. Szadkowski, G. Pappenberger, A. Cramer, W. P. Stemmer, C. Yanofsky and K. Kirschner, *Biochemistry*, 2000, **39**, 880–889; (b) J. H. Lebbink, T. Kaper, P. Bron, J. van der Oost and W. M. de Vos, *Biochemistry*, 2000, **39**, 3656–3665; (c) A. Lönn, M. Gardonyi, W. van Zyl, B. Hahn-Hägerdal and R. C. Otero, *Eur. J. Biochem.*, 2002, **269**, 157–163; (d) D. Sriprapundh, C. Vieille and J. G. Zeikus, *Protein Eng.*, 2003, **16**, 683–690; (e) H.-J. Kang, K. Uegaki, H. Fukada and K. Ishikawa, *Extremophiles*, 2007, **11**, 251–256; (f) C. Q. Zhong, S. Song, N. Fang, X. Liang, H. Zhu, X. F. Tang and B. Tang, *Biotechnol. Bioeng.*, 2009, **104**, 862–870; (g) C. M. Theriot, X. Du, S. R. Tove and A. M. Grunden, *Appl. Microbiol. Biotechnol.*, 2010, **87**, 1715–1726; (h) S. Hayashi, S. Akanuma, W. Onuki, C. Tokunaga and A. Yamagishi, *Biochemistry*, 2011, **50**, 8583–8593; (i) J. Zhang, H. Shi, L. Xu, X. Zhu and X. Li, *PLoS One*, 2015, **10**, e0133824.
- (a) L. Giver, A. Gershenson, P.-O. Freskgard and F. H. Arnold, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 12809–12813; (b) F. H. Arnold, L. Giver, A. Gershenson, H. Zhao and K. Miyazaki, *Ann. N. Y. Acad. Sci.*, 1999, **870**, 400–403.
- (a) V. G. Eijssink, S. Gåseidnes, T. V. Borchert and B. van den Burg, *Biomol. Eng.*, 2005, **22**, 21–30; (b) A. S. Bommarius and M. F. Paye, *Chem. Soc. Rev.*, 2013, **42**, 6534–6565; (c) M. J. Liszka, M. E. Clark, E. Schneider and D. S. Clark, *Annu. Rev. Chem. Biomol. Eng.*, 2012, **3**, 77–102.
- (a) V. Nguyen, C. Wilson, M. Hoemberger, J. B. Stiller, R. V. Agafonov, S. Kutter, J. English, D. L. Theobald and D. Kern, *Science*, 2017, **355**, 289–294; (b) R. Wolfenden, *Cell. Mol. Life Sci.*, 2014, **71**, 2909–2915; (c) R. Nussinov and P. G. Wolynes, *Phys. Chem. Chem. Phys.*, 2014, **16**, 6321–6322.
- Discovery and early studies of TeSADH/TbSADH: (a) R. J. Lamed and J. G. Zeikus, *Biochem. J.*, 1981, **195**, 183–190; (b) E. Keinan, E. K. Hafeli, K. K. Seth and R. Lamed, *J. Am. Chem. Soc.*, 1986, **108**, 162–169; (c) D. S. Burdette, C. Vieille and J. G. Zeikus, *Biochem. J.*, 1996, **316**, 115–122; (d) O. Bogin, M. Peretz, Y. Hacham, Y. Korkhin, F. Frolow, A. J. Kalb (Gilboa) and Y. Burstein, *Protein Sci.*, 1998, **7**, 1156–1163; (e) O. Bogin, M. Peretz, Y. Hacham, Y. Burstein, Y. Korkhin and F. Frolow, *Protein Sci.*, 1998, **7**, 1156–1163; (f) D. S. Burdette, V. Tchernajenko and J. G. Zeikus, *Enzyme Microb. Technol.*, 2000, **27**, 11–18; (g) C. Heiss and R. S. Phillips, *Chem. Soc., Perkin Trans.*, 2000, **16**, 2821–2825.
- X-ray structures of TbSADH: (a) Y. Korkin, A. J. Kalb (Gilboa), M. Peretz, O. Bogin, Y. Burstein and F. Frolow, *J. Mol. Biol.*, 1998, **278**, 967–981; (b) C. Li, J. Heatwole, S. Soelaiman and M. Shoham, *Proteins: Struct., Funct., Genet.*, 1999, **37**, 619–627.
- (a) D. Seebach, M. F. Züger, F. Giovanni, B. Sonleitner and A. Fiechter, *Angew. Chem., Int. Ed. Engl.*, 1984, **23**, 151–152; (b) S. Diederichs, K. Linn, J. Lückgen, T. Klement, J.-H. Grosch, K. Honda, H. Ohtake and J. Büchs, *J. Mol. Catal. B: Enzym.*, 2015, **121**, 37–44.
- Examples of mutagenesis studies of TbSADH: (a) C. Heiss, M. Laivenieks, G. Zeikus and R. S. Phillips, *J. Am. Chem. Soc.*, 2001, **123**, 345–346; (b) K. I. Ziegelmann-Fjeld, M. M. Musa, R. S. Phillips, J. G. Zeikus and C. Vieille, *Protein Eng., Des. Sel.*, 2007, **20**, 47–55; (c) R. Agudo, G.-D. Roiban and M. T. Reetz, *J. Am. Chem. Soc.*, 2013, **135**, 1665–1668; (d) J. M. Patel, M. M. Musa, L. Rodriguez, D. A. Sutton, V. V. Popik and R. S. Phillips, *Org. Biomol. Chem.*, 2014, **12**, 5905–5910; (e) C. M. Nealon, T. P. Welsh, C. S. Kim and R. S. Phillips, *Arch. Biochem. Biophys.*, 2016, **606**, 151–156; (f) Z. Sun, R. Lonsdale, A. Ilie, G. Li, J. Zhou and M. T. Reetz, *ACS Catal.*, 2016, **6**, 1598–1605; (g) Z. Sun, G. Li and M. T. Reetz, *Tet. Lett.*, 2016, **57**, 3648–3651.
- Reviews of ADHs: (a) H. Gröger, W. Hummel, S. Borchert and M. Krauß, in *Enzyme Catalysis in Organic Synthesis*, ed. K. Drauz, H. Gröger and O. May, Wiley-VCH, Weinheim, 3rd edn, 2012, pp. 1035–1110; (b) K. Götz, L. Hilterhaus and A. Liese, in *Enzyme Catalysis in Organic Synthesis*, ed. K. Drauz, H. Gröger and O. May, Wiley-VCH, Weinheim, 3rd edn, 2012, pp. 1205–1223; (c) T. S. Moody and J. D. Rozzell, in *Organic Synthesis Using Biocatalysis*, ed. A. Goswami and J. D. Stewart, Elsevier, Amsterdam, 2016, pp. 149–186.
- Review of directed evolution of stereoselective enzymes with emphasis on iterative saturation mutagenesis: M. T. Reetz, *Angew. Chem., Int. Ed.*, 2011, **5**, 138–174.
- (a) M. T. Reetz and S. Wu, *Chem. Commun.*, 2008, 5499–5501; (b) A. G. Sandström, Y. Wikmar, K. Engström, J. Nyhlen and J.-E. Bäckvall, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 78–83; (c) Z. Sun, Y. Wikmark, J.-E. Bäckvall and M. T. Reetz, *Chem. – Eur. J.*, 2016, **22**, 5046–5054.
- (a) M. A. Maria-Solano, A. Romero-Rivera and S. Osuna, *Org. Biomol. Chem.*, 2017, **15**, 4122–4129; (b) A. Romero-Rivera, M. Garcia-Borrás and S. Osuna, *Chem. Commun.*, 2017, **53**, 284–297.
- J. D. Bloom, S. T. Labthavikul, C. R. Otey and F. H. Arnold, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 5869–5874.



Electronic Supplementary Material (ESI) for ChemComm.
This journal is © The Royal Society of Chemistry 2017

Inducing High Activity of a Thermophilic Enzyme at Ambient Temperatures by Directed Evolution

Guangyue Li, Miguel A. Maria-Solano, Adrian Romero-Rivera, Silvia Osuna and
Manfred T. Reetz**

Methods

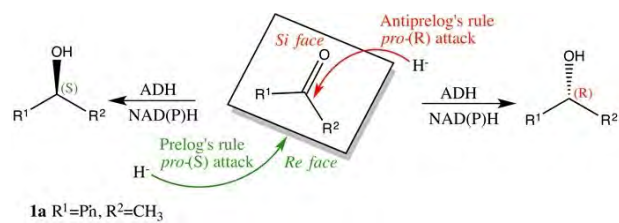
Docking acetophenone (1a) into WT TbSADH

The X-ray structure¹ of TbSADH was used as the basis for docking calculations. The models of WT TbSADH was prepared as in a previous study.² Substrate acetophenone (**1a**) was prepared for docking using ChemDraw. Docking to the WT TbSADH was performed using autodock vina. Ten docking poses were requested and a constraint was applied such that only the docking poses in which the substrate coordinates to the active site zinc ion were saved.

Molecular dynamics (MD) simulations.

MD simulations in explicit water were performed using AMBER 16 package⁴ and starting from the PDB structure: 1YKF.¹ The A85G/I86A variant was generated using the mutagenesis tool included in PyMOL (<http://www.pymol.org>). Parameters for substrate **1a** for the MD simulations were generated within the *antechamber* module of AMBER 16 using the general AMBER force field (GAFF),⁵ with partial charges set to fit the electrostatic potential generated at the B3LYP/6-31G(d) level by the restrained electrostatic potential (RESP) model.⁶ The charges were calculated according to the Merz-Singh-Kollman scheme^{7, 8} using Gaussian 09.⁹ Amino acid protonation states were predicted using the H++ server (<http://biophysics.cs.vt.edu/H++>).¹⁰ We have used the bonded model for Zn metal center, the residues of the first coordination sphere and either the substrate or a water molecule (*apo* state) bound.¹¹ In particular we used the Seminario approach¹² to obtain the metal parameters needed for the simulation as implemented in Prof. Ryde program.¹³ The optimization, frequencies and charge calculations to obtain the parameters was done at the B3LYP/6-31G(d) level using Gaussian 09.⁷ The parameters for NAD(P)H were extracted from previous studies by Prof. Ryde.^{14,15} The wild-type enzyme (PDB: 1YKF) and variant were solvated in a pre-equilibrated truncated cuboid box with a 10-Å buffer of TIP3P¹⁶ water molecules using the AMBER16 *leap* module, resulting in the addition of *ca.* 9,000 solvent molecules. The system was neutralized by addition of explicit counterions (Na⁺ and Cl⁻). All calculations were done using a modification of the *ff99SB* force field (*ff94SB*).¹⁷ A two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on solute by a harmonic potential with a force constant of 500 kcal mol⁻¹ Å⁻², and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems are gently heated using six 50-ps steps, incrementing the temperature 50 K each step (0–300 K, 30°C and 0–315 K, 45°C) under constant volume and periodic boundary conditions. Water molecules were treated with the SHAKE algorithm such that the angle between the hydrogen atoms is kept fixed. Long-range electrostatic effects were modeled using the particle-mesh-Ewald method.¹⁸ An 8-Å cutoff was applied to Lennard-Jones and electrostatic interactions. Harmonic restraints of 10 kcal/mol were applied to the solute, and the Langevin equilibration scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was then equilibrated without restraints for 2 ns with a 2-fs timestep at a constant pressure of 1 atm and temperature of 300 K. After the systems were equilibrated in the NPT ensemble, 5 independent two hundred nanosecond MD simulations were performed under the NVT ensemble and periodic-boundary conditions at 30 and 45°C in the substrate-bound and *apo* states. Therefore, an accumulated simulation time of 1 microsecond has been obtained for each variant (WT and A85G/I86A) at each temperature (30 and 45°C) in both *apo* and substrate-bound states.

Schemes, Tables and Figures



Scheme S1. Prelog and anti-Prelog selectivity for model ketone reductions catalyzed by TbSADH.

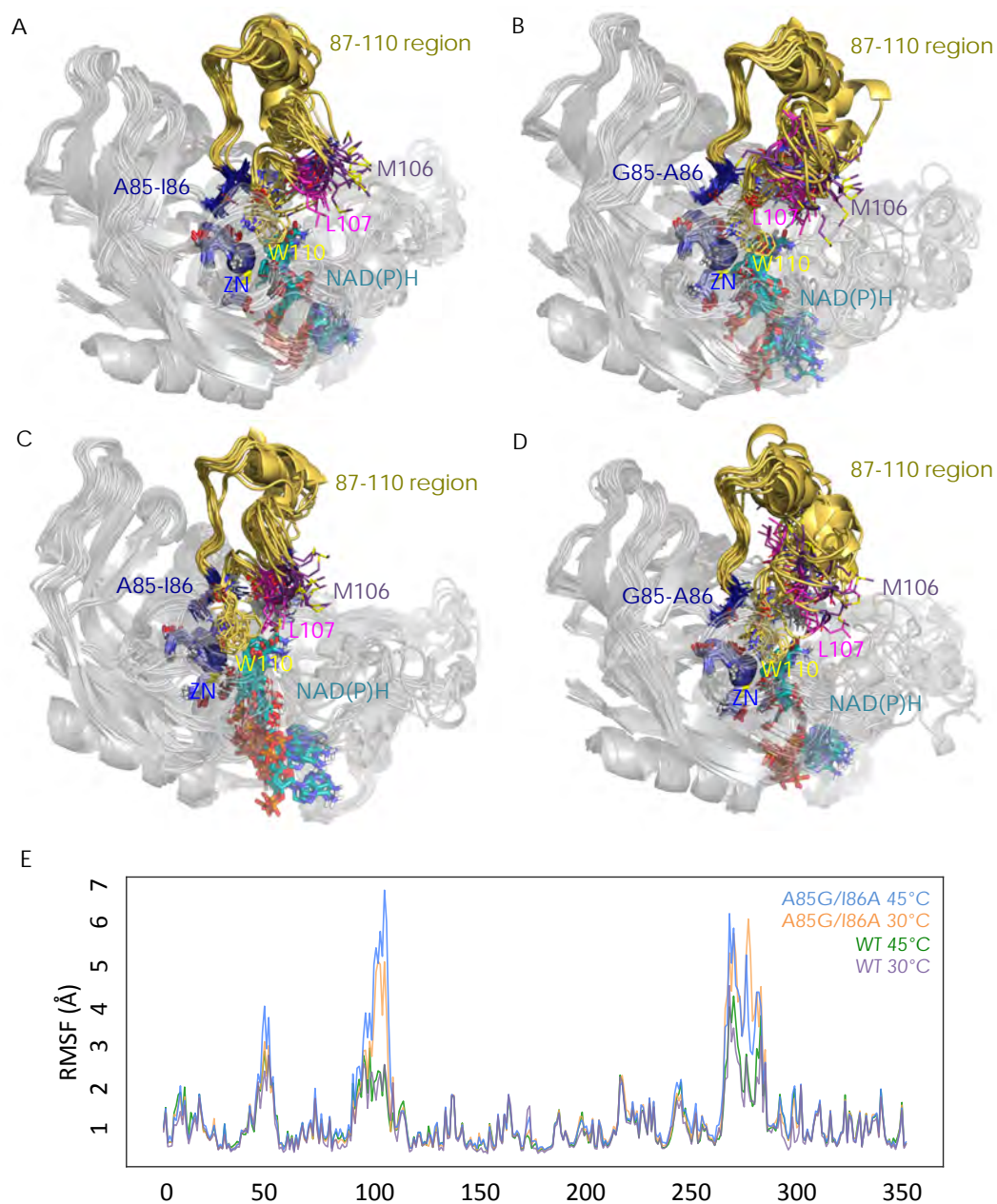


Figure S7. Overlay of some representative MD snapshots for WT (A 30°C, C 45°C) and A85G/I86A variant (B 30°C, D 45°C) in the *apo* state. Average values of Root Mean Square Fluctuation (RMSF) of all residues computed from the MD simulations (where the cofactor is not displaced from the active site) in *apo* state (E)

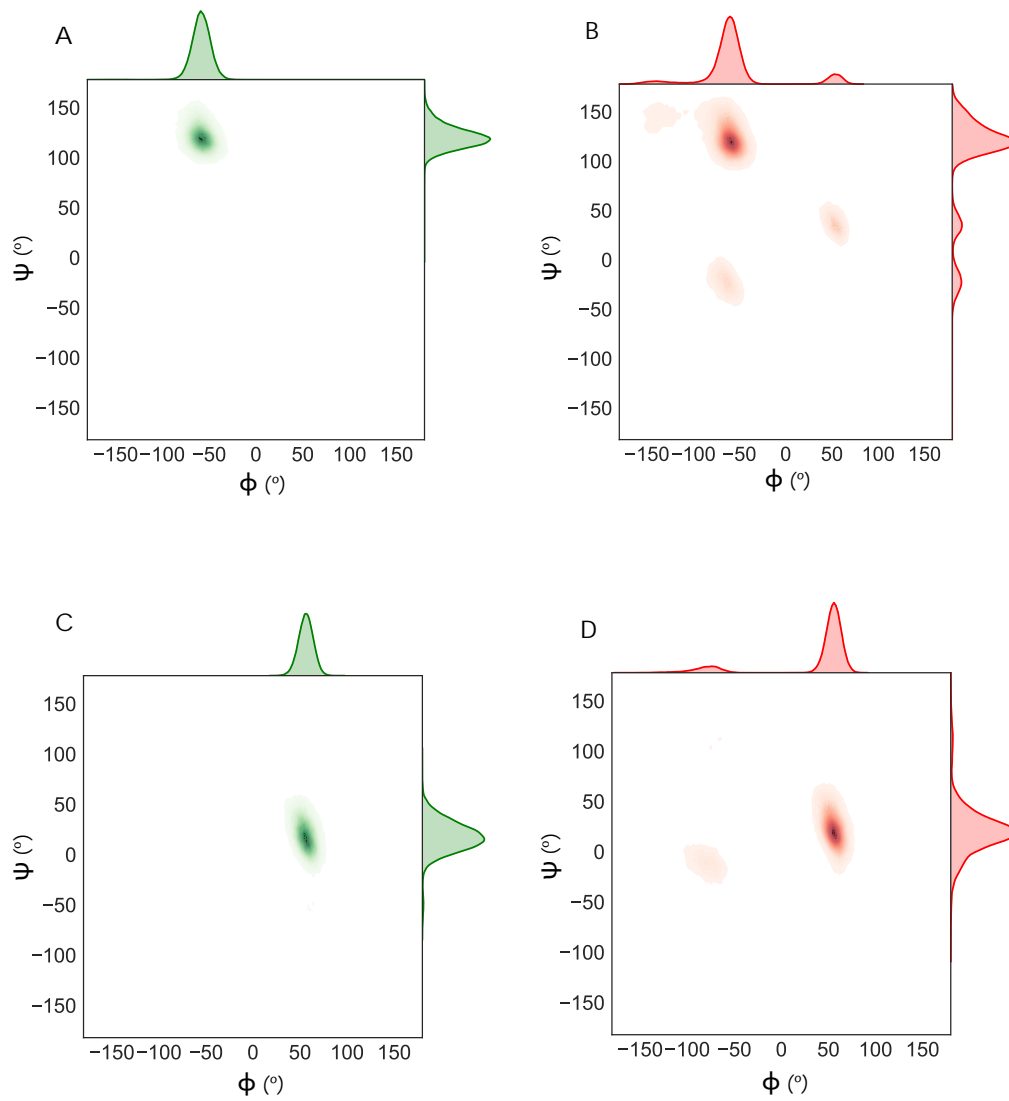


Figure S8. Representation of the Ramachandran plots residues M106 (**A**, **B**), and L107 (**C**, **D**) for WT en) and A85G/I86A variant (red) for all MD simulations in the *apo* state at 30°C.

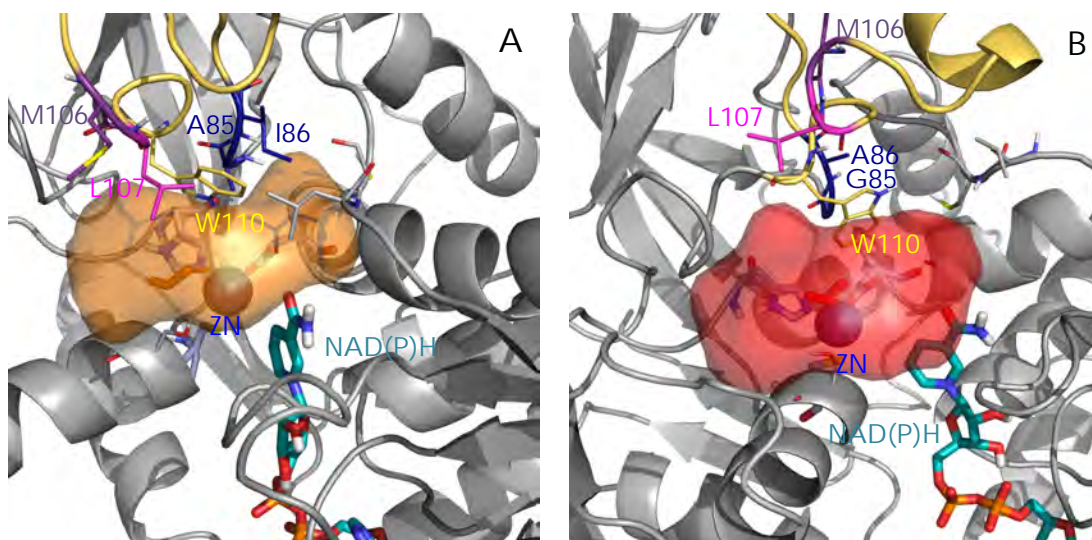


Figure S9. Active site volume representation of the most populated cluster from the MD simulations in the state for WT (A), and A85G/I86A variant (B). These calculations have been performed with POVME²⁰

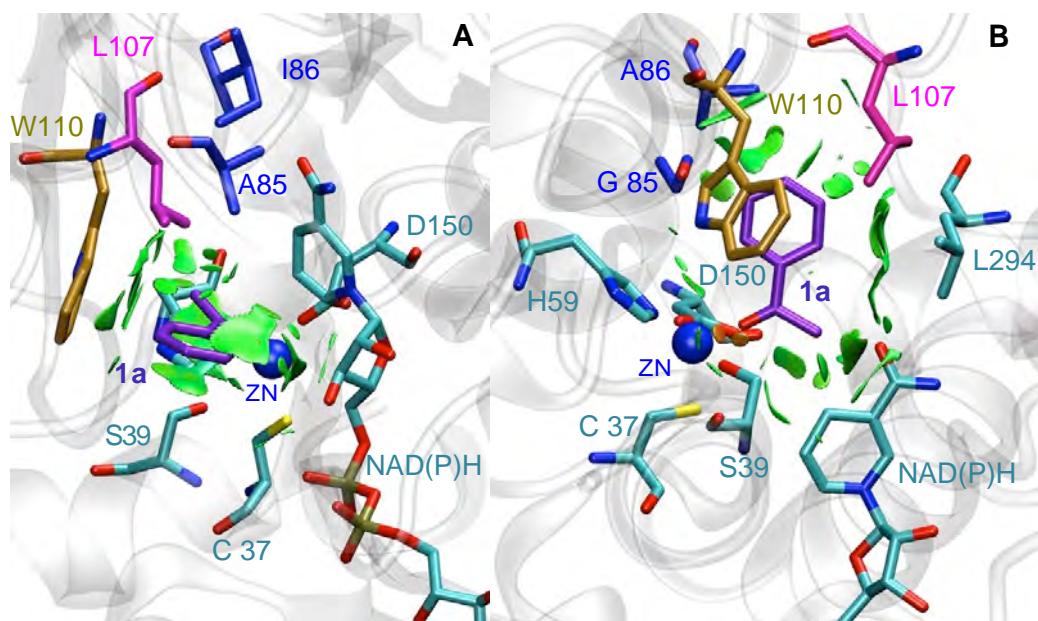


Figure S10. Representations of the most important non-covalent interactions (in green) between the substrate **1a** and the active site for WT (A) and A85G/I86A variant (B), computed with the computational tool NCIplot.²¹

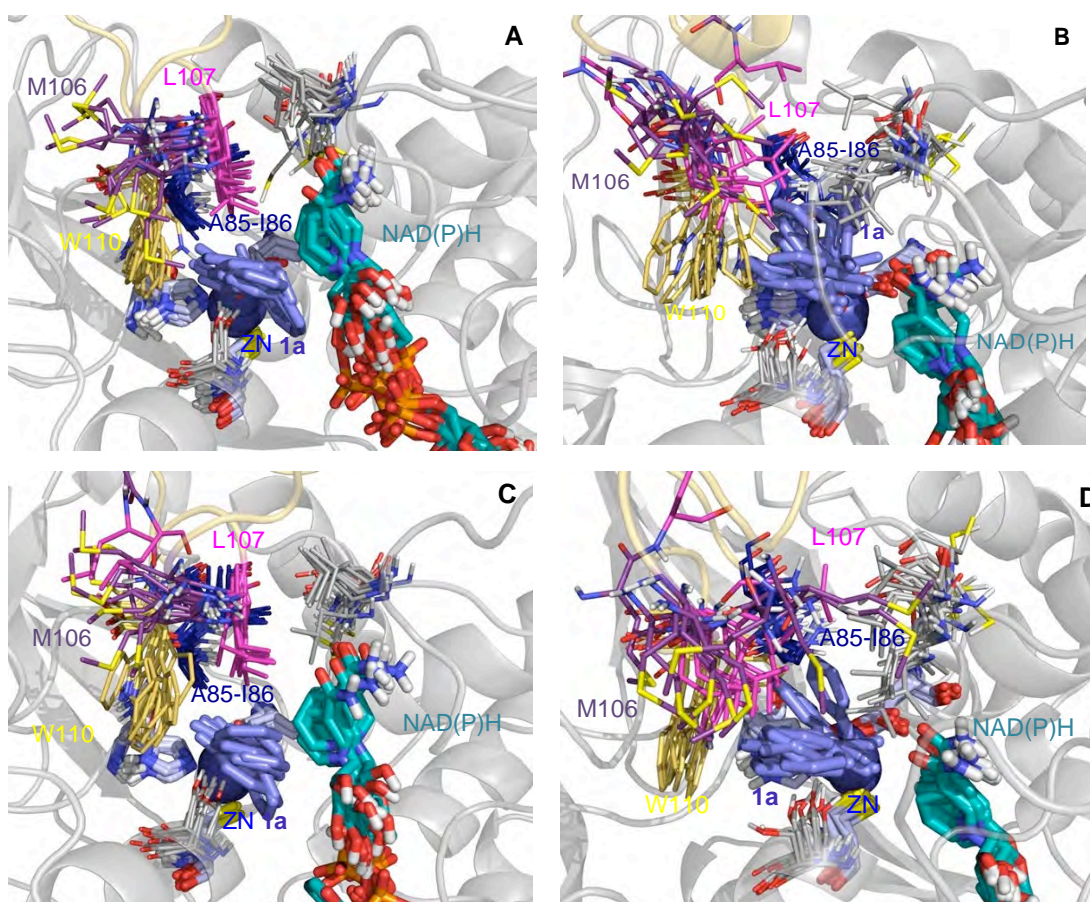


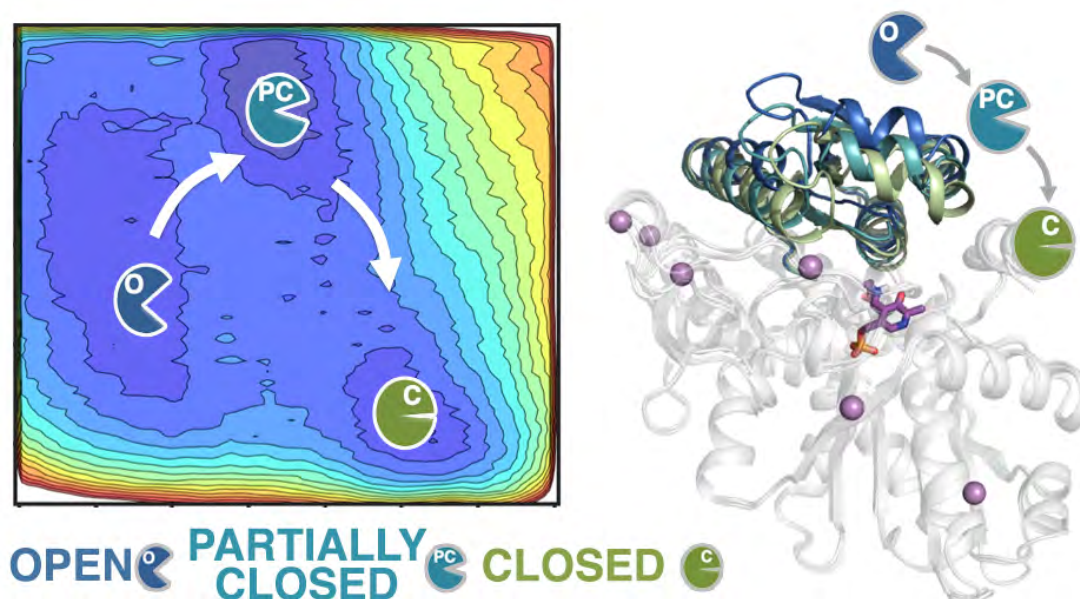
Figure S11. Overlay of some representative MD snapshots for WT with **1a** in *pro*-(S) conformation (**A** 30°C, **C** 45°C) and A85G/I86A variant with **1a** in *pro*-(R) conformation (**B** 30°C, **D** 45°C).

References

1. Korkhin, Y.; Kalb, A. J.; Peretz, M.; Bogin, O.; Burstein, Y.; Frolow, F. *J. Mol. Biol.* **1998**, *278*, 967.
2. Sun, Z.; Lonsdale, R.; Ilie, A.; Li, G.; Zhou, J.; Reetz, M. T. *ACS Catal.* **2016**, *6*, 1598.
3. Bougioukou, D. J.; Kille, S.; Taglieber, A.; Reetz, M. T. *Adv. Synth. Catal.* **2009**, *351*, 3287-3305.
4. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; et al. AMBER 16, University of California, San Francisco, 2016.
5. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157-1174.
6. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269-10280.
7. Bessler, B.; Merz Jr, K.; Kollman, P. *J. Comput. Chem.* **1990**, *11*, 431-439.
8. Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129-145.
9. M. J. Frisch GWT, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian 09, Revision A. 02. Gaussian. Inc: Wallingford, CT 2009.
10. Anandkrishnan, R.; Aguilar, B.; Onufriev, A. V. *Nucleic Acids Res.* **2012**, *40*, W537-W541.
11. G. V. Dhoke, M. D. Davari, U. Schwaneberg, M. Bocola. *ACS Catal.* **2015**, *5*, 3207-3215.
12. Seminario, J.M. *Int. J. Quantum. Chem.* **1996**, *60*, 1271-1277.
13. Hu, L.; Ryde, U. *J. Chem. Theory. Comp.* **2011**, *7*, 2452-2463.
14. Ryde, U. *Protein Sci.* **1995**, *4*, 1124-1132.
15. Ryde, U. *J. Chem. Phys.* **1983**, *79*, 926-935.
16. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. chem. phys.* **1983**, *79*, 926-935.
17. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Structure, Function, and Bioinformatics*, **2006**, *65*, 712-725.
18. Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089-10092.
19. Sun, Z.; Li, G.; Ilie, A.; Reetz, M. T. *Tetrahedron. Lett.* **2016**, *57*, 3648.
20. J. D. Durrant, L. Votapka, J. Sørensen, R. E. Amaro, J. *Chem. Theory. Comp.* **2014**, *10*, 5047-5056.
21. a) J. Contreras-García, E. R. Johnson, S. Keinan, R. Chaudret, J.-P. Piquemal, D. N. Beratan, W. Yang, J. *Chem. Theory. Comp.* **2011**, *7*, 625-632. b) E. R. Johnson, S. Keinan, P. Mori-Sanchez, J. Contreras-Garcia, A. J. Cohen, W. Yang, *J. Am. Chem. Soc.* **2010**, *132*, 6498-6506.

Chapter 5. Allosteric properties and stand-alone function of tryptophan synthase (TrpS) enzymes

5.1 Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution



Maria-Solano, M.A.; Iglesias-Fernández, J.*; Osuna, S.* Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution, *J. Am. Chem. Soc.* **2019**, *141*, 13049-13056. [Chemistry, Multidisciplinary, 14.70, Q1]. DOI: [10.1021/jacs.9b03646](https://doi.org/10.1021/jacs.9b03646)

The permission to reuse this published article in this thesis has been granted from ACS Publications in print and electronic formats.

Abstract

Multimeric enzyme complexes are ubiquitous in nature and catalyze a broad range of useful biological transformations. They are often characterized by a tight allosteric coupling between subunits, making them highly inefficient when isolated. A good example is Tryptophan synthase (TrpS), an allosteric heterodimeric enzyme in the form of an $\alpha\beta\alpha$ complex that catalyzes the biosynthesis of L-tryptophan. In this study, we decipher the allosteric regulation existing in TrpS from *Pyrococcus furiosus* (PfTrpS), and how the allosteric conformational ensemble is recovered in laboratory-evolved stand-alone β -subunit variants. We find that recovering the conformational ensemble of a subdomain of TrpS affecting the relative stabilities of open, partially closed, and closed conformations is a prerequisite for enhancing the catalytic efficiency of the β -subunit in the absence of its binding partner. The distal mutations resuscitate the allosterically driven conformational regulation and alter the populations and rates of exchange between these multiple conformational states, which are essential for the multistep reaction pathway of the enzyme. Interestingly, these distal mutations can be a priori predicted by careful analysis of the conformational ensemble of the TrpS enzyme through computational methods. Our study provides the enzyme design field with a rational approach for evolving allosteric enzymes toward improved stand-alone function for biosynthetic applications.

Deciphering the Allosterically Driven Conformational Ensemble in Tryptophan Synthase Evolution

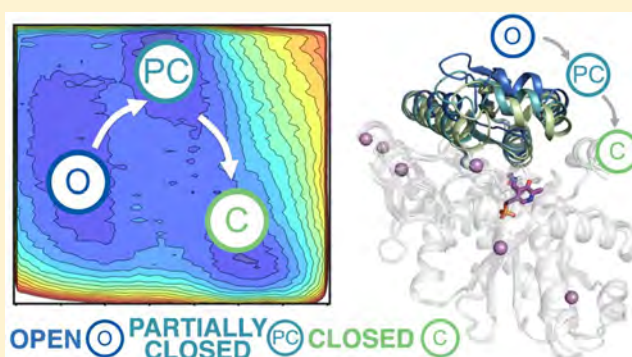
Miguel A. Maria-Solano,[†] Javier Iglesias-Fernández,^{*,†} and Sílvia Osuna^{*,†,‡}

[†]CompBioLab group, Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, Girona 17003, Spain

[‡]ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain

S Supporting Information

ABSTRACT: Multimeric enzyme complexes are ubiquitous in nature and catalyze a broad range of useful biological transformations. They are often characterized by a tight allosteric coupling between subunits, making them highly inefficient when isolated. A good example is Tryptophan synthase (TrpS), an allosteric heterodimeric enzyme in the form of an $\alpha\beta\beta\alpha$ complex that catalyzes the biosynthesis of L-tryptophan. In this study, we decipher the allosteric regulation existing in TrpS from *Pyrococcus furiosus* (*Pf*TrpS), and how the allosteric conformational ensemble is recovered in laboratory-evolved stand-alone β -subunit variants. We find that recovering the conformational ensemble of a subdomain of TrpS affecting the relative stabilities of open, partially closed, and closed conformations is a prerequisite for enhancing the catalytic efficiency of the β -subunit in the absence of its binding partner. The distal mutations resuscitate the allosterically driven conformational regulation and alter the populations and rates of exchange between these multiple conformational states, which are essential for the multistep reaction pathway of the enzyme. Interestingly, these distal mutations can be a priori predicted by careful analysis of the conformational ensemble of the TrpS enzyme through computational methods. Our study provides the enzyme design field with a rational approach for evolving allosteric enzymes toward improved stand-alone function for biosynthetic applications.



INTRODUCTION

Allostery is a central biological process in which two distinct sites within a biomolecule are functionally connected. Allosteric effects play a key role in protein regulation and cell signaling, and their functional significance has fostered many studies for unveiling the underlying forces that drive allostery.^{1–3} In enzymatic mechanisms, allosteric interactions often promote enzyme–substrate binding and product release, and directly affect catalytic turnover.^{4–6} Some studies suggest that allostery is an intrinsic characteristic of enzymes,⁷ given the fact that distal active site mutations often confer improved catalytic properties.^{8–11} The essential role played by remote mutations in tuning enzyme activity also indicates that allostery could be exploited for the engineering of new enzyme variants.¹²

Allosteric regulation present in multimeric enzyme complexes makes the isolated subunits, that is, in the absence of their protein partner, highly inefficient. This is indeed the case for Tryptophan synthase (TrpS; EC 4.2.1.20). TrpS is a heterodimeric enzyme complex composed of α -subunits (TrpA) and β -subunits (TrpB) in an $\alpha\beta\beta\alpha$ arrangement that presents an intricate allosteric communication network between TrpA and TrpB.^{13–15} TrpA catalyzes the retro-aldol cleavage of indole-3-glycerol phosphate (IGP) producing

glyceraldehyde-3-phosphate (G3P) and indole; the latter is able to diffuse through an internal TrpA–TrpB tunnel to reach the TrpB subunit (see Figure 1).

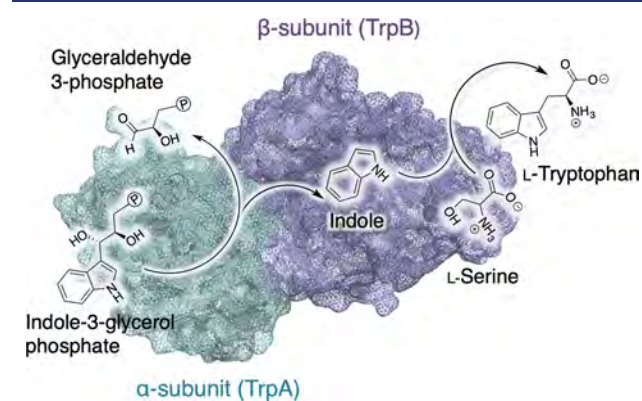


Figure 1. Overview of the Tryptophan synthase (TrpS) mechanism. The enzyme is a heterodimeric complex formed by two subunits: TrpA (shown in teal) and TrpB (in purple).

Received: April 4, 2019

Published: July 29, 2019

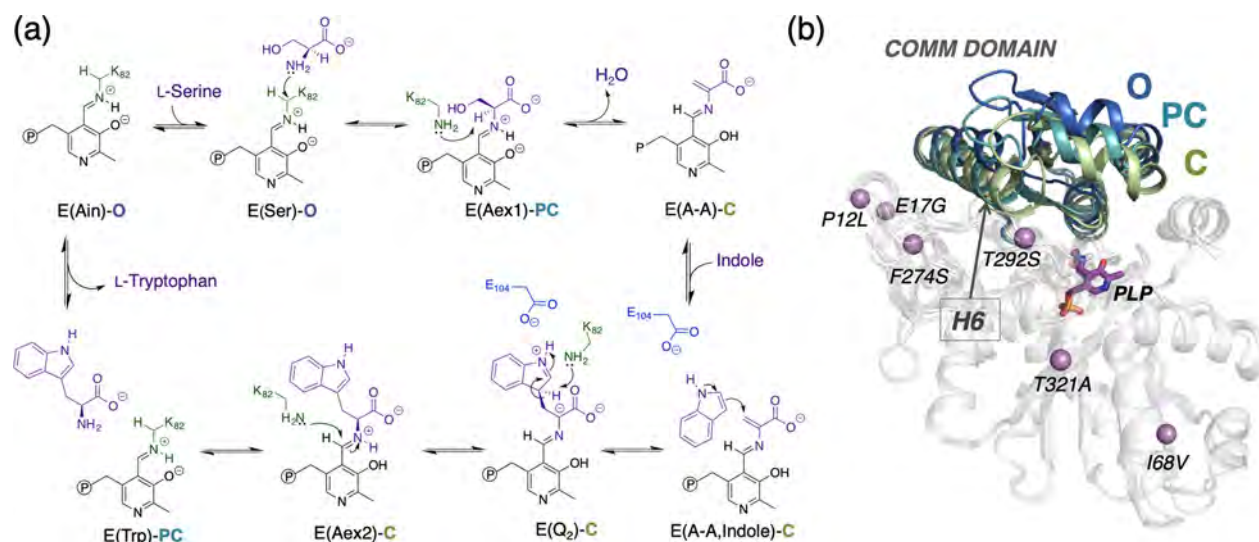


Figure 2. (a) Tryptophan synthase reaction mechanism of TrpB subunit showing the conformational states of the COMM domain according to the available X-ray data at each reaction intermediate. The degree of closure of the COMM domain is represented by colored labels in blue (open, O), teal (partially closed, PC), and green (closed, C). (b) Overlay of representative X-ray structures showing the COMM domain (97–184 residues) in O, PC, and C states. The pyridoxal phosphate (PLP) cofactor is shown in purple, the *Pf*TrpB^{OB2} stand-alone DE mutation positions are marked with purple spheres, and the COMM domain α -helix H6 (residues 174–164) is highlighted.

The TrpB resting state is characterized by a pyridoxal phosphate (PLP)-cofactor covalently linked to the K82 active site residue, forming a Schiff base intermediate (E(Ain)). After transamination with L-serine (E(Ser)), an external aldimine intermediate (E(Aex1)) is formed. This intermediate undergoes deprotonation at C α , assisted by K82, which is followed by a rapid elimination of the Aex1 hydroxyl group to form an electrophilic amino acrylate intermediate (E(A-A)). In the dimeric complex, indole formed in TrpA reaches the TrpB active site and reacts with E(A-A) to form a quinonoid intermediate (E(Q₂)), which after proton extraction (to recover indole's aromatic character) generates E(Q₃) (not shown in Figure 2a). At this point, protonation at C α of Q₃ by K82 forms the E(Aex2) intermediate, which undergoes a second transamination reaction to finally release the L-tryptophan (E(Trp)) product and restore the enzyme resting state (see Figure 2).¹⁵

Previous studies along the catalytic mechanism identified different open and closed conformations of the enzyme in both subunits, which were based on static X-ray structures. These open-to-closed transitions can be defined by the TrpA loop (residues 163–176) that gets ordered and the slow motion of the rigid COMM domain in TrpB (residues 97–184, see Figure 2 and Table S1). Both the TrpA loop and the TrpB COMM domain are part of the active site cavity of each subunit and modulate solvent exposure to prevent substrate loss to the media. Besides, the COMM domain contains the α -helix H6 (residues 174–164) that is directly involved in noncovalent interactions with the indole moiety of the TrpB reactant intermediates. Moreover, E104 located in the COMM domain has been reported to play a role in the stabilization of charge redistribution that takes place during the nucleophilic attack of indole in the A-A intermediate (see Figure 2).¹⁵

Tryptophan synthase has found applications in many fields of synthetic chemistry, in particular, for the production of noncanonical amino acids (NCAAs).^{16–18} The use of TrpS for industrial purposes is hampered by its multimeric structure and the low activity of TrpB as stand-alone enzyme. Detailed

insights were obtained in studies by Prof. Arnold and co-workers, who applied Directed Evolution (DE) to a thermophilic TrpB from *Pyrococcus furiosus* (*Pf*TrpB). They optimized the enzyme for stand-alone function,¹⁹ and later on for the production of a variety of Trp derivatives.^{19–25} The most efficient stand-alone catalyst was achieved by introducing up to six activating mutations, which were located far away from active site positions (see Figure 2b). Note that P12L, E17G, and F274S are located close to the TrpA–TrpB protein interface. Analysis of spectroscopic data suggested that *Pf*TrpB stand-alone variants and the *Pf*TrpS complex were better in stabilizing closed conformations of the COMM domain upon substrate binding than isolated *Pf*TrpB.²⁴ However, despite showing drastic differences in catalytic efficiency, X-ray data failed to find a connection between COMM domain closure and stabilization of the enzyme. In particular, the COMM domain structure is almost identical among different organisms (e.g., *Salmonella typhimurium* and *Pyrococcus furiosus*), isolated *Pf*TrpB enzyme, and *Pf*TrpB stand-alone variants, although all of them diverge in functionality. These observations suggest that the origin behind their different catalytic efficiencies could be attributed to alterations in the enzyme conformational ensemble induced by distal active site mutations. Such effects have not been explored yet, although they are crucial to understand how the stand-alone functionality was achieved.

In this work, we elucidate how the different reaction intermediates and distal mutations introduced in laboratory-evolution alter the allosterically driven conformational ensemble of *Pf*TrpS. Surprisingly, the introduced distal mutations increase the conformational heterogeneity of the COMM domain; hence, the *Pf*TrpB enzyme has the ability to access the different COMM domain conformations, which are essential for efficient catalysis in the absence of its binding partner. Through careful inspection of the conformational ensemble of *Pf*TrpS with our recently developed SPM tool,¹² we were able to identify the most important positions to recover the allosterically driven conformational ensemble, which coincide with the mutations introduced in laboratory-

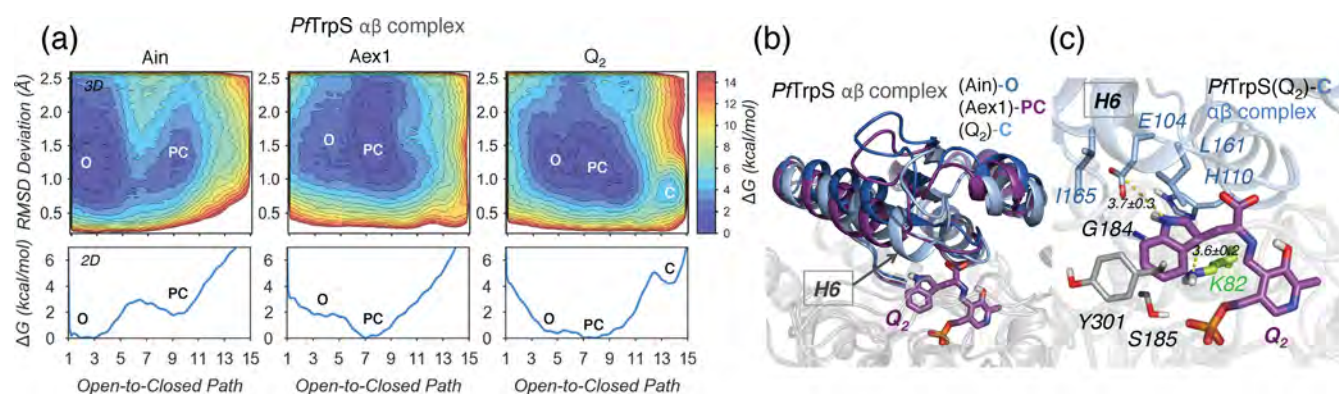


Figure 3. (a) Free energy landscape (FEL) associated with the COMM domain open-to-closed (O-to-C) conformational exchange of the *PfTrpS* complex enzyme at Ain, Aex1, and Q_2 reaction intermediates. (b) Overlay of the *PfTrpS* metastable conformations of the open (O) state at Ain intermediate, partially closed (PC) at Aex1, and closed (C) at Q_2 , respectively, showing the entire O-to-C sampled transition. (c) Detailed active site view of the *PfTrpS* metastable conformation of the C state at Q_2 intermediate (shown in purple). Active site residues are shown in gray, except for those included in the COMM domain (shown in blue), and the catalytic K82 proton transfer residue (green). The catalytic distances (in Å) between charge–charge stabilization E104- Q_2 and proton transfer K82- Q_2 are also represented.

evolution. Our study shows clearly that stand-alone versions of allosterically regulated enzymes can be rationally designed by targeting the recovery of the allosterically driven conformational ensemble.

RESULTS AND DISCUSSION

Available structural data show that the TrpB COMM domain is able to explore open (O), partially closed (PC), and closed (C) conformations along the multistep TrpB catalytic pathway, due to the allosteric regulation exerted by TrpA (see Figure 2). Considering only the TrpB subunit, X-ray studies revealed that its resting state (i.e., E(Ain)) is characterized by the O COMM domain conformations (1V8Z),²⁶ which are shifted toward PC states at the external aldimine intermediate E(Aex1) (5DW0).¹⁹ All subsequent reaction intermediates (i.e., from the electrophilic amino acrylate E(A–A) to E(Aex2)) were crystallized in C states (4HN4²⁷ and 3CEP²⁸). A recent X-ray structure (5DW3)¹⁹ indicated that the PC conformation is recovered once Trp is formed at E(Trp), preparing the enzyme for product release and the next turnover (see Figures 1 and 2 and Table S1 for more structural data).

Allosteric transitions, such as the TrpA-triggered O-to-C exchange of the COMM domain in TrpB, are relatively slow domain motions that take place on time scales larger than our currently accessible simulation times.²⁹ Indeed, initial 500 ns standard MD simulations of the *PfTrpS* in the $\alpha\beta$ complex, the isolated *PfTrpB* wild-type, and the stand-alone *PfTrpB*^{OB2} enzyme variant in multiple reaction intermediates (Ain, Aex1, and A–A) failed to sample the entire allosteric transition, with no clear RMSD differences observed between the studied systems (Figure S2). To overcome this limitation, we employed enhanced sampling techniques. In particular, we applied the metadynamics approach^{30,31} to reconstruct the free energy landscape (FEL) associated with the COMM domain O-to-C transition of the *PfTrpS* $\alpha\beta$ complex, isolated *PfTrpB* wild-type, and evolved stand-alone *PfTrpB*^{OB2} variant (see details in the Supporting Information). Several intermediates along the catalytic cycle were modeled. In particular, we selected E(Ain), E(Aex1), E(A–A), and E(Q_2) to evaluate the O–PC–C conformational exchange of the COMM domain found in X-ray data, but also to reproduce the multistep mechanism under study (see Figure 2).

Population Shift toward Closed Conformations along the Allosteric *PfTrpS* Catalytic Pathway.

To elucidate the allosterically driven conformational ensemble of *PfTrpS* $\alpha\beta$ complex, we reconstructed the FEL associated with the conformational dynamics of the COMM domain for each reaction intermediate (see Figure 3a). As expected from X-ray data, in the resting state of the enzyme, *PfTrpS*-Ain, the O conformational state is highly favored, in agreement with its functional role in Ser binding. However, less stable PC states (ca. 2 kcal/mol higher in energy) are also visited with an associated O-to-PC transition energy barrier of only ca. 3 kcal/mol. As the *PfTrpS* enzymatic reaction progresses, a population shift occurs toward the stabilization of PC states (see Figure 3a). After the reaction with serine in the external aldimine Aex1 intermediate, the open O state is destabilized by ca. 2 kcal/mol with respect to the PC state, which becomes the most stable conformation. In contrast to Ain and Aex1, the quinonoid Q_2 intermediate generated after indole coupling samples all possible conformations of the COMM domain: O and PC states are almost equally stabilized, while the C state is ca. 5 kcal/mol higher in energy. The associated PC-to-C barrier is ca. 6 kcal/mol. This suggests that the adoption of the fully closed COMM domain conformation is the limiting factor, in agreement with the spectroscopic data for *PfTrpS*.^{19,24} Such closed active states form an optimized network of hydrophobic interactions between the enzyme and the indole moiety (see Figure 3c). Several side-chain residues, including Y301, S185, and COMM domain G184, H110, L161, and I165 define this network.

Comparison of *PfTrpS*(Q_2)-O and -C metastable structures shows that the helix H6 closure is needed for forming CH \cdots CH and CH $\cdots\pi$ interactions between L161 and I165 with the indole moiety and also a hydrogen bond with the E104 residue (Figure 3b,c). The C state of *PfTrpS*(Q_2) shows a highly preorganized active site with E104 and the proton acceptor K82 properly positioned for catalysis together with the indole moiety establishing many noncovalent interactions with the active site pocket (see Figures 3c and S8a,b). The high stability of O states at the Q_2 intermediate suggests that the COMM domain of *PfTrpS*(Q_2) is already prepared for product release and recovery of the native state of the enzyme for the next cycle. Altogether, these findings highlight the crucial role of the allosterically driven conformational ensemble of the COMM

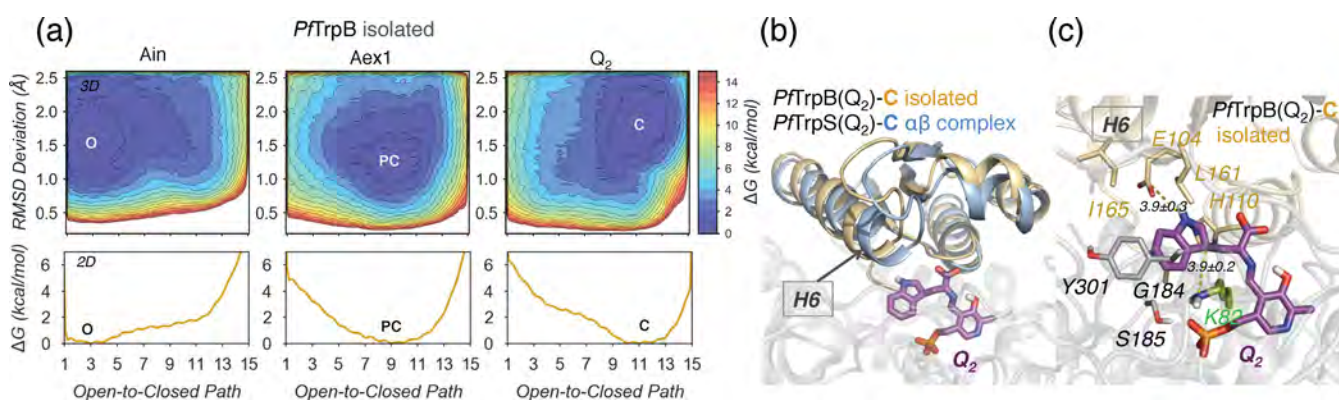


Figure 4. (a) Free energy landscape (FEL) associated with the COMM domain open-to-closed (O-to-C) conformational exchange of the *Pf*TrpB isolated enzyme at Ain, Aex1, and Q₂ reaction intermediates. (b) Overlay of the metastable conformations of the closed (C) states at Q₂ intermediate for *Pf*TrpB (in orange) and *Pf*TrpS (blue). (c) Detailed active site view of the *Pf*TrpB metastable conformation of the C state at Q₂ intermediate (shown in purple). Active site residues are shown in gray, except for those included in the COMM domain (shown in orange), and the catalytic K82 proton transfer residue (green). The catalytic distances (in Å) between charge–charge stabilization E104–Q₂ and proton transfer K82–Q₂ are also represented.

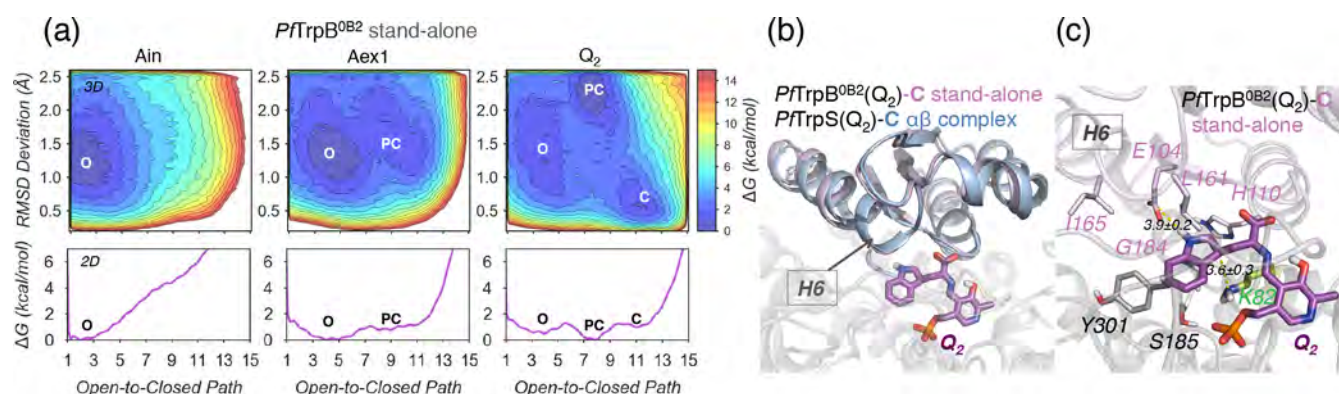


Figure 5. (a) Free energy landscape (FEL) associated with the COMM domain open-to-closed (O-to-C) conformational exchange of the *Pf*TrpB^{OB2} enzyme at Ain, Aex1, and Q₂ reaction intermediates. (b) Overlay of the metastable conformations of the closed (C) states at Q₂ intermediate for *Pf*TrpB^{OB2} (in pink) and *Pf*TrpS (blue). (c) Detailed active site view of the *Pf*TrpB^{OB2} metastable conformation of the C state at Q₂ intermediate (shown in purple). Active site residues are shown in gray, except for those included in the COMM domain (shown in violet), and the catalytic K82 proton transfer residue (green). The catalytic distances (in Å) between charge–charge stabilization E104–Q₂ and proton transfer K82–Q₂ are also represented.

domain of *Pf*TrpS for efficiently optimizing the multiple steps along its catalytic cycle.

Isolated *Pf*TrpB Displays Restricted COMM Domain Heterogeneity and Unproductive Closure. Experimental data showed that, in the absence of the allosteric partner *Pf*TrpA, *Pf*TrpB activity decreases 3-fold (k_{cat} of 0.31 and 1.0 s⁻¹ for isolated *Pf*TrpB and *Pf*TrpS, respectively).¹⁹ Our reconstructed FELs corresponding to isolated *Pf*TrpB display some similarities to the *Pf*TrpS system (see Figures 3a and 4a). Ser binding at the Aex1 intermediate shifts the conformational ensemble from O toward PC states. Similarly, a population shift toward C states at the Q₂ intermediate is observed. Contrary to the situation for the *Pf*TrpS complex, a single energy minimum is found at the Aex1 and Q₂ intermediates of *Pf*TrpB. In fact, the COMM domain is not able to escape from O states at Ain, PC at Aex1, and C at Q₂ intermediates as the other states are inaccessible. Therefore, *Pf*TrpB in the absence of *Pf*TrpA allosteric regulation has a very limited conformational heterogeneity of the COMM domain, which hampers the multistep reaction pathway. It is also worth mentioning that the stable C states at the *Pf*TrpB–Q₂ intermediate are highly deviated from the reference O-to-C

conformational path (i.e., RMSD larger than 1.5 Å; see Figure 4a). A detailed structural analysis of the isolated wild-type *Pf*TrpB(Q₂) as compared to the *Pf*TrpS(Q₂) complex in C states indicates that the isolated *Pf*TrpB enzyme cannot efficiently sample catalytically competent C states; this is in particular true for the key COMM H6 closure (Figures 4b,c and S8c). Furthermore, the proton transfer catalytic distance K82–Q₂ is also longer than that in *Pf*TrpS (3.9 ± 0.3 Å vs 3.6 ± 0.3 Å). Our simulations have therefore shown that, in the absence of its *Pf*TrpA allosteric partner, the *Pf*TrpB COMM domain displays a restricted conformational landscape, which lacks the ability to easily access O, PC, and C states existing in the allosterically driven conformational ensemble of *Pf*TrpS.

Activating Distal Mutations for Stand-Alone Function Recovers COMM Domain Heterogeneity. *Pf*TrpB was evolved for stand-alone function generating a new variant *Pf*TrpB^{OB2}, which displays a considerably improved catalytic constant with respect to both the isolated wild-type *Pf*TrpB and the *Pf*TrpS complex (k_{cat} of 2.9, 0.31, and 1.0 s⁻¹ for *Pf*TrpB^{OB2}, *Pf*TrpB, and *Pf*TrpS, respectively). It is also worth mentioning that the activity of the evolved *Pf*TrpB^{OB2} decays dramatically in the presence of *Pf*TrpA (k_{cat} of 0.04 s⁻¹).¹⁹

Intrigued by the restricted conformational dynamics of the COMM domain as found in isolated *Pf*TrpB, we decided to explore whether distal mutations introduced in laboratory evolution were able to recover the allosterically driven conformational ensemble of *Pf*TrpS. By comparing the reconstructed FELs for stand-alone *Pf*TrpB^{OB2} and *Pf*TrpS complex along the different reaction intermediates (see Figures 3a and 5a), it becomes evident that the *Pf*TrpB^{OB2} variant recovers the conformational heterogeneity of the COMM domain, characteristic of the allosterically regulated dimeric enzyme. However, interesting differences between both systems are found to be crucial for rationalizing their catalytic activities.

In the resting state (A_{in}), *Pf*TrpB^{OB2} has only the O state accessible. However, as the reaction progresses, a population shift toward PC and C states occurs, as was also observed in the dimeric *Pf*TrpS complex. At A_{ex1}, O and PC states have comparable relative stabilities and are separated by a small energy barrier of ca. 1 kcal/mol, which allows a fast O-to-PC conformational exchange. Similar to the *Pf*TrpS system, at the Q₂ state the allosterically driven conformational ensemble containing O, PC, and C states is recovered. Nevertheless, a substantially lower barrier is observed for the O-to-PC-to-C transition of ca. 2 kcal/mol as compared to *Pf*TrpS complex. This rather small energy barrier allows *Pf*TrpB^{OB2} to easily adopt the catalytically competent C conformation from PC and O states. This high stability of the catalytically relevant C state contrasts with the *Pf*TrpS system where the closed state is ca. 5 kcal/mol higher in energy. Such a difference in the stability of the C state explains the improved catalytic efficiency of the evolved stand-alone variant. The C state of stand-alone *Pf*TrpB^{OB2}(Q₂) has an almost identical degree of closure of the COMM domain as the *Pf*TrpS catalytically competent conformation, and a similar catalytic K82-Q₂ proton transfer distance (see Figures 5b,c and S9). This indicates that the C state of the stand-alone *Pf*TrpB^{OB2}(Q₂) variant is properly preorganized for the reaction.

A remarkable difference between the dimeric *Pf*TrpS(Q₂) complex and stand-alone *Pf*TrpB^{OB2}(Q₂) is found at the PC state, which in the case of the evolved variant is highly digressed from the original path (i.e., RMSD > 1.5 Å, see Figure 5a); therefore, we denote this as the novel PC state. The large deviation arises from an unexpected large-scale conformational change of 14 Å that positions R159, adjacent to the H6 helix of the COMM domain, toward the active site (see Figure 6a). This novel PC conformation has not been previously observed by means of X-ray crystallography. Interestingly, R159 takes over the position previously occupied by L161 at H6 of *Pf*TrpB^{OB2}(Q₂)-C, establishing a cation- π interaction with the indole moiety (Figure 6b). We hypothesize that this novel conformation of R159 may play a role in the catalytic cycle, most probably in properly positioning serine and/or indole for the reaction (see the Supporting Information for a detailed discussion and Figures S10 and 11).

Our findings indicate that stand-alone *Pf*TrpB^{OB2} recovers the COMM domain heterogeneity, thus making O and active C states accessible again. The great stabilization of active C states together with the novel PC conformation displayed by the *Pf*TrpB^{OB2} enzyme variant unravel the increase in catalytic efficiency with respect to the allosterically regulated *Pf*TrpS complex.

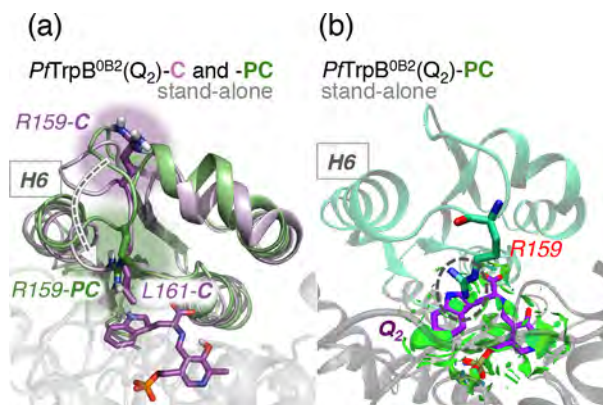


Figure 6. (a) Overlay of the *Pf*TrpB^{OB2} metastable conformations of the closed (C, in pink) and novel partially closed (PC, in green) states at the Q₂ reaction intermediate. The novel PC state revealed by MD simulations presents R159 from the COMM domain located in the active site close to the position previously occupied by L161 in the C state. (b) Representation of the noncovalent interactions (computed with NCIplot)³² at the novel PC state of *Pf*TrpB^{OB2}, highlighting (green surfaces) the cation- π interaction between R159 and the indole moiety of Q₂.

Another relevant aspect is the experimentally observed inactivation of the evolved *Pf*TrpB^{OB2} in the presence of *Pf*TrpA. To study this inactivation, we reconstructed the FEL for the *Pf*TrpA-*Pf*TrpB^{OB2}(Q₂) complex. Surprisingly, our simulations show that the presence of *Pf*TrpA does not restrict the *Pf*TrpB^{OB2} COMM domain heterogeneity, as it is also able to sample the O-to-C exchange. However, the formation of the dimeric complex with *Pf*TrpA induces a population shift toward unproductive closed states (i.e., highly deviated from the reference path), similar to those observed in the isolated *Pf*TrpB system (see Figure S12). Thus, *Pf*TrpA truncates the efficient conformational ensemble of the stand-alone *Pf*TrpB^{OB2} yielding nonproductive closed conformational states of the COMM domain.

COMM Domain Heterogeneity as an Essential Factor in Indole Active-Site Accessibility. Available X-ray structures after E(A_{ex1}) formation display C conformations of the COMM domain. However, our analysis of the allosterically driven conformational ensemble of *Pf*TrpS complex and that of stand-alone *Pf*TrpB^{OB2} provided evidence for a high flexibility of the COMM domain and the ability to visit O, PC, and the catalytically relevant C states. The question that remains is what is the specific role of O conformational states of the COMM domain after Ser binding? One possibility would be to assist in either indole binding or Trp release after the reaction. Experimentally, the Michaelis constant for indole binding in the stand-alone *Pf*TrpB^{OB2} enzyme variant (8.7 μ M) was improved with respect to isolated *Pf*TrpB (77 μ M), but also as compared to the enzyme complex *Pf*TrpS (20 μ M).¹⁹

To elucidate the changes in indole binding and the role played by the COMM domain O states, we reconstructed the FELs at the electrophilic amino acrylate E(A-A) intermediate (Figures 7a,b and S13), and analyzed the available indole substrate access tunnels with the CAVER software.³³ At this A-A intermediate, *Pf*TrpS complex and stand-alone *Pf*TrpB^{OB2} can easily access both O and C states, which are separated by relatively small energy barriers. The analysis of indole access tunnels in both O and C states reveals two

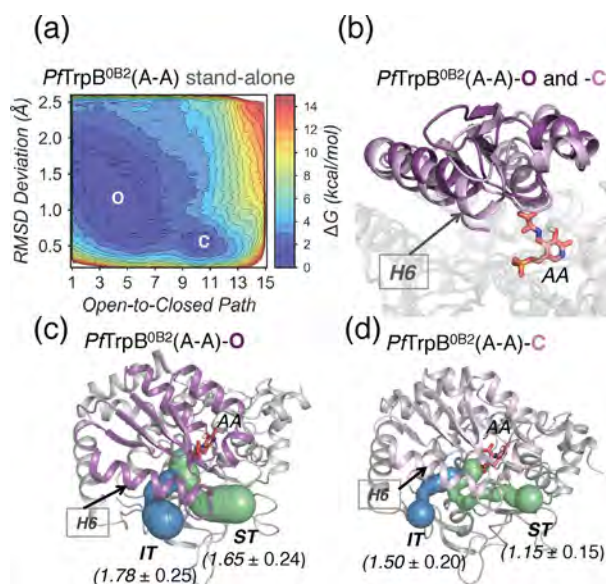


Figure 7. (a) Free energy landscape (FEL) associated with the COMM domain open-to-closed (O-to-C) conformational exchange of stand-alone *PfTrpB*^{OB2} enzyme at A–A reaction intermediate. (b) Overlay of metastable conformations of the O (dark purple) and C state (light purple) at A–A intermediate of *PfTrpB*^{OB2}. (c and d) Internal (IT, in blue) and secondary (ST, in green) tunnels of *PfTrpB*^{OB2} at the O and C states at A–A reaction intermediate computed with CAVER 3.0.³³ The averaged bottleneck radii (in Å) are also shown.

different entry pathways (Figures 7c,d and S13c): the previously described internal tunnel (IT) that connects TrpA and TrpB subunits in *PfTrpS* complex (shown in blue in Figures 7c,d and S13c), and a secondary tunnel (ST)

connecting the active site with a novel entry path not described before (shown in green). C states of the COMM domain yield a narrow bottleneck tunnel radius hampering indole diffusion outside the active site, thus capturing it for efficient catalysis (see Figure 7d). Therefore, the differences in indole binding should be related to O COMM domain states. Interestingly, the isolated wild-type *PfTrpB* is not able to sample the O state, which results in indole access through PC conformations that have a much narrower bottleneck radius (see Figure S13c). This leads to less favorable K_M values for *PfTrpB*, as observed experimentally.

At the O state of *PfTrpS* complex, indole diffusion occurs along the internal TrpA–TrpB tunnel, suggesting that the secondary tunnel (green in Figures 7c,d and S13C) may play a role in Ser binding and/or Trp release. For the stand-alone *PfTrpB*^{OB2} variant, both tunnels show a large bottleneck radius; thus no tunnel preference for indole entrance to the active site is found (see Figure 7c). Altogether, these calculations indicate that the recovery of the allosterically regulated COMM conformational ensemble of *PfTrpS*, especially O state accessibility, is also key for indole binding.

Distal Mutations for Stand-Alone Function Can Be Predicted Computationally.

Our group has recently shown that distal mutations found by DE in the case of multistep retro-aldolase enzymes can be identified with residue-by-residue correlation and proximity analysis tools.¹² Intrigued by the possibility of predicting distal positions for stand-alone function, we applied our Shortest Path Map (SPM) method.¹² This computational tool identifies those pairs of residues that have a higher contribution to the conformational dynamics of the enzyme (see Figure 8 and computational details). We focused our analysis on the *PfTrpS*(Q₂) metadynamics trajectory because of the complete O-to-C conformational exchange sampled in it (see Figure S14 for SPM analysis at

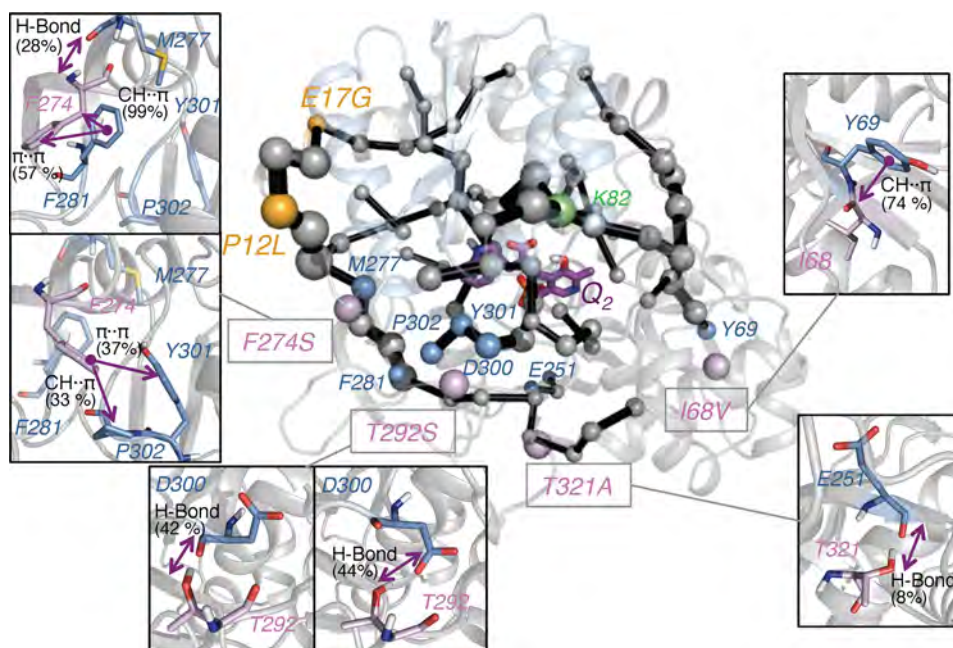


Figure 8. Identification of the amino acids that contribute to the open-to-closed (O-to-C) conformational exchange in *PfTrpS* at (Q₂) intermediate through Shortest Path Map (SPM) analysis.¹² The sizes of the spheres and black edges are indicative of the importance of the position for the *PfTrpS* conformational dynamics. Positions mutated via DE are marked in orange (if they are included in the SPM), or in pink (if they directly interact with SPM residues). SPM residues that interact with the DE positions are marked with blue spheres. For each mutation that interacts with SPM residues, a zoom is provided to show the type of noncovalent interaction and the percentage of interaction time during the simulation.

other reaction intermediates). *PfTrpB*^{OB2} presents six mutations: P12L, E17G, I68V, T292S, F274S, and T321A, from which two were directly predicted by the SPM tool, and three were directly interacting with a SPM position. The specific effect of each isolated mutation on the enzyme activity is not known, except for two of them: T292S and P12L. The most beneficial mutation T292S (3-fold increase in k_{cat} with respect to *PfTrpB*)¹⁹ was previously suggested to modulate COMM domain closure based on the T292–D300 interaction observed in X-ray data.¹⁵ In our metadynamics simulations, the hydrogen bond between T292 and D300 is maintained 86% of the time (see Figure 8). Although position 292 is not directly included in our computed SPM path, position 300 is predicted as key for the COMM domain conformational dynamics. This indicates that by altering the D300 position interactions (for instance, the T292–D300 interaction) the COMM domain closure can be modulated. Interestingly, P12L distal mutation, which was found to have a slight impact on the k_{cat} of the enzyme, is directly identified with high contributions in our SPM analysis (see orange spheres in Figure 8). Similarly, the distal site E17G is also predicted by SPM, suggesting a role on COMM domain conformational heterogeneity. DE positions F274 and I68, although not strictly included in the SPM path, make direct and stable noncovalent interactions with already SPM predicted positions (see blue and pink spheres in Figure 8). For instance, F274 highly forms CH $\cdots\pi$ and $\pi\cdots\pi$ interactions with the SPM residues F281 (maintained 99% of the simulation time), Y301 (37%), and P302 (33%), as well as a hydrogen bond with M277 (28%). Similarly, DE position I68 makes CH $\cdots\pi$ interactions with Y69 (74%), included in SPM. The only DE position that has a minor role in the COMM domain conformational dynamics and makes negligible interactions with SPM residues is T321.

Our new proposed methodology makes use of metadynamics simulations to enforce the sampling of the allosterically regulated O-to-C transition, and identifies which residues present a higher contribution to the O-to-C COMM domain conformational exchange through inter-residue correlation calculations. With this new computational approach, distal positions involved in the allosteric transition can be identified, thus providing a set of key positions for the generation of smart libraries for stand-alone function. This new proposed protocol can be applied to any allosterically regulated system of interest. This study also provides further evidence for the key role played by the enzyme conformational dynamics in the evolution of enhanced catalytic activities, especially in challenging multistep mechanisms such as the one catalyzed by TrpS.³⁴

CONCLUSIONS

Recovering the allosterically driven conformational ensemble existing in multimeric enzymes such as TrpS for stand-alone function is strikingly similar to the dramatic effect induced by distal mutations on the catalytic efficiency of some enzymes. Only those ensembles of conformations that are preactivated for catalysis are selected and stabilized along the evolutionary process. Understanding the differences between both processes is highly appealing for the rational design of enzymes. The present study demonstrates that fine-tuned control of the allosterically driven conformational ensemble of *PfTrpS* plays a key role along its catalytic cycle. By altering the relative stabilities of open, partially closed, and closed conformational states of the COMM domain, each reaction step along the

catalytic pathway can be efficiently optimized. Our free energy calculations on the conformational exchange of the COMM domain indicate that the rate for the open-to-closed conformational transition is relatively fast (in the nanosecond to microsecond time scale) in comparison with the reaction steps and turnover time scale (millisecond to second). However, such transitions are essential for preorganizing the active site pocket to accommodate the different substrates, and efficiently catalyzing Ser and indole coupling for Trp production. Distal mutations, introduced experimentally for converting *PfTrpB* into an efficient stand-alone variant, recover the allosterically regulated conformational ensemble of *PfTrpS*. This enables access to open, partially closed, and closed states of the COMM domain. In the absence of such mutations, the isolated *PfTrpB* lacks the COMM domain conformational heterogeneity, which is required for the challenging multistep catalytic pathway. By careful analysis of the open-to-closed conformational exchange of the COMM domain, and the residues that contribute more to the exchange, distal mutations introduced via directed evolution can be predicted with our recently developed SPM tool. This study shows that, by evaluating the native allosterically regulated conformational ensemble, and the residues that have a higher contribution to the allosteric conformational transition, proficient stand-alone enzyme variants could be rationally designed. The hypothesis that many enzymes are intrinsically regulated allosterically⁷ is inspiring, as it also suggests that our novel computational approach proposed here might be of general use in the computational enzyme design field.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.9b03646.

Computational methods as well as additional tables and figures (PDF)

AUTHOR INFORMATION

Corresponding Authors

*silvia.osuna@udg.edu

*javier.iglesias@udg.edu

ORCID

Miguel A. Maria-Solano: 0000-0002-7837-0429

Javier Iglesias-Fernández: 0000-0001-7773-2945

Silvia Osuna: 0000-0003-3657-6469

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the Generalitat de Catalunya for the emerging group CompBioLab (2017 SGR-1707) and Spanish MINECO for project PGC2018-102192-B-I00. We are grateful for the computer resources, technical expertise, and assistance provided by the Barcelona Supercomputing Center - Centro Nacional de Supercomputación. M.A.M.-S. was supported by the Spanish MINECO for a Ph.D. fellowship (BES-2015-074964), and J.I.-F. was supported by the European Community for Marie Curie fellowship (H2020-MSCA-IF-2016-753045). S.O. is grateful for the funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program

(ERC-2015-StG-679001). We thank Prof. Swart, Dr. Feixas, and Dr. Kress for the helpful discussions and comments on the manuscript.

REFERENCES

- (1) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. *Nature* **2014**, *508* (7496), 331–9.
- (2) Nussinov, R.; Tsai, C. J.; Ma, B. *Annu. Rev. Biophys.* **2013**, *42*, 169–89.
- (3) Freiburger, L. A.; Baettig, O. M.; Sprules, T.; Berghuis, A. M.; Auclair, K.; Mittermaier, A. K. *Nat. Struct. Mol. Biol.* **2011**, *18* (3), 288–94.
- (4) Lisi, G. P.; Loria, J. P. *Curr. Opin. Struct. Biol.* **2017**, *47*, 123–130.
- (5) Tzeng, S. R.; Kalodimos, C. G. *Nature* **2012**, *488* (7410), 236–40.
- (6) Nussinov, R. *Chem. Rev.* **2016**, *116* (11), 6263–6.
- (7) Gunasekaran, K.; Ma, B.; Nussinov, R. *Proteins: Struct., Funct., Genet.* **2004**, *57* (3), 433–443.
- (8) Obexer, R.; Godina, A.; Garrabou, X.; Mittl, P. R. E.; Baker, D.; Griffiths, A. D.; Hilvert, D. *Nat. Chem.* **2017**, *9* (1), 50–56.
- (9) Currin, A.; Swainston, N.; Day, P. J.; Kell, D. B. *Chem. Soc. Rev.* **2015**, *44* (5), 1172–1239.
- (10) Morley, K. L.; Kazlauskas, R. J. *Trends Biotechnol.* **2005**, *23* (5), 231–237.
- (11) Jiménez-Osés, G.; Osuna, S.; Gao, X.; Sawaya, M. R.; Gilson, L.; Collier, S. J.; Huisman, G. W.; Yeates, T. O.; Tang, Y.; Houk, K. N. *Nat. Chem. Biol.* **2014**, *10* (6), 431–436.
- (12) Romero-Rivera, A.; Garcia-Borrás, M.; Osuna, S. *ACS Catal.* **2017**, *7* (12), 8524–8532.
- (13) Hyde, C. C.; Ahmed, S. A.; Padlan, E. A.; Miles, E. W.; Davies, D. R. *J. Biol. Chem.* **1988**, *263*, 17857–71.
- (14) Lee, S. J.; Ogasahara, K.; Ma, J. C.; Nishio, K.; Ishida, M.; Yamagata, Y.; Tsukihara, T.; Yutani, K. *Biochemistry* **2005**, *44* (34), 11417–11427.
- (15) Dunn, M. F. *Arch. Biochem. Biophys.* **2012**, *519* (2), 154–166.
- (16) Barry, S. M.; Kers, J. A.; Johnson, E. G.; Song, L. J.; Aston, P. R.; Patel, B.; Krasnoff, S. B.; Crane, B. R.; Gibson, D. M.; Loria, R.; Challis, G. L. *Nat. Chem. Biol.* **2012**, *8* (10), 814–816.
- (17) Kieffer, M. E.; Repka, L. M.; Reisman, S. E. *J. Am. Chem. Soc.* **2012**, *134* (11), 5131–5137.
- (18) Patel, R. N. *Biomolecules* **2013**, *3* (4), 741–777.
- (19) Buller, A. R.; Brinkmann-Chen, S.; Romney, D. K.; Herger, M.; Murciano-Calles, J.; Arnold, F. H. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (47), 14599–14604.
- (20) Herger, M.; van Roye, P.; Romney, D. K.; Brinkmann-Chen, S.; Buller, A. R.; Arnold, F. H. *J. Am. Chem. Soc.* **2016**, *138* (27), 8388–8391.
- (21) Murciano-Calles, J.; Romney, D. K.; Brinkmann-Chen, S.; Buller, A. R.; Arnold, F. H. *Angew. Chem., Int. Ed.* **2016**, *55* (38), 11577–11581.
- (22) Buller, A. R.; van Roye, P.; Murciano-Calles, J.; Arnold, F. H. *Biochemistry* **2016**, *55* (51), 7043–7046.
- (23) Romney, D. K.; Murciano-Calles, J.; Wehrmuller, J. E.; Arnold, F. H. *J. Am. Chem. Soc.* **2017**, *139* (31), 10769–10776.
- (24) Buller, A. R.; van Roye, P.; Cahn, J. K. B.; Scheele, R. A.; Herger, M.; Arnold, F. H. *J. Am. Chem. Soc.* **2018**, *140* (23), 7256–7266.
- (25) Boville, C. E.; Scheele, R. A.; Koch, P.; Brinkmann-Chen, S.; Buller, A. R.; Arnold, F. H. *Angew. Chem., Int. Ed.* **2018**, *57* (45), 14764–14768.
- (26) Hioki, Y.; Ogasahara, K.; Lee, S. J.; Ma, J.; Ishida, M.; Yamagata, Y.; Matsuura, Y.; Ota, M.; Ikeguchi, M.; Kuramitsu, S.; Yutani, K. *Eur. J. Biochem.* **2004**, *271* (13), 2624–35.
- (27) Niks, D.; Hilario, E.; Dierkers, A.; Ngo, H.; Borchardt, D.; Neubauer, T. J.; Fan, L.; Mueller, L. J.; Dunn, M. F. *Biochemistry* **2013**, *52* (37), 6396–6411.
- (28) Barends, T. R. M.; Domratcheva, T.; Kulik, V.; Blumenstein, L.; Niks, D.; Dunn, M. F.; Schlichting, I. *ChemBioChem* **2008**, *9* (7), 1024–1028.
- (29) Henzler-Wildman, K.; Kern, D. *Nature* **2007**, *450* (7172), 964–972.
- (30) Barducci, A.; Bonomi, M.; Parrinello, M. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (5), 826–843.
- (31) Laio, A.; Gervasio, F. L. *Rep. Prog. Phys.* **2008**, *71* (12), 22.
- (32) Contreras-Garcia, J.; Johnson, E. R.; Keinan, S.; Chaudret, R.; Piquemal, J. P.; Beratan, D. N.; Yang, W. J. *Chem. Theory Comput.* **2011**, *7* (3), 625–632.
- (33) Chovancova, E.; Pavelka, A.; Benes, P.; Strnad, O.; Brezovsky, J.; Kozlikova, B.; Gora, A.; Sustr, V.; Klvana, M.; Medek, P.; Biedermannova, L.; Sochor, J.; Damborsky, J. *PLoS Comput. Biol.* **2012**, *8* (10), e1002708.
- (34) Maria-Solano, M. A.; Serrano-Hervas, E.; Romero-Rivera, A.; Iglesias-Fernandez, J.; Osuna, S. *Chem. Commun.* **2018**, *54* (50), 6622–6634.

Deciphering the Allosterically-driven Conformational Ensemble in Tryptophan Synthase Evolution.

Miguel A. Maria-Solano,^[a] Javier Iglesias-Fernández,^{*[a]} and Sílvia Osuna^{*[a,b]}

^a Miguel A. Maria-Solano, Javier Iglesias-Fernández, and Prof. Sílvia Osuna

Department of Chemistry

Institute of Computational Chemistry and Catalysis

University of Girona

Girona 17003, Catalonia, Spain.

E-mail: Silvia.osuna@udg.edu

^bProf. S. Osuna

ICREA

Barcelona 08010, Catalonia, Spain

Supporting information

SI Table of Contents

Computational methods.....	S5
SI Tables.....	S6
SI Figures.....	S7
References.....	S17

Computational methods:

Molecular Dynamics Simulations

System preparation: The crystal structure of the open *PfTrpS* enzyme, with PDB accession code 1WDW, was used as a starting structure for all the simulations. The *PfTrpS* heterodimeric complex used for this study contains one *PfTrpA* subunit and one *PfTrpB* subunit. The *PfTrpB* wild-type isolated structure was generated by manually removing the *PfTrpA* subunit from the *PfTrpS* PDB. Point mutations in the *PfTrpB*^{OB2} variant were introduced with the RosettaDesign software¹. MD simulation parameters for the reaction intermediates (IGP, G3P, Ain, Aex1, A-A, Q₂) were generated with the antechamber module of AMBER16² using the general amber force-field (GAFF)³ with partial charges set to fit the electrostatic potential generated at the HF/6-31G(d) level with the RESP model.⁴ These charges were calculated using the Gaussian09 software package. Different reaction intermediates were introduced to the open *PfTrpS*, *PfTrpB*, and *PfTrpB*^{OB2} structures by alignment to available X-ray structures (see Table S1). A total of 12 systems (3 enzyme variants with 4 different reaction intermediates, Ain, A-A, Aex1, Q₂) were generated. For the *PfTrpS*, IGP was introduced in the *PfTrpA* subunit at Ain, Aex1 and A-A *PfTrpB* reaction intermediates whereas G3P at Q₂.

To study the inactivation effect of *PfTrpA* on the evolved *PfTrpB*^{OB2} enzyme variant, a system containing both subunits was set-up by structural alignment of the *PfTrpB*^{OB2} to the wild-type structure with PDB code 1WDW. The dimeric *PfTrpB*₂ system was also set-up from the 1WDW structures by manually removing the *PfTrpA* subunits. The reaction intermediates studied were introduced as described above.

Molecular Dynamics Simulations: Long-timescale MD-simulations were performed using an in-house GPU-cluster. Each enzyme system was immersed in a pre-equilibrated cubic box with a 10-Angstrom buffer of TIP3P water molecules,⁵ resulting in the addition of approximately 15.000 water molecules. Afterwards, the systems were neutralized by the addition of explicit counterions (Na⁺ or Cl⁻). All subsequent calculations were done using a modification of the amber99 force field (ff14SB).⁶ A two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on solute by a harmonic potential with a force constant of 500 kcal mol⁻¹Å⁻², and the second stage is an unrestrained minimization of all the atoms in the simulation cell. All systems were gently heated using seven 50 ps steps, incrementing the temperature 50 K each step (0-350 K) under constant-volume and periodic-boundary conditions. Water molecules are treated with the SHAKE algorithm such that the angle between the hydrogen atoms is kept fixed. Long-range

electrostatic effects are modeled using the particle-mesh-Ewald method.⁷ An 8Å cutoff was applied to Lennard-Jones and electrostatic interactions. Decreasing harmonic restraints were applied to the protein (210, 165, 125, 85, 45, 10 kcal/mol Å²) during the thermal equilibration, with the Langevin scheme used to control and equalize the temperature. The time step is kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system is then equilibrated without restraints for 2 ns with a 2 fs timestep at a constant pressure of 1 atm and temperature of 350 K. After equilibration, a 500 ns of production MD simulation was performed for each system in the NVT ensemble and periodic-boundary conditions. Production runs were performed with the AMBER16 software.²

Well-tempered Metadynamics simulations with Path Collective Variables

Conventional MD simulations can only sample limited time scales, therefore, obviating important information regarding the conformational dynamics of the enzymes. To overcome the time scale gap, while keeping full atomistic resolution of our systems, we relied on enhanced sampling techniques, in particular the metadynamics approach.⁸⁻¹⁰ Metadynamics simulations allow to reconstruct the free energy landscape (FEL) as a function of a few degrees of freedom, often referred to as collective variables (CVs). In this work we used path CVs, an extremely powerful approach to study transitions between different conformational states. Here, a path of conformations from open (O) to closed (C) states was obtained by linear interpolation between the X-ray available data. Specifically, $s(R)$ represents the progression along the path, while $z(R)$ measures the distance from the ideal path provided. In this work, a path of 15 conformations from an open ($s(R) = 1$) to closed state ($s(R)=15$), was generated. Guided by structural information we restricted the path of structures to the alpha carbons of the COMM domain (residues 97-184) and a region located at the base of COMM domain (residues 282-305), see Figure S1. The λ parameter was computed as 2.3 multiplied by the inverse of the mean square displacement between successive frames, 80. Metadynamics enhances the sampling of the conformational space by adding external energy potentials to a selected set of collective variables (CVs). This bias potential gradually overcomes energy barriers allowing for efficient exploration of different conformational states. After certain simulation time, the biasing potential corresponds to the negative of the free energy surface and, therefore, all possible states are equally sampled. More exhaustive discussions of the method can be found elsewhere.¹¹⁻¹²

The PLUMED2 software package¹³ together with the GROMACS 5.1.2 code¹⁴ were used to carry out the metadynamics simulations. Here, the well-tempered¹⁵ version of metadynamics was used to improve convergence of the FES reconstruction. Metadynamics simulations were started from the equilibrated structures obtained from previous classical simulations of the *PfTrpS*, *PfTrpB*, and *PfTrpB*^{OB2} enzyme variants, all in four different reaction intermediates (Ain, Aex1, A-A, and Q₂). *PfTrpA-PfTrpB*^{OB2} and *PfTrpB*₂ system were only studied in the Q₂ and Ain intermediates, respectively. Initial Gaussian potentials of height 0.15 kcal mol⁻¹, deposited every 2 ps of MD simulation at 350 K, were gradually decreased on the basis of the well-tempered adaptative bias with a bias factor of 10. The adaptive Gaussian width scheme,⁹ in which hills variance adapt to local properties of the free-energy surface, was used. The multiple-walker extension,¹⁶ which uses several replicas of the same system biasing identical CVs, was used to increase the sampling of the conformational space and to increase the convergence of individual

free-energy profiles. In this approach, multiple walkers are run in parallel and each walker replica reads the Gaussian potentials deposited by the others during the simulation time, in such a way they are all dependent of each other. The free energy landscape associated with the metadynamics CVs is estimated by summing the Gaussian potentials deposited by all walker replicas as a function of the CVs values.

After an initial metadynamics run, we extracted ten snapshots for each system covering approximately all the conformational space available. Then, multiple-walkers metadynamics simulations with 10 replicas were computed. Each replica was run for 50 – 100 ns, giving a total of 500-1000 ns of simulation time per system (i.e. accumulated simulation time of *ca.* 7 microseconds, see Figure S3). The convergence of the recovered FEL was evaluated by monitoring the energy difference between selected regions of the conformational surface along simulation time (see S4 and S5). In particular, the regions selected are the local energy minima (e.g. energy differences between **O** and **C** local energy minima). For the systems where only one local energy minima was found, the energy differences were computed between the local energy minima and a higher in energy region. Finally, a set of structures from each local energy minima were clustered to obtain representative metastable conformations (Figure S7). The local energy minima and the associated representative structures were labeled as open (**O**), partially closed (**PC**) and closed (**C**) accordingly with the $s(R)$ CV values; (**O**)=1-5, (**PC**)=5-10 and (**C**)=10-15.

Molecular Dynamics Simulations to study the role of Arg159

A starting configuration of *Pf*TrpB^{OB2} enzyme variant with Arg159 pointing to the active site (IN conformation) was obtained from the metastable structure *Pf*TrpB^{OB2}(Q₂)-PC obtained from the metadynamics simulations. Conventional Arg159 conformation (OUT conformation) was obtained from previously described molecular dynamics simulations at the A_{in} reaction intermediate. Ser and Trp initial conformations were obtained from X-ray structures with pdb accession codes 5IXJ and 5DW3, respectively. Threonine molecule present in 5IXJ was manually converted into Serine. Five systems were set-up to study the effect of Arg159 novel conformation in the catalytic cycle of TrpB: (1) R159 IN conformation at A_{in}, (2) R159 IN conformation at Ser, (3) R159 IN conformation at Trp (4) R159 OUT conformation at Ser, and (5) R159 OUT conformation at Trp. Five replicas of 800 ns were run for each system using the computational set-up described above.

The results indicate that R159 is stable in a PC novel conformation when L-Ser or L-Trp are bound in the *Pf*TrpB^{OB2} active site. On the contrary, absence of the substrate or product (*Pf*TrpB^{OB2}(A_{in})) destabilizes this conformation as R159 tends to leave the active site in three out of a total of five simulation replicas (Figure S10). Analysis of the catalytic distance between Ser and PLP cofactor reveals the stabilization of the substrate molecule within the active site of the enzyme in this PC novel conformation. Histogram analysis of the catalytic distance highlighted an increased population of conformations at short catalytic distances (< 5 Å), when R159 points towards the active site of the enzyme (Figure S11 A). Interestingly, slightly longer distances (> 8 Å) corresponding to product release are only observed when the R159 residue points towards the solvent (Figure S11 B). Although more conformation

sampling is required, these results suggest that this novel conformation of Arg159 plays a role in L-Ser positioning for the catalytic reaction.

CAVER Analysis

The program CAVER 3.0¹⁷ was used to analyze the available tunnels for substrate entrance to the *Pf*TrpS, *Pf*TrpB, and *Pf*TrpB^{OB2} active site at A-A intermediate. 100 snapshots from each local energy minima from the metadynamics trajectories were selected and aligned for analysis. For this study, a spherical probe of 0.9 Å radius was selected with a weighting coefficient of 1, and clustering threshold of 12.0. The starting point for the calculation was chosen at indole active site coordinates by alignment of the metastable structures at A-A intermediate with the X-ray structure (PDB ID 4HPX), which contains the A-A intermediate co-crystalized with an indole analogue.

Shortest Path Map analysis

The first step of the Shortest Path Map (SPM) calculation relies on the construction of a graph based on the computed mean distances and correlation values observed along the MD simulations. For each residue of the protein a node is created and centered on the C-alpha if a neighboring residue displays a mean distance of less than 6 Å along the simulation time. The length of the line connecting both residues is drawn according to their correlation value ($d_{ij} = -\log |C_{ij}|$). Larger correlation values (closer to 1 or -1) will have shorter edge distances, whereas less correlated residue pairs (values closer to 0) will have edges with long distances. At this point, we make use of Dijkstra algorithm as implemented in graph module¹⁸ to identify the shortest path lengths. The algorithm goes through all nodes of the graph and identifies which is the shortest path to go from the first until the last protein residue. The method therefore identifies which are the edges of the graph that are shorter, i.e. more correlated, and that are more frequently used for going through all residues of the protein, i.e. they are more central for the communication pathway. More details about our SPM tool can be found in our recent publication in ACS Catalysis.

19

Hydrogen bond and aromatic interaction analysis.

Conserved hydrogen bonds and aromatic interactions along the metadynamics simulations for the *Pf*TrpS enzyme at Q₂ intermediate were analyzed with the cptraj module of the AmberTools16.² For aromatic interactions, only the hydrogens of Phe, Tyr, and Trp residues with an angle and distance cutoff of 30° and 5 Å, respectively, were considered.

Metadynamics calculation of the *Pf*TrpB in the dimer form (*Pf*TrpB₂).

Although *Pf*TrpB enzyme exists as a dimer in solution,²⁰ calculations were performed with monomeric *Pf*TrpB to reduce the computational cost associated to the study. This simplification was based on the following reasoning:

- The higher *Pf*TrpB₂ stability (compared to mesophilic TrpB enzymes), is mainly caused by the great number of hydrogen bonds involved in the main chains of *Pf*TrpB (monomeric form) instead of the hydrophobic interactions, which are indeed the stabilization forces in the TrpB-TrpB interface.²⁰

- The dimeric TrpB-TrpB interface is situated far away from the H6 of the COMM domain and the active site of the enzyme.
- There is no allosteric communication reported between TrpB subunits.

Howbeit, a metadynamics simulation of the *Pf*TrpB₂ complex enzyme at the Ain reaction intermediate was performed, under the same methodological conditions as described above, to validate the results. Figure S15 shows how the FELs of the *Pf*TrpB and *Pf*TrpB₂ complex are remarkably similar. In both cases the O conformational states are highly favored, the PC states are low in energy and the C states are inaccessible. Therefore, we conclude that monomeric TrpB is a suitable entity to study the allosteric effects exerted by *Pf*TrpA and DE mutations.

Table S1. Tryptophan synthase X-Ray crystallographic data collected from bibliography.

PDB ID	Enzyme	Subunit states	TrpA ligand	TrpB ligand	Ref.
1WDW	<i>Pf</i> TrpS	α O β O	-	Ain	21
1V8Z	<i>Pf</i> TrpB	β O	-	Ain	20
5DVZ	<i>Pf</i> TrpB	β O	-	Ain	22
5IXJ	<i>Pf</i> TrpB	β O	-	Ain, L-Thr	23
6AMC	<i>Pf</i> TrpB ^{4D11}	β O	-	Ain	24
6AM7	<i>Pf</i> TrpB ^{2b9}	β O	-	Ain	24
6CUZ	<i>Pf</i> TrpB ^{7E6}	β PO	-	Ain	25
6CUV	<i>Pf</i> TrpB ^{7E6}	β O	-	Ain	25
		β C	-	A-A analogue	
6CUT	<i>Pf</i> TrpB ^{7E6}	β C	-	A-A-analogue	25
5DW0	<i>Pf</i> TrpB	β PC	-	Aex1	22
5VM5	<i>Pf</i> TrpB ^{2b9}	β PC	-	Aex1	24
		β C	-	A-A	
6AMH	<i>Pf</i> TrpB ^{4D11}	β PC	-	Aex1	24
6AMI	<i>Pf</i> TrpB ^{4D11}	β PC	-	L-Trp	24
5DW3	<i>Pf</i> TrpB	β PC	-	L-Trp	22
5T6M	<i>Pf</i> TrpB	β PC	-	β -MeTrp	26
6AM8	<i>Pf</i> TrpB ^{2b9}	β C	-	Aex2	24
		β PC	-	L-Trp	
4HT3	<i>St</i> TrpS	α C β PC	F9F	Ain	27
2CLL	<i>St</i> TrpS	α C β PC	F9F	Aex1	28
4HN4	<i>St</i> TrpS	α C β C	F9F	A-A	27
4HPX	<i>St</i> TrpS	α C β C	F9F	A-A, benzimidazole	27
3CEP	<i>St</i> TrpS	α C β C	IGP	Q analogue	29
3PR2	<i>St</i> TrpS	α C β C	F9F	Q analogue	30
5CGQ	<i>St</i> TrpS	α C β PC	F9F	L-Trp	

Supplemental Figures

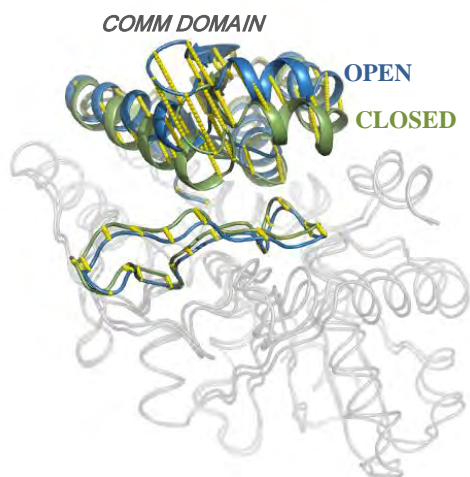


Figure S1. Schematic representation of the Open-to-Closed (O-to-C) path generated from the open (PDB ID: 1WDW) in blue to the closed (PDB ID: 3CEP) in green X-ray structures. The alpha carbon atoms included in the path (i.e. COMM domain (97-184) and 282-305 region) are shown as yellow spheres.

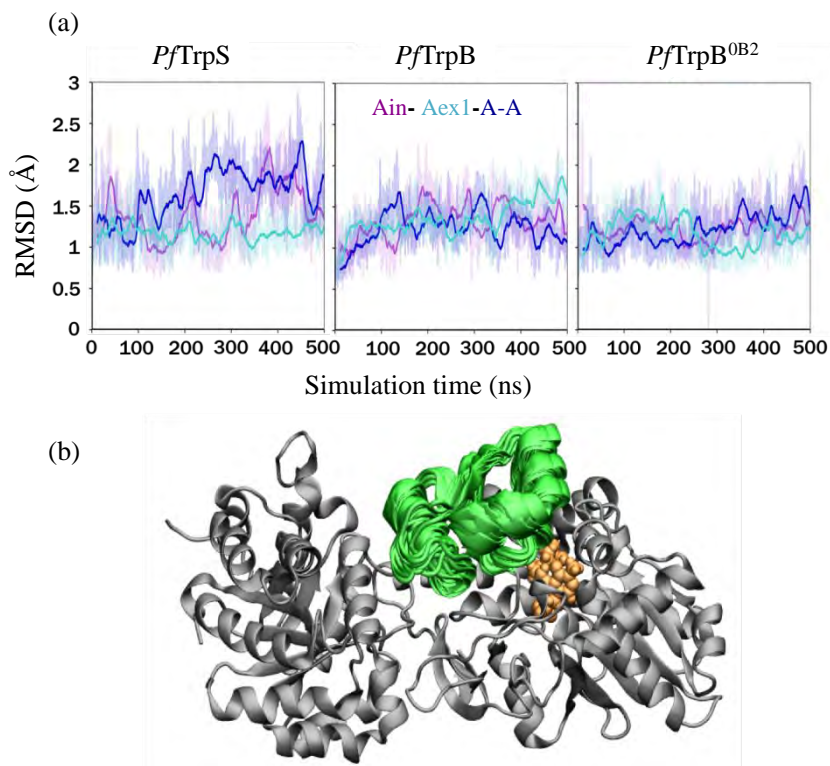


Figure S2. Molecular Dynamics simulations of *PfTrpS*, *PfTrpB*, and *PfTrpB^{OB2}* with different reaction intermediates (Ain, A-A, Aex1). (a) RMSD values calculated for the COMM domain within 500 ns conventional MD simulations. (b) Superimposition of conformations from the MD simulation of the *PfTrpS*-Ain system. The protein structure, the COMM domain, and the PLP cofactor are colored in grey, green, and orange respectively.

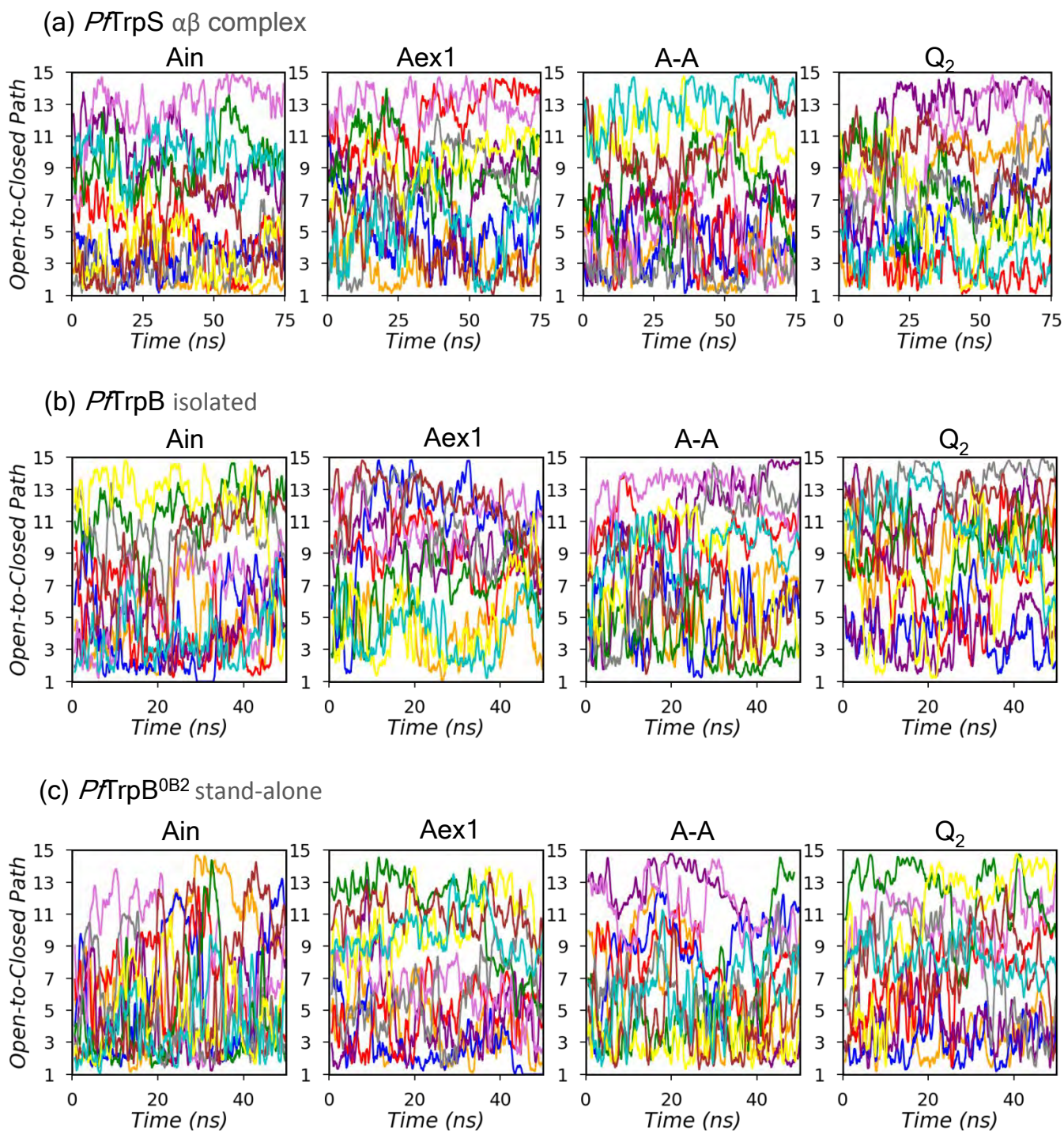


Figure S3. Representation of the Open-to-Closed (**O**-to-**C**) path of conformations (1-to-15) sampled along the metadynamics simulations of the *Pf*TrpS $\alpha\beta$ complex (a), *Pf*TrpB wild-type isolated (b) and *Pf*TrpB^{OB2} stand-alone (c) enzymes at different reaction intermediates (Ain, Aex1, A-A and Q₂). The different color lines represent each metadynamics walker replica (1-10) run in parallel. As it is shown, the (**O**-to-**C**) path conformational space is highly sampled with multiple crossing events among the walker replicas.

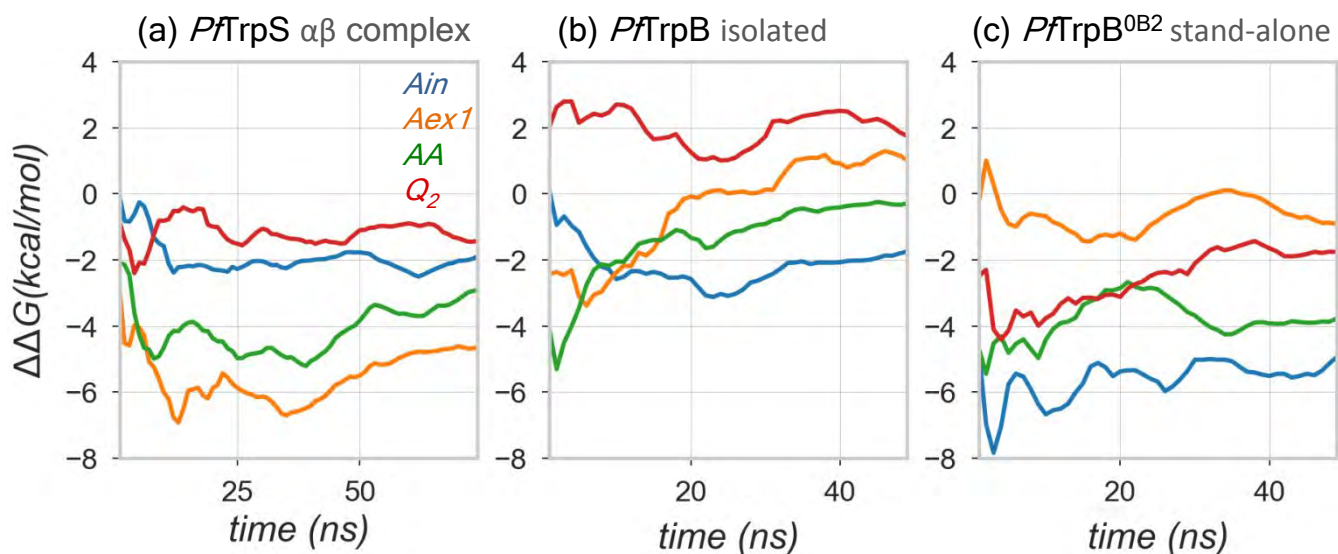


Figure S4. Estimate of the differences in energy between selected regions of the FEL surface along the metadynamics simulations for the *PfTrpS* $\alpha\beta$ complex (a), *PfTrpB* wild-type isolated (b), and *PfTrpB*^{OB2} stand-alone (c) enzymes at different intermediates (Ain, Aex1, A-A and Q₂). Each line represents the $\Delta\Delta G$ mean value of the 10 walker replicas.

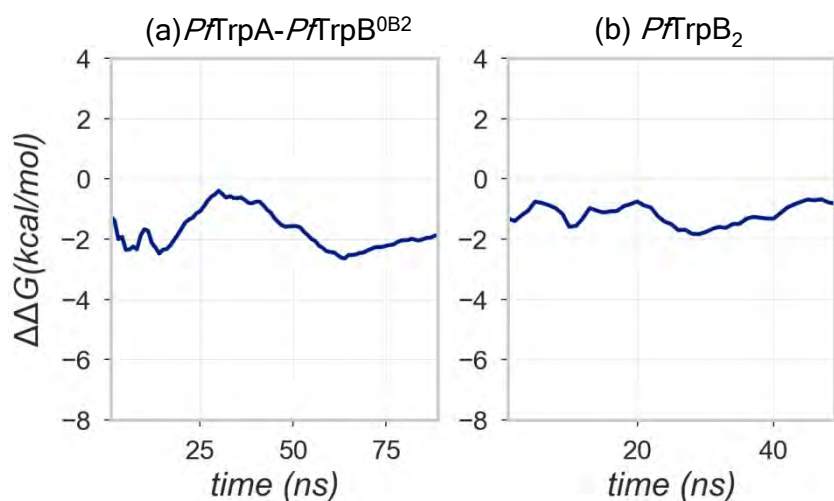


Figure S5. Estimate of the differences in energy between selected regions of the FEL surface along the metadynamics simulations for the *PfTrpA-PfTrpB*^{OB2} complex at Q₂ reaction intermediate (a), and the *PfTrpB*₂ complex at Ain reaction intermediate (b). The line represents the $\Delta\Delta G$ mean value of the 10 walker replicas.

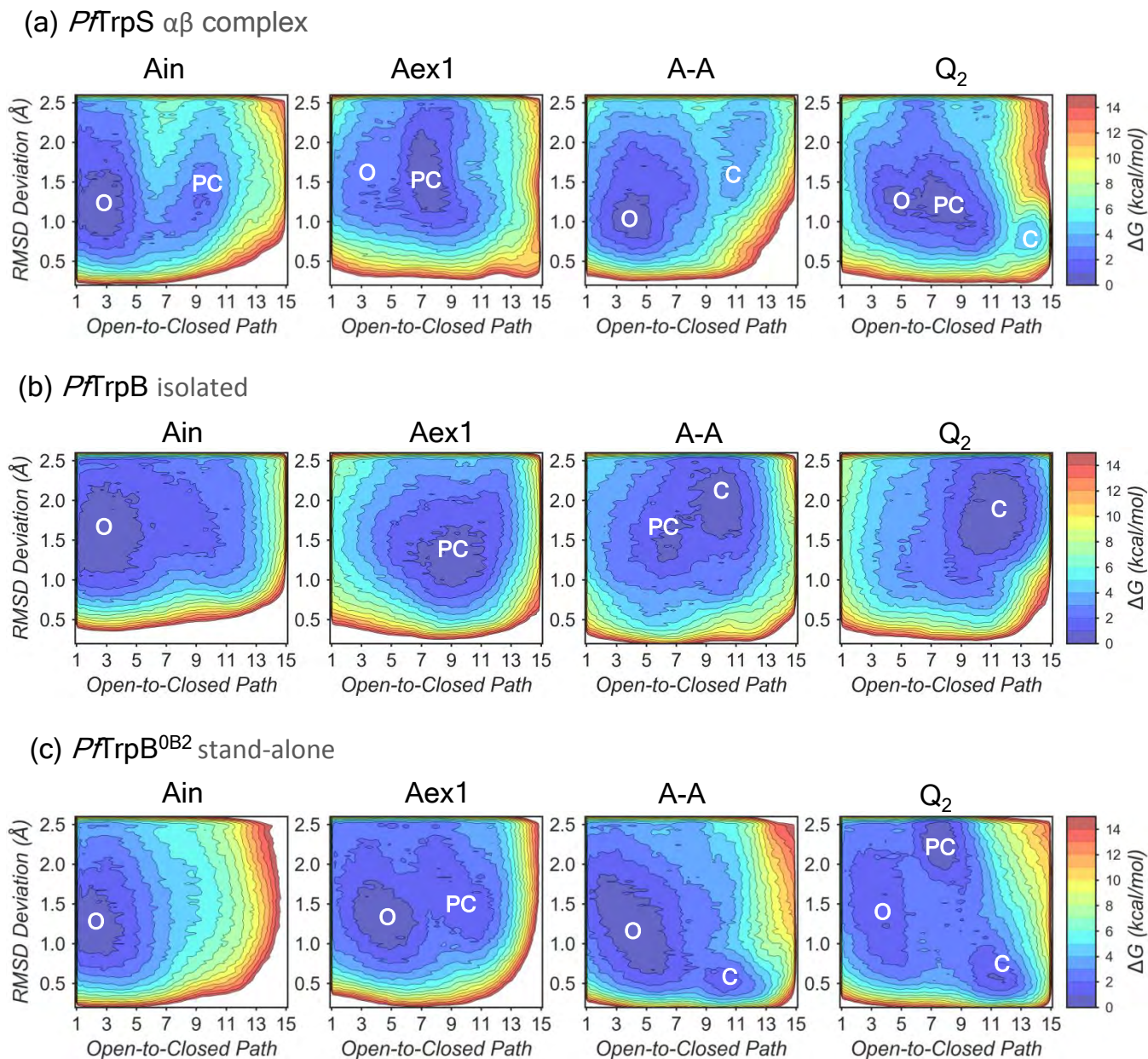
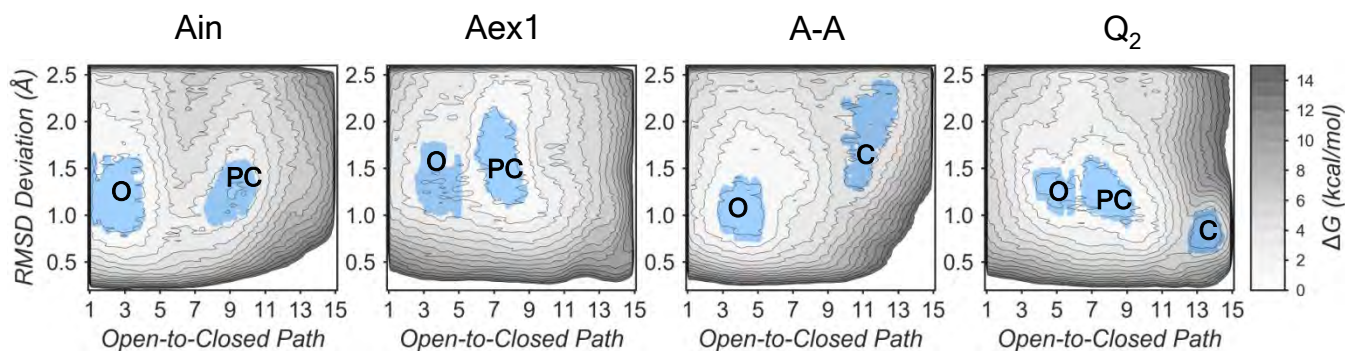
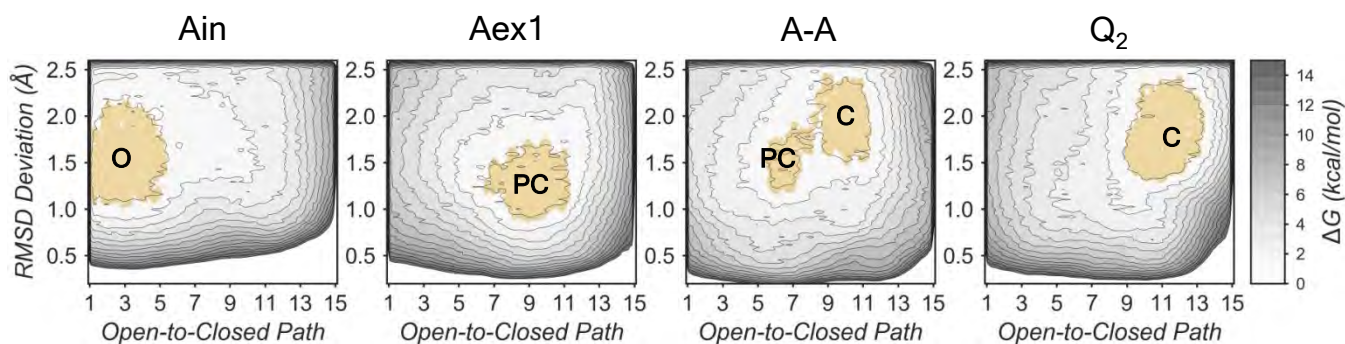


Figure S6. FELs associated with the COMM domain Open-to-Closed (O-to-C) conformational exchange of the *Pf*TrpS $\alpha\beta$ complex (a), *Pf*TrpB wild-type isolated (b) and *Pf*TrpB^{OB2} stand-alone (c) enzymes at different reaction intermediates (Ain, Aex1, A-A and Q₂). The FELs are estimated by summing the Gaussian potentials deposited as a function of the CVs during the metadynamics simulations.

(a) *Pf*TrpS $\alpha\beta$ complex



(b) *Pf*TrpB isolated



(c) *Pf*TrpB^{OB2} stand-alone

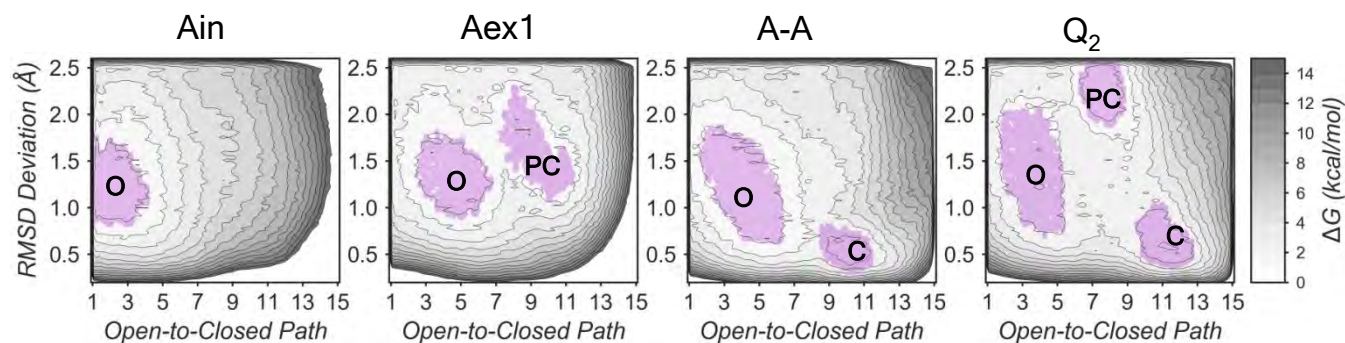


Figure S7. Projection of a set of conformations that correspond to the local energy minima coordinates from the metadynamics simulation on the FELs associated to the Open-to-Closed (O-to-C) conformational exchange of the *Pf*TrpS $\alpha\beta$ complex in blue dots (a), *Pf*TrpB wild-type isolated in yellow dots (b) and *Pf*TrpB^{OB2} stand-alone in violet dots (c). The representative metastable conformations of each local energy minima were obtained by clustering these sets of structures.

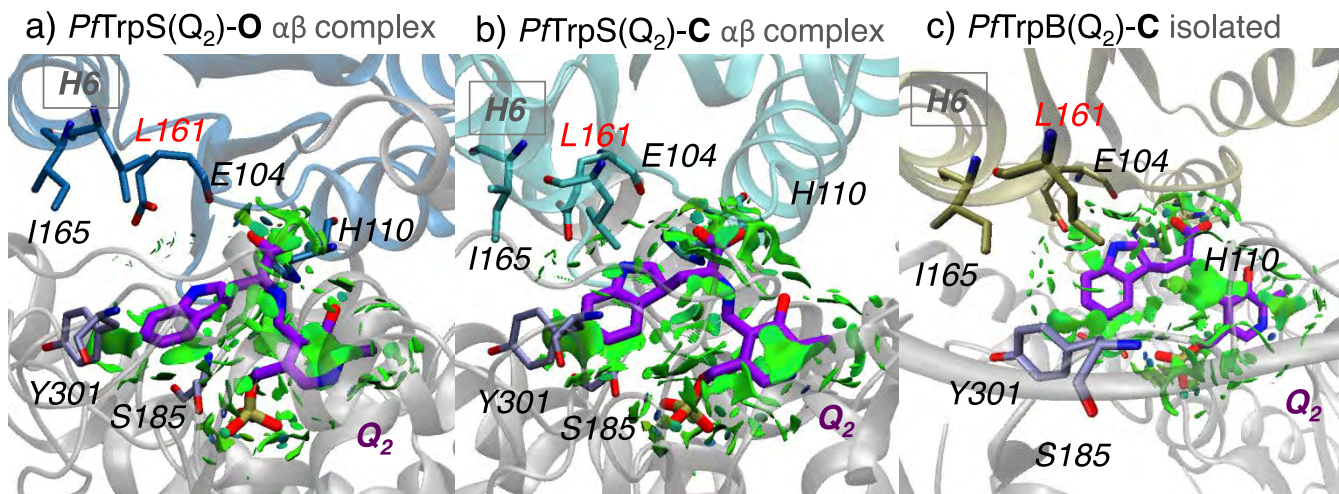


Figure S8. Schematic representation of the non-covalent interactions for the *PfTrpS(Q₂)-O* αβ complex, *PfTrpS(Q₂)-C* αβ complex and *PfTrpB(Q₂)-C* wild-type isolated representative metastable conformations from the metadynamics simulations, calculated with the computational tool NCIplot.³¹⁻³²

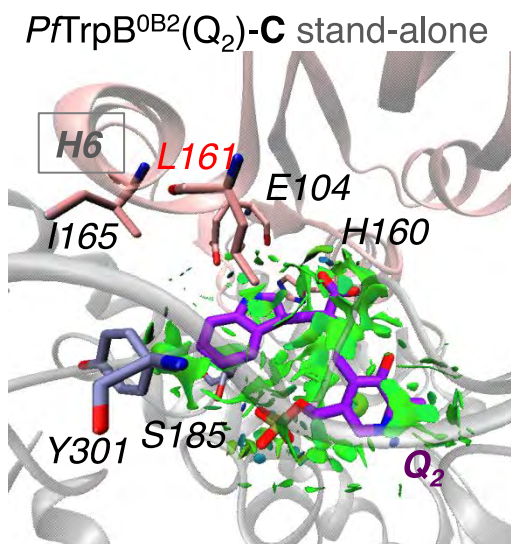


Figure S9. Schematic representation of the non-covalent interactions for the *PfTrpB^{OB2}(Q₂)-C* stand-alone representative metastable conformation from the metadynamics simulations, calculated with the computational tool NCIplot.³¹⁻³²

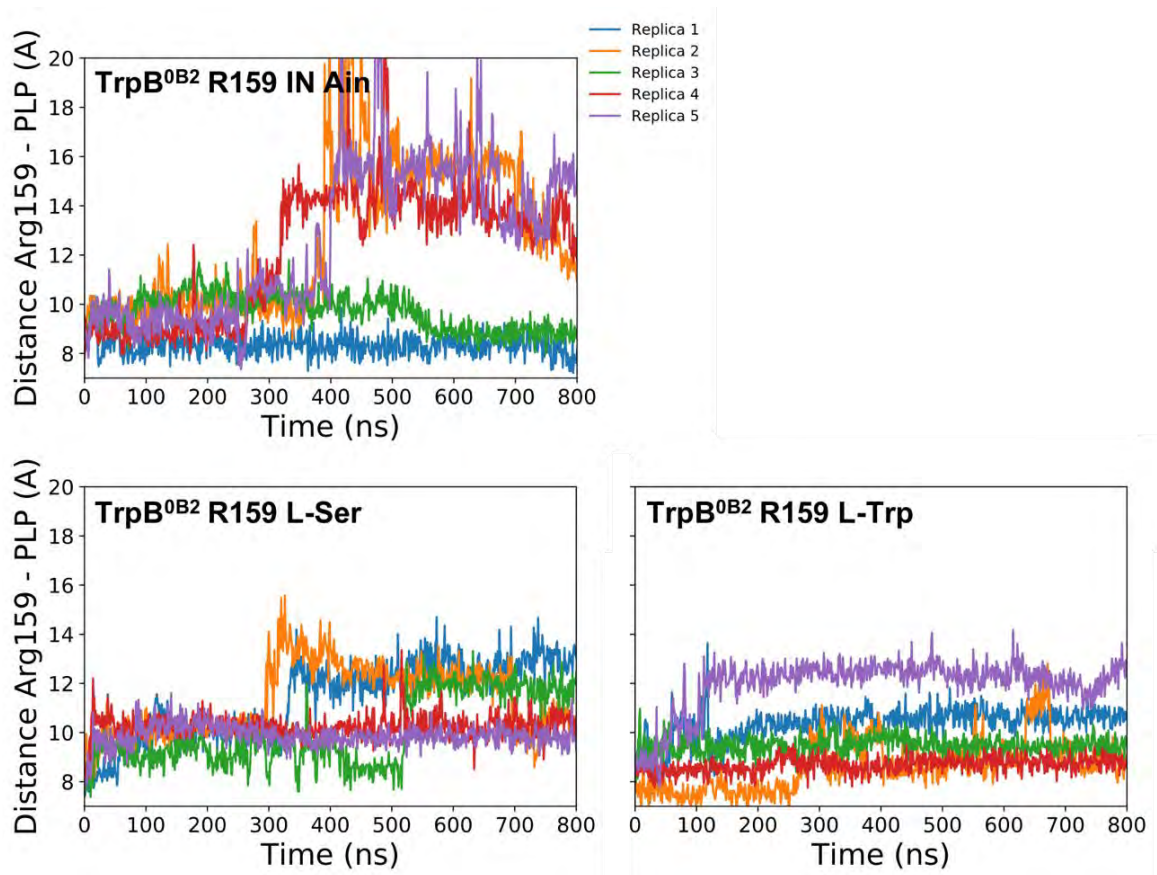


Figure S10. Distance between Arg159 and PLP residues along simulation time for the TrpB^{0B2} Ain and L-Ser/L-Trp bound states. Each simulation replica is shown in a different color.

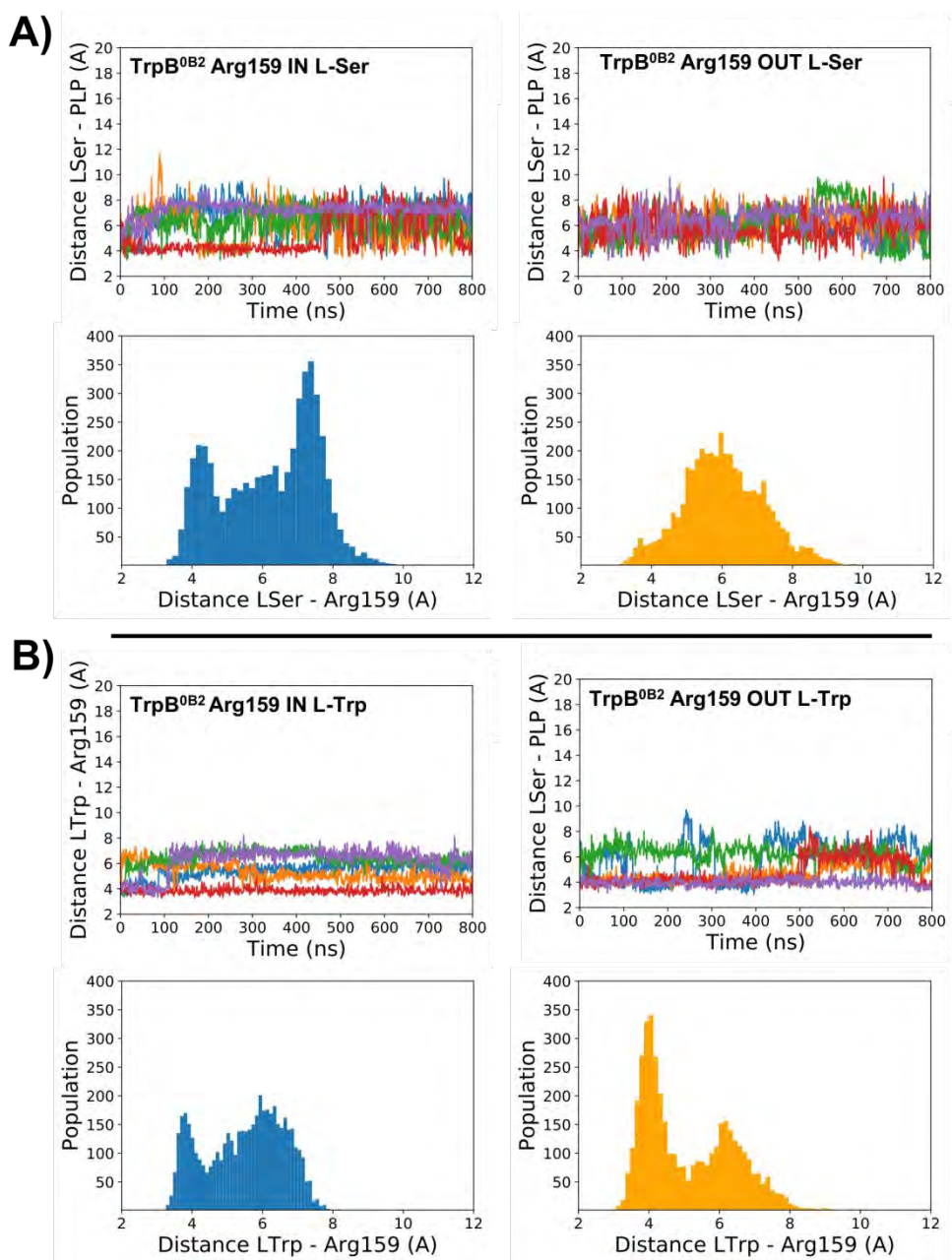


Figure S11. Distance between L-Ser (A) / L-Trp (B) and PLP residues along simulation time for the TrpB^{0B2} Arg159 residue inside and outside the active site. Each simulation replica is shown in a different color. Histogram analysis are provided for each system by considering the information of all simulation replicas altogether.

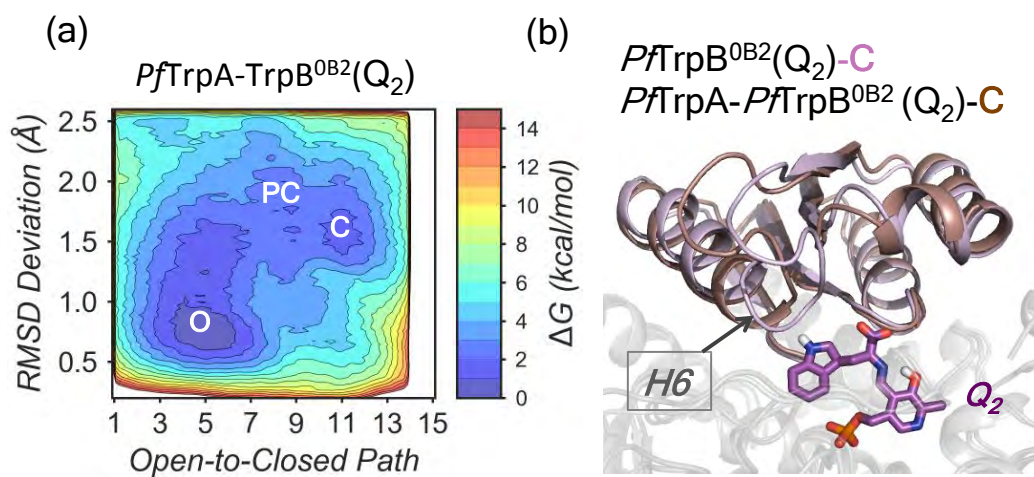


Figure S12. (a) Free energy landscape (FEL) associated to the COMM domain Open-to-Closed (O-to-C) conformational exchange of stand-alone *Pf*TrpA-*Pf*TrpB^{OB2} enzyme at Q₂ reaction intermediate. (b) Overlay of meta-stable conformations of the C states at Q₂ intermediate for-*Pf*TrpB^{OB2} dark (in pink) and *Pf*TrpA-*Pf*TrpB^{OB2} (brown).

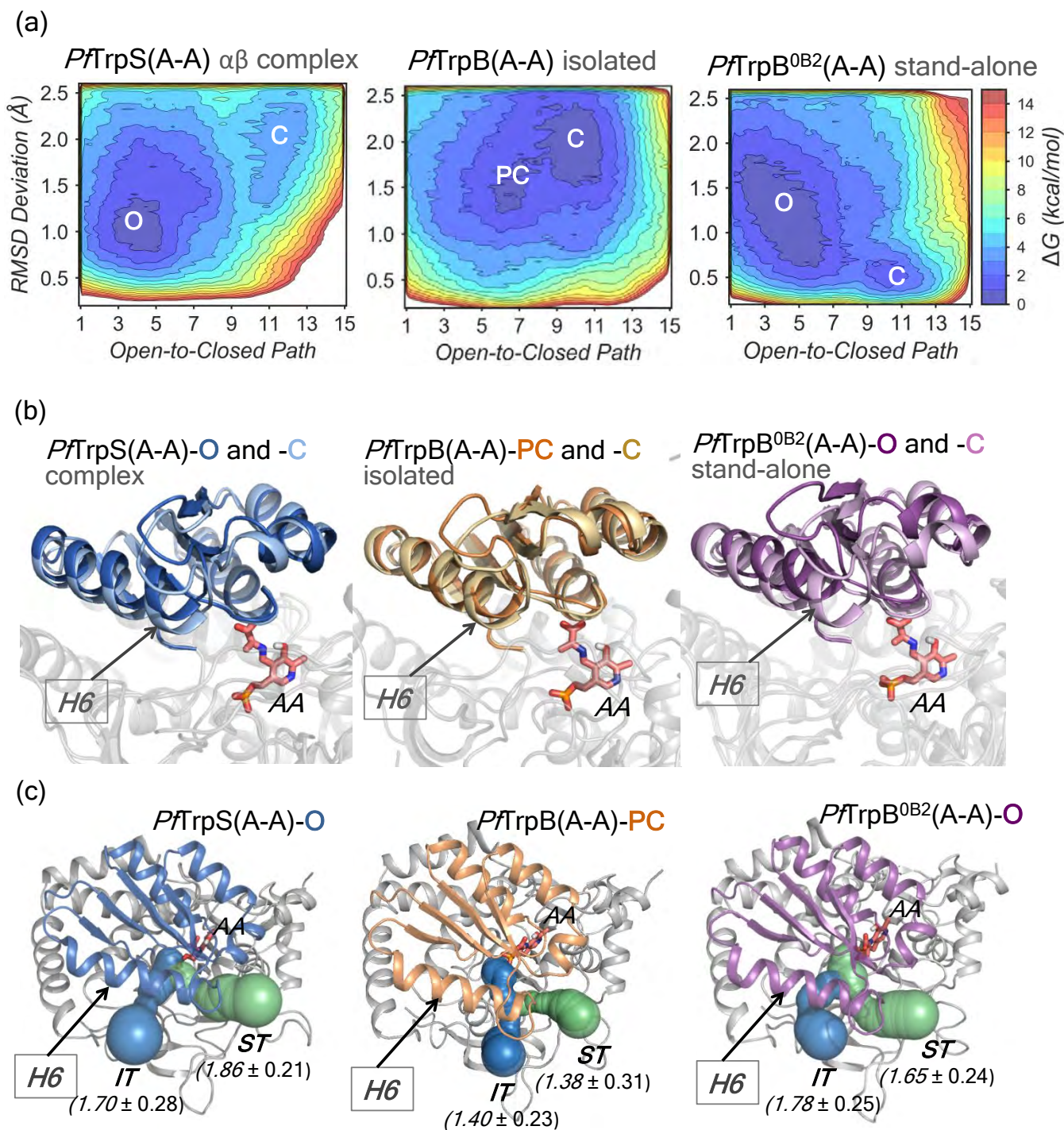


Figure S13. (a) Representation of the FELs associated with the COMM domain Open-to-Closed (O-to-C) transition for the *Pf*TrpS $\alpha\beta$ complex, *Pf*TrpB isolated and *Pf*TrpB^{OB2} stand-alone enzymes at A-A reaction intermediate. (b) Overlays of the *Pf*TrpS, *Pf*TrpB and *Pf*TrpB^{OB2} metastable open (O), partially closed (PC) and closed (C) conformations at A-A intermediate. (c) *Pf*TrpS, *Pf*TrpB and *Pf*TrpB^{OB2} metastable conformations of the open (O), and partially closed (PC) states at A-A reaction intermediate, together with the internal (IT, in blue) and the secondary (ST) tunnels computed with CAVER 3.0.¹⁷ The averaged bottleneck radii (in Å) are also shown.

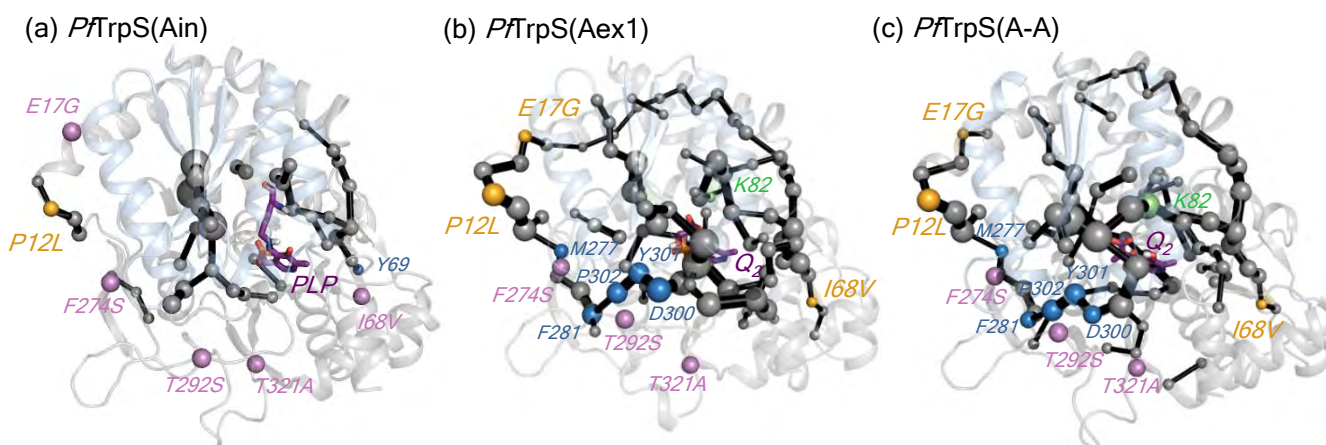


Figure S14. Identification of the amino-acids that contribute to the Open-to-Closed (**O**-to-**C**) conformational exchange in *PfTrpS* at Ain (a), Aex1 (b) and A-A (c) intermediates through Shortest Path Map (SPM) analysis.¹⁹ The size of the spheres and black edges are indicative of the importance of the position for the *PfTrpS* conformational dynamics. Positions mutated via DE are marked in orange (if they are included in the SPM), or in pink (if they directly interact with SPM residues). SPM residues that interacted with the DE positions are marked with blue spheres.

Note that in the absence of substrate (i.e. TrpS-Ain), SPM shows a small set of correlated motions. However, after substrate binding (i.e. Aex1, A-A and Q_2 intermediates) an increase in the number of correlated pathways connecting distal regions with the COMM domain is observed, which are conserved among the different reaction intermediates. In particular, one of the most conserved pathways found corresponds to the P12L and E17G distal regions that are located close the TrpA-TrpB interface, as well as the pathway that includes I68V. This analysis indicates that the COMM domain conformational dynamics is modulated by these different pathways, thus providing mutation points for achieving stand-alone function.

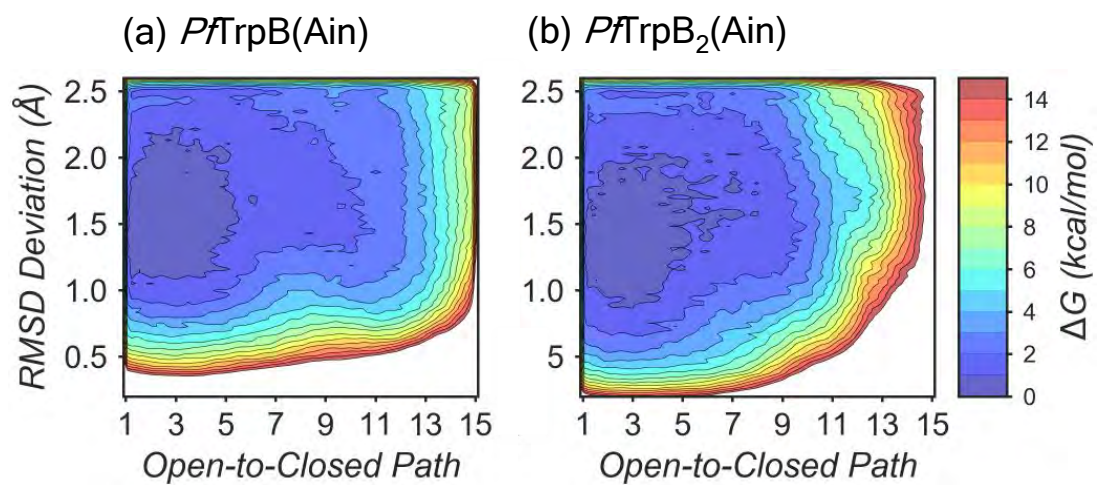


Figure S15. Representation of the FELs associated with the COMM domain Open-to-Closed (O-to-C) transition for the *PfTrpB* (a), *PfTrpB*₂ complex (b) enzymes at Ain reaction intermediate.

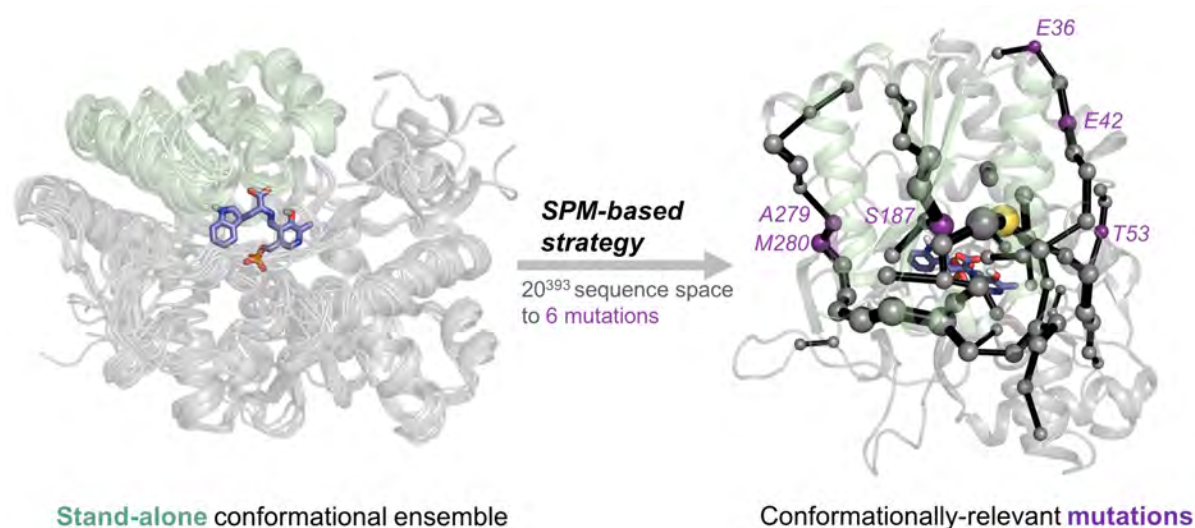
References :

- (1) Richter, F.; Leaver-Fay, A.; Khare, S. D.; Bjelic, S.; Baker, D., De Novo Enzyme Design Using Rosetta3. *PLOS ONE* **2011**, *6* (5), e19230.
- (2) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A., *AMBER 16, University of California, San Francisco, 2016*.
- (3) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004**, *25* (9), 1157-74.
- (4) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **1993**, *97* (40), 10269-10280.
- (5) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79* (2), 926-935.
- (6) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65* (3), 712-25.
- (7) Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, *98* (12), 10089-10092.
- (8) Laio, A.; Gervasio, F. L., Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **2008**, *71* (12), 22.
- (9) Branduardi, D.; Bussi, G.; Parrinello, M., Metadynamics with Adaptive Gaussians. *J Chem Theory Comput* **2012**, *8* (7), 2247-54.
- (10) Barducci, A.; Bonomi, M.; Parrinello, M., Metadynamics. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2011**, *1* (5), 826-843.
- (11) Alessandro, L.; Francesco, L. G., Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* **2008**, *71* (12), 126601.
- (12) Barducci, A.; Bonomi, M.; Parrinello, M., Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1* (5), 826-843.
- (13) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G., PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **2014**, *185* (2), 604-613.
- (14) Abraham, M. H., B.; Van del Spoel, D.; Lindahl, E., *GROMACS 5.1.2, University of Groningen, 2016*.
- (15) Barducci, A.; Bussi, G.; Parrinello, M., Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys Rev Lett* **2008**, *100* (2), 020603.
- (16) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M., Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J Phys Chem B* **2006**, *110* (8), 3533-9.
- (17) Chovancova, E.; Pavelka, A.; Benes, P.; Strnad, O.; Brezovsky, J.; Kozlikova, B.; Gora, A.; Sustr, V.; Klvana, M.; Medek, P.; Biedermannova, L.; Sochor, J.; Damborsky, J., CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol* **2012**, *8* (10), e1002708.

- (18) Csardi, G.; Nepusz, T., The igraph software package for complex network research. *InterJournal, Complex Systems* **2006**, 1695 (5), 1-9.
- (19) Romero-Rivera, A.; Garcia-Borràs, M.; Osuna, S., Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catal.* **2017**, 7 (12), 8524-8532.
- (20) Hioki, Y.; Ogasahara, K.; Lee, S. J.; Ma, J.; Ishida, M.; Yamagata, Y.; Matsuura, Y.; Ota, M.; Ikeguchi, M.; Kuramitsu, S.; Yutani, K., The crystal structure of the tryptophan synthase beta subunit from the hyperthermophile *Pyrococcus furiosus*. Investigation of stabilization factors. *Eur J Biochem* **2004**, 271 (13), 2624-35.
- (21) Lee, S. J.; Ogasahara, K.; Ma, J. C.; Nishio, K.; Ishida, M.; Yamagata, Y.; Tsukihara, T.; Yutani, K., Conformational changes in the tryptophan synthase from a hyperthermophile upon alpha(2)beta(2) complex formation: Crystal structure of the complex. *Biochemistry* **2005**, 44 (34), 11417-11427.
- (22) Buller, A. R.; Brinkmann-Chen, S.; Romney, D. K.; Herger, M.; Murciano-Calles, J.; Arnold, F. H., Directed evolution of the tryptophan synthase beta-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, 112 (47), 14599-14604.
- (23) Herger, M.; van Roye, P.; Romney, D. K.; Brinkmann-Chen, S.; Buller, A. R.; Arnold, F. H., Synthesis of beta-Branched Tryptophan Analogues Using an Engineered Subunit of Tryptophan Synthase. *J. Am. Chem. Soc.* **2016**, 138 (27), 8388-8391.
- (24) Buller, A. R.; van Roye, P.; Cahn, J. K. B.; Scheele, R. A.; Herger, M.; Arnold, F. H., Directed Evolution Mimics Allosteric Activation by Stepwise Tuning of the Conformational Ensemble. *J. Am. Chem. Soc.* **2018**, 140 (23), 7256-7266.
- (25) Boville, C. E.; Scheele, R. A.; Koch, P.; Brinkmann-Chen, S.; Buller, A. R.; Arnold, F. H., Engineered Biosynthesis of beta-Alkyl Tryptophan Analogues. *Angew. Chem. Int. Ed.* **2018**, 57 (45), 14764-14768.
- (26) Buller, A. R.; van Roye, P.; Murciano-Calles, J.; Arnold, F. H., Tryptophan Synthase Uses an Atypical Mechanism To Achieve Substrate Specificity. *Biochemistry* **2016**, 55 (51), 7043-7046.
- (27) Niks, D.; Hilario, E.; Dierkers, A.; Ngo, H.; Borchardt, D.; Neubauer, T. J.; Fan, L.; Mueller, L. J.; Dunn, M. F., Allostery and Substrate Channeling in the Tryptophan Synthase Bienenzyme Complex: Evidence for Two Subunit Conformations and Four Quaternary States. *Biochemistry* **2013**, 52 (37), 6396-6411.
- (28) Ngo, H.; Kimmich, N.; Harris, R.; Niks, D.; Blumenstein, L.; Kulik, V.; Barends, T. R.; Schlichting, I.; Dunn, M. F., Allosteric regulation of substrate channeling in tryptophan synthase: Modulation of the L-Serine reaction in stage I of the ss-reaction by alpha-site ligands. *Biochemistry* **2007**, 46 (26), 7740-7753.
- (29) Barends, T. R. M.; Domratheva, T.; Kulik, V.; Blumenstein, L.; Niks, D.; Dunn, M. F.; Schlichting, I., Structure and mechanistic implications of a tryptophan synthase quinonoid intermediate. *ChemBioChem* **2008**, 9 (7), 1024-1028.
- (30) Lai, J. F.; Niks, D.; Wang, Y. C.; Domratheva, T.; Barends, T. R. M.; Schwarz, F.; Olsen, R. A.; Elliott, D. W.; Fatmi, M. Q.; Chang, C. E. A.; Schlichting, I.; Dunn, M. F.; Mueller, L. J., X-ray and NMR Crystallography in an Enzyme Active Site: The Indoline Quinonoid Intermediate in Tryptophan Synthase. *J. Am. Chem. Soc.* **2011**, 133 (1), 4-7.
- (31) Johnson, E. R.; Keinan, S.; Mori-Sanchez, P.; Contreras-Garcia, J.; Cohen, A. J.; Yang, W., Revealing noncovalent interactions. *J Am Chem Soc* **2010**, 132 (18), 6498-506.
- (32) Contreras-Garcia, J.; Johnson, E. R.; Keinan, S.; Chaudret, R.; Piquemal, J. P.; Beratan, D. N.; Yang, W., NCIPLOT: a program for plotting non-covalent interaction regions. *J. Chem. Theory. Comput.* **2011**, 7 (3), 625-632.



5.2 Rational prediction of distal activity-enhancing mutations in tryptophan synthase



Maria-Solano, M.A.*; Kinateder, T.; Iglesias-Fernández, J.; Sterner, R.*; Osuna, S.* Rational prediction of distal activity-enhancing mutations in tryptophan synthase, [to be submitted].

The work included in this chapter has been carried out in collaboration with an experimental group led by Reinhard Sterner. The free energy landscape calculations and their combined analysis with the Shortest Path Map tool for the rational design of stand-alone enzyme variants was performed by our group, while the reconstruction of the phylogenetic tree and the experimental validation of the designed variants by the Sterner group.

Abstract

Allostery is a central mechanism for the regulation of multi-enzyme complexes. The mechanistic basis that drives allosteric regulation is poorly understood, but harbors key information for enzyme engineering. In the present study, we focus on the tryptophan synthase complex that is composed of TrpA and TrpB subunits, which allosterically activate each other. Specifically, we develop a rational approach for identifying key amino acid residues of TrpB distal from the active site. In particular, we predict positions crucial for shifting the inefficient conformational ensemble of the isolated TrpB to a productive ensemble through intra-subunit allosteric effects. The experimental validation of the new conformationally-driven TrpB design demonstrates its superior stand-alone activity in the absence of TrpA, comparable to those enhancements obtained after multiple rounds of experimental laboratory evolution. Our work evidences that the current challenge of distal active site prediction for enhanced function in computational enzyme design can be ultimately addressed.

Rational prediction of distal activity-enhancing mutations in tryptophan synthase

Miguel A. Maria-Solano,^{[a]#} Thomas Kinateder,^{[b]#} Javier Iglesias-Fernández,^[a] Reinhard Sterner,^{[b]*} and Sílvia Osuna^{[a,c]*}

[a] CompBioLab group, Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, Girona, Spain

[b] Institute of Biophysics and Physical Biochemistry, Regensburg Center for Biochemistry, University of Regensburg, Universitätsstrasse 31, 93053 Regensburg, Germany

[c] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

#: These authors contributed equally to the work

ABSTRACT

Allostery is a central mechanism for the regulation of multi-enzyme complexes. The mechanistic basis that drives allosteric regulation is poorly understood, but harbors key information for enzyme engineering. In the present study, we focus on the tryptophan synthase complex that is composed of TrpA and TrpB subunits, which allosterically activate each other. Specifically, we develop a rational approach for identifying key amino acid residues of TrpB distal from the active site. In particular, we predict positions crucial for shifting the inefficient conformational ensemble of the isolated TrpB to a productive ensemble through intra-subunit allosteric effects. The experimental validation of the new conformationally-driven TrpB design demonstrates its superior stand-alone activity in the absence of TrpA, comparable to those enhancements obtained after multiple rounds of experimental laboratory evolution. Our work evidences that the current challenge of distal active site prediction for enhanced function in computational enzyme design can be ultimately addressed.

INTRODUCTION

Enzymes are some of the most sophisticated biomolecules that exist on Earth. They achieve impressive rate accelerations thanks to their highly preorganized active site pocket, while exhibiting remarkable conformational flexibility key for their function, regulation and evolution.¹⁻⁷ Enzymes are dynamic biological entities, being their catalytic activity directly related to their structure and the broad ensemble of conformations they sample in solution.⁴⁻⁶ This conformational equilibrium can be shifted, for example, by the binding of a ligand to a given site. This in turn influences the binding or the turnover of a substrate at the active site of the enzyme, a phenomenon that is called “allostery”.^{8, 9} Likewise, the introduction of an amino acid substitution in the protein sequence not only induces an evident structural change but also a redistribution of the conformational ensemble, which in turn can potentially impact catalytic activity.^{4, 6, 10, 11} Indeed, it has been proven that allosteric effects are not restricted to effector binding, but instead single point mutations or covalent attachment (e.g. phosphorylation), among others can induce similar responses.^{9, 12, 13}

Identifying mutations that modulate enzyme activity is the primary goal of enzyme engineering. One approach to enzyme engineering is Directed Evolution (DE), which has been applied to a myriad of enzyme systems successfully identifying active site and distal mutations, providing access to impressive tailor-made enzyme variants at the expense of large and expensive screening efforts.¹⁴⁻¹⁷

Rational design emerged as an attractive alternative to decrease the screening efforts to a reduced number of promising enzyme variants based on prior structural knowledge and computational approaches.¹⁸⁻²¹ Given the sophisticated nature of enzyme catalysis, multiple computational strategies and protocols have been developed in recent years for computational enzyme design.²⁰ The evaluation of the conformational landscape of enzymes along distinct natural and DE evolutionary pathways has evidenced that the introduced mutations progressively tune the conformational ensemble, stabilizing key conformational states for the novel function.^{4, 6, 10, 20} Of note is that the mutations introduced with DE are often located distal from the active site pocket, which given the vast sequence space are computationally challenging to predict.^{20, 22, 23} In addition to that, the computational prediction of which remote mutations can induce the desired population shift to favor the key conformational ensemble for novel functionality is an extremely difficult task.²⁰ Our group has recently shown that active site and distal positions targeted by DE can be computationally identified through the coupling of MD simulations with cross correlation methods such as the Shortest Path Map (SPM).^{20, 24} SPM has been applied for identifying DE mutations in the retro-aldolase, monoamine oxidase and tryptophan synthase enzymes suggesting its potential application for smart library construction for enzyme design.^{20, 24}

Tryptophan synthase (TrpS) is an excellent model system for studying allosteric properties. TrpS is a heterodimeric enzyme complex formed by α (TrpA) and β (TrpB) subunits in an $\alpha\beta\alpha$ arrangement. The functional unit is formed by a TrpA and an associated TrpB subunit (Fig. 1a).^{25, 26} TrpA catalyzes the retro-aldol cleavage of indole-3-glycerol phosphate (IGP) producing glyceraldehyde-3-phosphate (G3P) and indole, which diffuses along an internal tunnel towards the TrpB active site.²⁷ TrpB is a pyridoxal phosphate (PLP) cofactor dependent enzyme that catalyzes the production of L-Tryptophan by condensation of indole and L-serine in a multistep reaction mechanism, which mainly comprises: (1) formation of a Schiff base intermediate (Ain) at the resting state by covalent attachment of PLP cofactor to the catalytic lysine, (2) transamination with L-Ser, (3) indole coupling, and (4) formation of several quinonoid intermediates (Q) to finally release L-Trp. This complex multi-step mechanism involves multiple proton donor/abstraction steps assisted by the catalytic lysine (Supplementary Scheme 1).²⁸ Of relevance is the tight allosteric coupling between TrpA and TrpB along the catalytic itinerary.^{29, 30} TrpA and TrpB catalyze different reactions that are synchronized (i.e. TrpA tunes the TrpB conformational ensemble and vice versa). This fine tuning of the conformational ensemble involves open-to-closed transitions of the rigid COMM domain that forms a lid covering the TrpB active site (Fig. 1b) and an active site loop of TrpA, as shown by X-ray and computational data.^{26, 31, 32} Given the tight allosteric communication exerted between subunits, both TrpA and TrpB are much less efficient when isolated, which hampers TrpB industrial application for non-canonical amino-acids production.³³⁻³⁸ Arnold and coworkers addressed this limitation by applying DE to optimize activity of TrpB from the TrpS of *Pyrococcus furiosus* for stand-alone function (i.e. recovery of the catalytic activity in the absence of the allosteric protein partner TrpA).^{33, 34} Interestingly, the most evolved variant (*pfTrpB*^{0B2}) was even more efficient than the original *pfTrpS* complex (2.9-fold increase in k_{cat}), and contained 5 out of the 6 mutations located distal from the active site. This manifests that the recovery of activity exerted by the distal mutations is induced through allosteric effects.^{33, 34} Intrigued by the allosteric regulation induced

by distal mutations, we explored the conformational energy landscape of the *pfTrpS* enzyme complex, the *pfTrpB* isolated enzyme and the stand-alone *pfTrpB*^{OB2} evolved variant.³¹ Free energy calculations revealed that the DE mutations in *pfTrpB*^{OB2} recovered the allosterically driven conformational ensemble of the *pfTrpS* complex, allowing the exploration of open, partially closed and closed conformations of the COMM domain, which is required for the multi-step catalytic pathway. The *pfTrpB* stand-alone activity was thus achieved through the recovery of the conformational ensemble present in the *pfTrpS* complex. In fact, the allosterically driven conformational ensemble was not only recovered but also improved, as a higher stability of catalytically productive closed states was found in the case of *pfTrpB*^{OB2}. This explained the *pfTrpB*^{OB2} superior activity with respect to the *pfTrpS* complex. In contrast, isolated *pfTrpB* showed a restricted COMM domain conformational heterogeneity and catalytically unproductive closed states. Careful analysis of the *pfTrpS* conformational ensemble through SPM correlation-based tools elucidated the enzyme pathways most contributing to the TrpS conformational dynamics, which interestingly included some important DE positions.^{20, 31} This suggests that the identified positions with SPM can potentially alter the enzyme conformational dynamics, and thus its stand-alone activity. However, multiple positions are identified and there is a lack of information on which specific amino-acid substitution should be introduced for achieving an efficient conformational ensemble for stand-alone function.

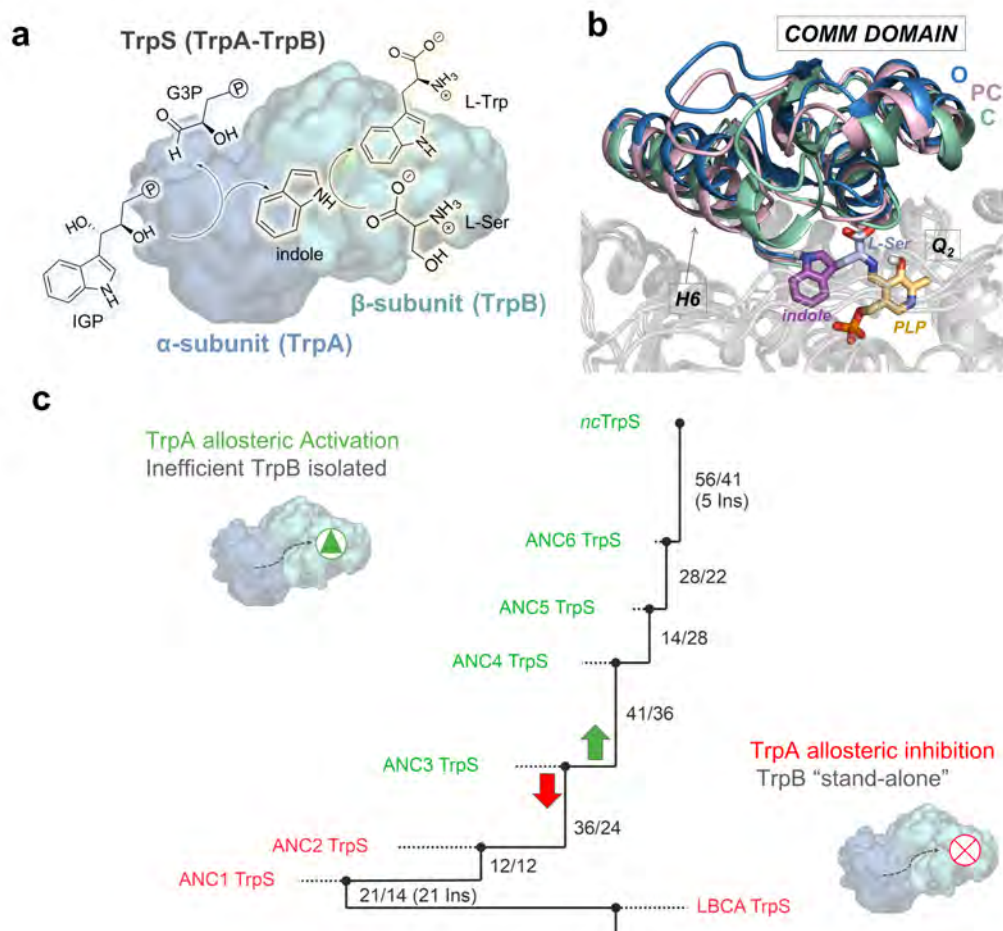


Fig. 1 | Overview of Tryptophan Synthase (TrpS) enzyme. **a**, The functional unit of TrpS consists of a heterodimer, which is formed by TrpA (blue) and TrpB (green). TrpA catalyzes the cleavage of indole-3-glycerol phosphate (IGP) to glyceraldehyde-3-phosphate (G3P) and indole, which in TrpB reacts with activated L-Ser in a multistep mechanism to yield L-Trp (see Supplementary Scheme 1). **b**, Overlay of *pf*TrpS metastable conformations from previous computational exploration showing the transition of the COMM domain (residues 97-184) from an open (blue, O), to a partially closed (pink, PC) to a closed conformation (green, C). Highlighted are the α -helix H6 of the COMM domain (residues 174-164) and the reaction intermediate Q₂ in the active site, which is colored as a function of its origin molecule (PLP cofactor in orange, L-Ser in blue and indole in purple).³¹ **c**, The phylogenetic tree shows the path from the last bacterial common ancestor (LBCA) TrpS over six intermediate nodes (ANC1 TrpS to ANC6 TrpS) to the extant *Neptuniibacter caesariensis* TrpS.³⁹ Numbers next to each edge indicate the number of mutations accumulated in TrpA and TrpB with respect to the previous node. While LBCA-TrpB gets deactivated by TrpA and exhibits stand-alone function, the allosteric effect of TrpA is reverted along the phylogenetic tree with a switch between ANC2 TrpB and ANC3 TrpB to an allosteric activation, as observed in extant *nc*TrpB.

An orthogonal *in silico* method to analyze functional transitions in enzyme evolution is ancestral sequence reconstruction (ASR).⁴⁰⁻⁴² In a previous work, we reconstructed the TrpS phylogenetic tree and identified a shift in the allosteric modulation exerted by TrpA on TrpB activity.^{39, 43} The analysis of the steady state kinetic parameters of the last bacterial common ancestor (LBCA) revealed high stand-alone activity of LBCA-TrpB and its allosteric inhibition in the presence of TrpA. Along the phylogenetic tree, this inhibition was gradually inverted towards allosteric activation existing in modern TrpB (Fig. 1c).

This inversion of the allosteric effect exerted by TrpA on TrpB between ANC2 and ANC3 provides a perfect starting point for an SPM-based design. Specifically, we wanted to identify residues within the allosteric network of TrpB that are able to rescue the missing allosteric activation from TrpA and predict mutations that convey stand-alone function in the context of the inefficient ANC3 TrpB. To this end, we intended to explore the conformational ensemble of the stand-alone LBCA TrpB enzyme system, and identify key positions by means of our developed SPM correlation-based tool. Sequence comparison of the identified positions along the phylogenetic tree further reduces the number of potential mutations and provides the specific amino-acid substitutions for stand-alone function. This approach decreases the experimental screening to one single mutant and includes the rational prediction of both active site and distal mutations. Our study presents a computational enzyme design approach that is not restricted to active site mutations and demonstrates that the challenge to rationally predict distal mutations can be ultimately addressed by exploring the conformational energy landscape of enzymes in combination with cross correlation and bioinformatic tools.

RESULTS:

Reconstruction of Ancestral TrpS conformational ensembles. As shown in previous studies, natural evolution has altered the need of TrpS to be allosterically regulated.³⁹ As opposed to modern TrpB, the ancestral LBCA TrpB was found to operate less efficiently (in terms of k_{cat}) in the presence of TrpA.⁴³ The allosteric inhibition imparted by TrpA suggests that the ancestral TrpB in complex is less efficient in accessing the catalytically productive conformational states required for enhanced activity.³¹ Interestingly, LBCA TrpB affinity towards L-Ser substrate was enhanced in the heterocomplex form (Supplementary Table 1). To provide the molecular basis for stand-alone activity and higher affinity towards L-Ser, we decided to computationally reconstruct the free energy landscape (FEL) of LBCA TrpB in the presence (i.e. heterocomplex TrpS) and absence of TrpA (Fig. 2). We employed metadynamics simulations to reconstruct the FEL associated with the open-to-closed transition of the COMM domain (see Supplementary Fig. 1) at the resting state (i.e. E(Ain)) and at the Q₂ intermediate (i.e. quinonoid intermediate formed after indole coupling, see Supplementary Scheme 1). The reconstructed FEL of the LBCA TrpB(Ain) in the absence of TrpA, indicates that TrpB(Ain) mostly visits partially closed (PC) conformational states of the COMM domain (Fig. 2a). This is altered in the presence of TrpA, which clearly induces a shift in the FEL stabilizing open (O) states with similar deviations from the reference path (Fig. 2a and 2b, on the left). At the resting state, closed (C) states are inaccessible for both systems. The analysis of the access tunnels to the active site for L-Ser binding through CAVER calculations (see Fig. 2c and Supplementary Fig. 2) indicates that the PC conformational ensemble of the isolated LBCA TrpB has a substantially narrower tunnel bottleneck than the accessible O states of the complex. This finding indicates that the O conformational ensemble improves L-Ser accessibility to the active site, thus explaining the enhanced K_M^{L-Ser} values displayed by the LBCA TrpS complex.

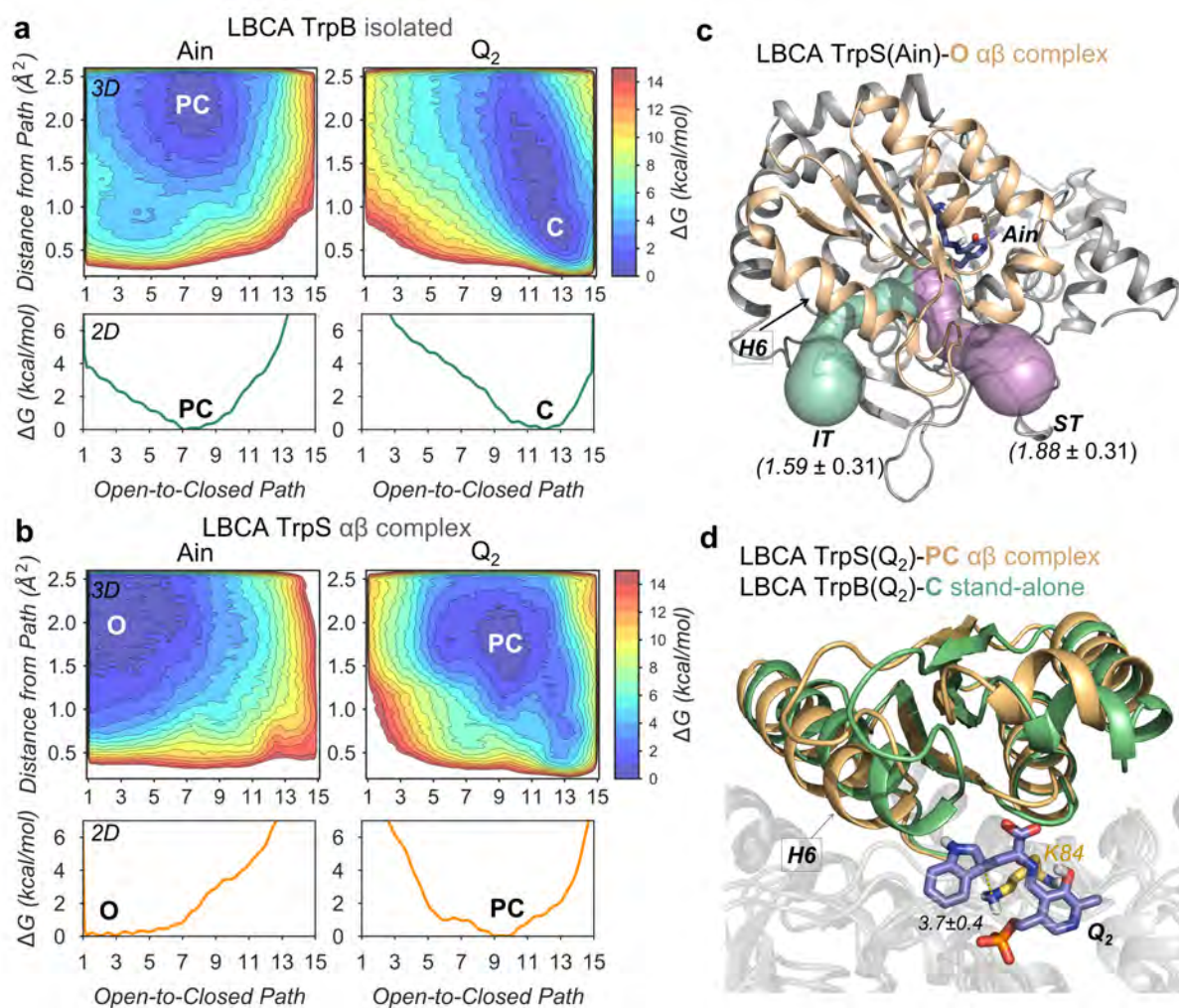


Fig. 2 | Computational exploration of the LBCA conformational ensemble. Free energy landscape (FEL) associated with the COMM domain open-to-closed (O-to-C) conformational transition of the LBCA TrpB (**a**) and LBCA TrpS (**b**) at Ain and Q_2 reaction intermediates. The x-axis corresponds to the progression along the reference O-to-C path generated from X-Ray data, while the y-axis to the mean square deviation (MSD) distance from the reference path **c**, Tunnels access of LBCA TrpS-O state at Ain reaction intermediate for the L-Ser substrate, computed with CAVER 3.0. The averaged bottleneck radii (in \AA) for the internal TrpA-TrpB tunnel (IT, green) and the secondary tunnel (ST, violet) found are also shown. **d**, Overlays of the metastable conformations of the partially closed (PC) state of LBCA-TrpS (orange) and the closed (C) state (green) of LBCA-TrpB at Q_2 reaction intermediate. The catalytic proton transfer distance (in \AA) between the K84 (yellow) residue and the Q_2 reaction intermediate (slate) is also represented.

More interesting is the fact that TrpA was found to inhibit the TrpB catalytic efficiency, as isolated LBCA TrpB displays a *ca.* 8.4-fold k_{cat} higher value. As we show in our previous study,³¹ the catalytic activity of TrpS can be estimated by evaluating its ability to visit catalytically competent C states of the COMM domain. The catalytically-relevant closed conformational ensemble displays an efficient active site preorganization by means of optimized non-covalent interaction networks and short catalytic distances between the Q_2 intermediate and the conserved catalytic K84 that acts as proton acceptor. In particular, the H6 COMM domain helix was found to play an important role in the closure to form non-covalent interactions with the indole moiety of Q_2 . In the present work, the reconstructed FEL associated to the COMM domain open-to-closed (O-to-C) transition for LBCA TrpB (Fig. **2a,d**) indicates that at the

Q₂ intermediate, the catalytically productive C conformational ensemble is indeed accessible for efficient catalysis. The structural characterization of the visited C conformational states shows LBCA TrpB adopts catalytically productive COMM domain closure with appropriate K84-Q₂ proton transfer distances (Fig. 2d and Supplementary Fig. 3 and 4), as discussed above. This evidences that LBCA TrpB has stand-alone properties derived from the exploration of stable catalytically competent C conformations in the absence of TrpA. On the contrary, LBCA TrpA alters the conformational landscape of TrpB as it induces a shift towards PC conformations hampering the ability of the COMM domain to complete the O-to-C transition for achieving catalytically productive C states (Fig. 2b). As expected, PC conformations of LBCA TrpB in the presence of TrpA do not exhibit a competent closure of the COMM domain, in particular this is notorious for the H6 region. Besides, the K84-Q₂ proton transfer distances are larger (Fig 2d. and Supplementary Fig. 3). In summary, our results indicate that the destabilization of the competent C LBCA TrpB ensemble is the main responsible of the allosteric inhibition exerted by the LBCA TrpA protein partner. It is worth mentioning that we estimated a similar effect (i.e. destabilization of the competent C ensemble) for the allosteric inhibition exerted by *pf*TrpA on the laboratory-evolved stand-alone *pf*TrpB^{OB2}. Another interesting aspect of LBCA conformational dynamics is its limited conformational heterogeneity (i.e. a narrow set of states are sampled), especially if compared with the previously studied allosteric *pf*TrpS complex and the laboratory-evolved stand-alone *pf*TrpB^{OB2} catalyst. A high degree of conformational heterogeneity was observed for the latter cases, which explored the complete O-to-C transition at Q₂ intermediate. The lack of O states of the COMM domain at the Q₂ intermediate for LBCA-TrpB suggests a more rigid COMM as the reaction evolves, and an infrequent transition towards O state, thus suggesting that product release might be rate limiting.

Computational prediction of distal active site mutations for stand-alone function. The mutations introduced along an evolutionary pathway progressively tune the conformational ensemble of enzymes towards novel function.^{4, 6, 10, 20} In this context, distal active site mutations have been shown to play a crucial role in natural and laboratory evolvability.^{12, 22} Their prediction considering the vast protein sequence space that yields a targeted function is, however, an extremely challenging task in computational enzyme design.²⁰ We have recently reported that molecular dynamics coupled to correlation-based tools are promising methodologies for the identification of both active site and distal positions targeted in non-rational laboratory evolution experiments.^{20, 24} In particular, we successfully developed and applied the Shortest Path Map (SPM) method for identifying the enzyme pathways that most contribute to the conformational dynamics of the *pf*TrpS enzyme. Of relevance is that the identified positions contained or make persistent non-covalent interactions with residues targeted in the laboratory evolution of the *pf*TrpS enzyme for stand-alone function.³¹ SPM identifies important positions for the enzyme conformational dynamics, thus reducing the potential number of mutational hotspots.

Inspired by the previous work on the TS ancestral reconstruction, we focused our computational design on the ancestral ANC3 TrpB scaffold.^{39, 43} This enzyme corresponds to the third node of the phylogenetic tree and exhibits reversion of allosteric inhibition towards activation along the evolution pathway (see Fig. 1c). In other words, ANC3 TrpB is the first enzyme that is allosterically dependent on

TrpA, thus being highly inefficient as stand-alone catalyst (Supplementary Table 2). The absence of TrpA decreases ANC3 TrpB activity 30.2-fold in terms of k_{cat} , suggesting a reduced conformational O-to-C ensemble. Given the success of SPM in identifying key positions for the enzyme conformational dynamics, we decided to apply our computational methodology to rationally engineer an ANC3 TrpB catalyst towards stand-alone activity. Our initial reference protein was LBCA TrpB, as it exhibits stand-alone properties thanks to its ability to adopt stable and efficient closed states of the COMM domain. The SPM analysis of the LBCA TrpB SPM identified 74 possible hotspots that potentially regulate the enzyme conformational dynamics (74 out of 413 residues, i.e. 18% of the total enzyme). This number is too large for an efficient rational design of ANC3 TrpB, as it is unclear, which positions should be targeted and which substitutions should be introduced to establish stand-alone function. We solved this problem by analyzing the sequence conservation between LBCA TrpB and the targeted ANC3 TrpB system for the 74 SPM positions (see the workflow followed in Fig. 3). Comparing the residues at each of the 74 SPM positions reduced the number of sites to 6 and specified the nature of the mutation to the amino acid found in LBCA TrpB. Interestingly, 5 out of 6 positions were located far away from the active site.

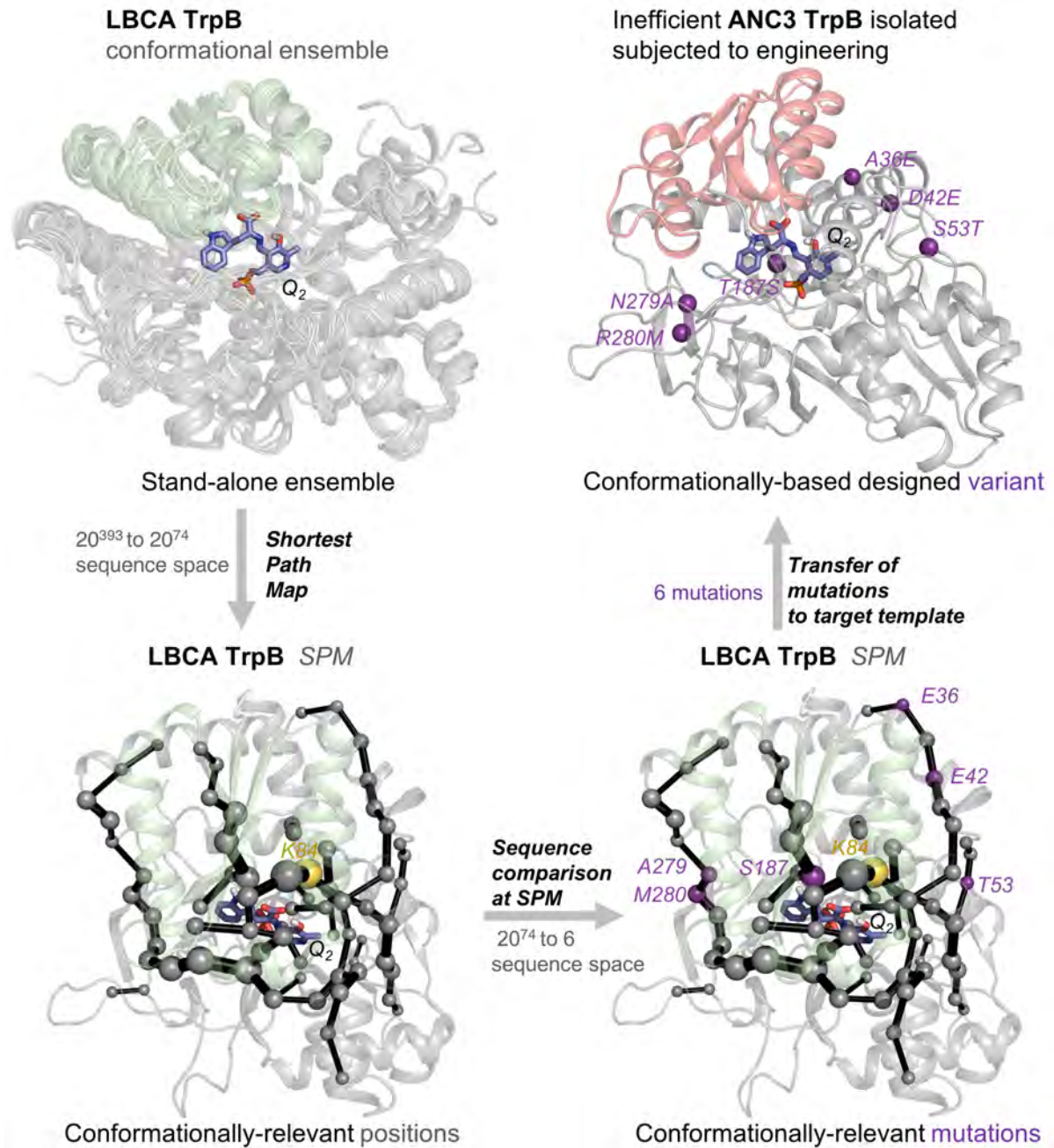


Fig. 3 | SPM-based computational workflow for the rational design of SPM6 TrpB enzyme variant. By analyzing the conformational ensemble of the stand-alone LBCA TrpB with high catalytic activity (upper left ensemble) through the SPM, we identified positions (grey spheres, lower left structure) within allosteric pathways (black edges) in the enzyme that most contribute to the LBCA TrpB conformational dynamics in the Q2 intermediate. Thereby the size of each edge and node corresponds to the relevance for conformational dynamics; catalytic K84 is highlighted in yellow. Excluding residues that do not participate in an allosteric pathway reduces the sequence space from 20^{393} to 20^{74} possible activity enhancing substitutions. Sequence comparison at the SPM positions between stand-alone LBCA TrpB and inefficient ANC3 TrpB reduces the sequence space to 6 mutations with respect to LBCA TrpB (lower right structure, purple residues), that were introduced into ANC3 TrpB (upper right structure, purple residues) and tested *in-vitro*.

Rational SPM-based ANC3 TrpB designs for stand-alone function. The application of the SPM method coupled to sequence comparison between two variants exhibiting rather high (LBCA TrpB) or low (ANC3 TrpB) stand-alone function reduced the SPM library to only 6 specific mutations in ANC3 TrpB: A56E, D62E, S73T, T207S, N299A and R300M. This ANC3 variant was termed SPM6 TrpB. Interestingly, none of the mutations are located at the COMM domain, 5 out of 6 mutations are located far away from the active site (ca. 18-29 Å), among which N299A and R300M are near the TrpA-TrpB protein interface and only S73T is located at the active site pocket (Fig. 3 and Supplementary Fig. 5). The computational screening of ANC3 TrpB, the ANC3 TrpS and the SPM6 TrpB enzyme variant by means of conventional molecular dynamics simulations suggested that both SPM6 TrpB and ANC3 TrpS are able to retain the closed conformation of the COMM domain. In contrast, isolated ANC3 TrpB explores additional non-productive conformations (Supplementary Fig. 6). This fast screening computational protocol suggests a rather low stability of the C state of the COMM domain in isolated ANC3 TrpB, which explains its low stand-alone catalytic activity (Fig. 4a). These computational insights encouraged us to experimentally test the SPM6 enzyme variant. As shown in Fig. 4 and Supplementary Table 3, SPM6 TrpB successfully enhances the catalytic activity with respect to ANC3 TrpB *by almost* one order of magnitude (7-fold increase in k_{cat}). The catalytic efficiencies for both, L-Ser and indole are also improved. It is worth emphasizing that a similar fold increase in stand-alone catalytic activity was achieved in *p*TrpB by means of multiple rounds of laboratory evolution.³³ Our SPM-based computational approach therefore provides the same order of improvement in stand-alone activity but by only testing one single rationally designed variant. Another interesting aspect to evaluate is whether the SPM6 mutations have an impact in the allosteric modulation exerted by TrpA. The catalytic activity of the ancestral ANC3 TrpB increases 30.2-fold in terms of k_{cat} thanks to the TrpA-triggered allosteric activation. Unexpectedly, the introduction of SPM6 mutations to ANC3 TrpB confer increased stand-alone activity while still retaining some TrpA allosteric activation (the activity of SPM6 TrpB is enhanced 5.5-fold in the presence of TrpA). This indicates that the SPM6 distal mutations tune the O-to-C conformational ensemble of SPM6 TrpB through long-range intra-subunit allosteric effects but these changes in the conformational landscape do not prevent TrpA allosteric activation. In fact, the combination of both inter-subunit and intra-subunit allosteric effects yields SPM6 TrpS complex displaying even higher catalytic activity than the ancestral ANC3 TrpS complex (i.e. 1.3-fold increase, Fig. 4a).

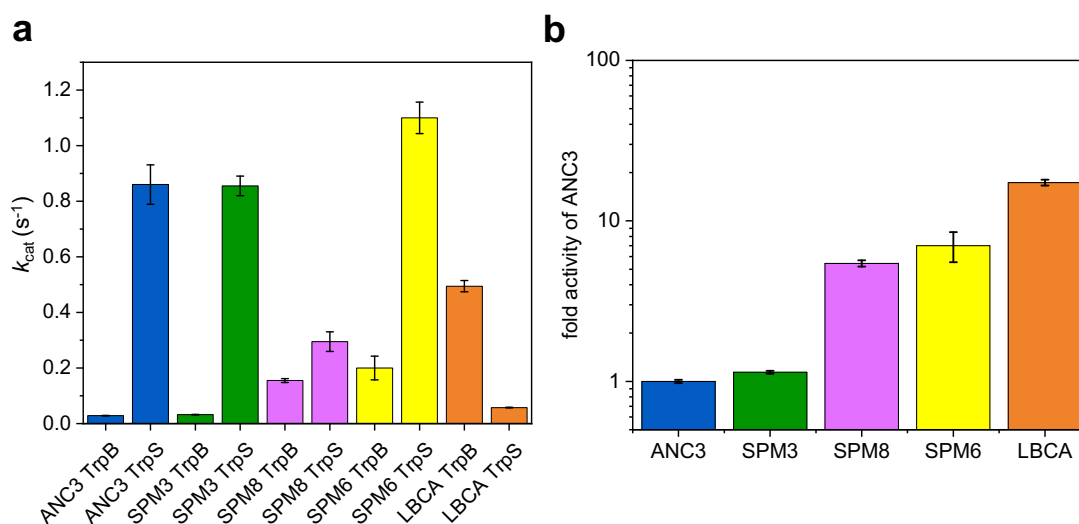


Fig. 4 | Illustration of the TrpB kinetic characterization. **a**, Activity changes of ANC3, SPM3, SPM8, SPM6 and LBCA isolated TrpB enzymes upon complexation with their corresponding TrpA protein partners, in terms of the average values for k_{cat} . **b**, TrpB fold activity at logarithmic scale of ANC3 respect to SPM3, SPM8, SPM6 and LBCA. The new TrpB designs SPM6 and SPM8 are 7 and 5.4-fold more active than ANC3 TrpB, while the reference stand-alone LBCA TrpB 17.4-fold. Errors bars indicate the standard deviation observed in two separate experiments.

In order to further test the SPM predictive power and the robustness of the strategy followed so far, we additionally targeted two other SPM based approaches. In the first one, we followed the same workflow as for the SPM6 design but used instead of LBCA TrpB the isolated ANC2 TrpB as stand-alone reference protein for the SPM pathway analysis. (Fig. 1c). After identifying the shortest path map in ANC2 TrpB and subsequent sequence comparison between ANC2 TrpB and ANC3 TrpB we identified 3 SPM positions and predicted 3 mutations in ANC3 TrpB: S73T, N299S and R300M. The corresponding variant was termed SPM3. This reduced number of non-conserved SPM positions makes sense since ANC2 and ANC3 are closer in the phylogenetic tree than LBCA and ANC3.

In the second approach, we conducted a SPM analysis on the ANC3 TrpS complex. The rationale behind taking this complex as a reference was that, while isolated ANC3 TrpB is poorly active and its allosteric communication is likely truncated, complexation with TrpA leads to high activity and likely a restored allosteric network. After identifying allosterically relevant SPM positions within ANC3 TrpS, we again compared the ANC3 TrpB sequence to LBCA TrpB in order to predict mutations that lead to a stand-alone catalyst. Following this protocol, we identified two extra positions as compared to SPM6 (SPM8): R53N and M187I, where R53N is far away the active site and M187I is located at the H6 helix of the COMM domain. It should be noted that the 3 positions of SPM3 and 6 out of 8 of SPM8 were previously identified in SPM6. Following the same computational MD-based screening protocol SPM3 and SPM8 TrpB variants were analyzed, which suggested a rather high stability of the C state of the COMM domain for enhanced activity (Supplementary Fig. 6). The experimental validation of the computational predictions for both SPM3 and SPM8 variants revealed enhanced catalytic activities of ANC3 TrpB when isolated. SPM8 TrpB exhibited a similar activity enhancement (5.4-fold in terms of k_{cat}) as SPM6 TrpB, whereas a quite modest enhancement was obtained for SPM3 TrpB in line with the

reduced number of mutations (1.1-fold). Regarding the inter allosteric effects exerted by TrpA, SPM3 and SPM8 variants also showed TrpA allosteric activation. In particular, SPM3 showed a similar degree of the k_{cat} increase in complex as ANC3 (26.3-fold), while SPM6 (5.5-fold) and SPM8 (1.9-fold) enzyme variants present a reduced predisposition to the TrpA allosteric activation (Fig. 4).

DISCUSSION

Allosteric regulation is a central biological process focused on the functional connection between distinct sites on either a single biological entity or among complex multimeric structures.^{9, 12, 44} This regulation of enzymatic function is not limited to effector or protein partner binding, as similar effects have been observed by covalent attachment or by introducing mutations located at distal positions of the enzyme active site.^{9, 12} The elucidation of the underlying mechanism and forces that drive allosteric regulation has the enormous potential of identifying key positions for regulating enzymatic function, which could be exploited in enzyme design.²⁰ The present study indeed demonstrates that distal active site positions, regulating the allosterically-driven conformational ensemble and thus the enzyme activity, can be successfully identified by means of correlation-based tools and sequence comparison. Given the vast sequence and conformational space, the rational prediction of mutations, especially those located at remote positions from the active site impacting enzymatic function is an extremely difficult task in the computational enzyme design field. Apart from that, the identification of the amino acid substitutions that optimize the enzyme conformational ensemble for a targeted enzyme function is extremely challenging. Our study focuses on the engineering of stand-alone function taking advantage of the substantial allosteric contributions that distal mutations were exerting on the laboratory-evolved variants.^{31, 33, 34} The exploration of the free energy landscape of the ancestrally reconstructed LBCA TrpS in complex and as stand-alone catalyst (LBCA TrpB), together with our previous findings³¹ on the wild-type *pf*TrpS complex, isolated *pf*TrpB and laboratory-evolved *pf*TrpB^{OB2} has elucidated the conformational ensemble that a stand-alone catalyst has to display for being efficient. This information is pivotal for fine-tuning the conformational ensemble and progressing towards the targeted enzyme design goal. We find that LBCA TrpB naturally adopts a stable catalytically productive COMM domain closure, which is hampered by the presence of the LBCA TrpA protein partner. LBCA TrpA therefore induces an allosteric inhibition of LBCA TrpB activity, which contrasts with the TrpA allosteric activation usually found in modern TrpB enzymes. In this study we exploit the intrinsic ability of LBCA TrpB to efficiently stabilize catalytically competent COMM domain closed conformations when isolated (i.e. crucial for stand-alone properties), and develop a novel computational enzyme design approach for achieving stand-alone function. In particular, we apply our SPM method to identify the enzyme pathways and positions that most contribute to the LBCA TrpB conformational dynamics. We hypothesized that these conformationally-relevant SPM positions could be potential hotspots for tuning the conformational ensemble of TrpA-dependent TrpB enzymes. The reconstruction of the phylogenetic tree from LBCA TrpS to the modern *nc*TrpS provided an intermediate variant ANC3 TrpB, which exhibits a high allosteric activation from ANC3 TrpA (i.e. ANC3 TrpB is highly inefficient when isolated). The application of SPM into LBCA TrpB reduced the sequence space from 393 to 74 hotspots, suggesting

that ca. 18% of the positions play a conformationally-relevant role. However, this still leads to a massive amount of enzyme variants to screen. Interestingly, the analysis of sequence conservation at the identified SPM positions between LBCA and ANC3 TrpB templates reduced this large number to only 6 positions. This approach assumes that the transfer of the non-conserved conformationally-relevant SPM mutations from the LBCA to the targeted ANC3 TrpB template will alter the enzyme conformational dynamics and induce the stabilization of the catalytically relevant closed state of the COMM domain. It is worth mentioning that among these 6 mutations 5 are distal from the active site and none is included in the COMM domain.

The fast-computational screening of the rationally designed enzyme including these 6 mutations indicated that SPM6 better stabilizes the closed conformational ensemble than the parent ANC3. Indeed, the experimental evaluation of SPM6 indicated the introduced mutations boosted the stand-alone catalytic activity of the inefficient isolated ANC3 TrpB enzyme near one order of magnitude. This enhancement by only testing a single variant is comparable to that observed for the laboratory evolved *pfTrpB*^{OB2} after three rounds of DE, that involved the screening of ca. 3,080 variants.³³ The observed enhancement of ANC3 TrpB stand-alone activity still does not completely recover the activity displayed by the ANC3 TrpS complex. The new SPM6 designed variant enhances the low initial 3% activity displayed by ANC3 TrpB up to ca. 23%. It should be also mentioned that the SPM6 design is based on the template scaffold LBCA-TrpB, whose catalytic activity is lower than that of ANC3 TrpS complex (LBCA TrpB activity is ca. 58% that of ANC3 TrpS). In the case of the DE *pfTrpB*^{OB2} enzyme variant, a 300 % of activity recovery was observed.³³

The partial recovery observed for SPM6, is in part due to the dramatic loss of activity displayed by ANC3 TrpB in the absence of TrpA (97% of activity loss), which is more moderate in *pfTrpB* (69%). These numbers indicate that the total recovery of ANC3 activity is more demanding from an engineering point of view, and suggest that the new generated SPM6 variant still presents some predisposition towards TrpA regulation. This evidences that the distal mutations introduced in SPM6 variant successfully enhanced the stand-alone activity of ANC3 TrpB activity through intra-subunit allosteric effects, however, they did not completely free TrpB from the inter-subunit allosteric regulation exerted by TrpA. To our surprise, SPM6 in complex with TrpA showed the most efficient turnover tested in this work, which indicates that the combination of intra- and inter-allosteric effects can operate synergistically to successfully tune the O-to-C conformational ensemble and achieve high catalytic efficiencies.

Another secondary insight gained from this work comes from the analysis of how the TrpS conformational landscape is altered and conserved along the natural evolutionary pathway. The exploration of the conformational ensemble and the identification of the key conformationally-relevant SPM positions of LBCA, ANC2 and ANC3 phylogenetic nodes and their comparison with the previously studied modern *pfTrpS* revealed that the main allosteric pathways are not significantly altered along evolution. Indeed, the comparison of the generated SPM paths for the different enzymes reveals a

rather high number of shared positions, thus suggesting similar TrpB correlated motions among variants. The conservation of the conformationally-relevant positions also explains the common positions targeted in the different SPM-based strategies for SPM3 and SPM8 designs. Our findings reinforce the original approach based on the LBCA SPM analysis as a robust computational strategy that could be exploited for the rational engineering of TrpB enzyme variants either for improved stand-alone or in complex function. It also evidences that conformational heterogeneity and, in particular, the use of ancestral conformationally-rich scaffolds corresponds to a successful strategy for designing desired enzymatic functions.^{42, 45}

The approach presented in this work highlights that the exploration of the enzyme conformational ensemble is essential for successful computational enzyme design. The detection of the key conformationally-relevant positions and the combined analysis of its conservation along ancestral phylogenetic trees harbors meaningful information for solving the current challenge in computational enzyme design of distal active site prediction for enhanced function.

METHODS:

Molecular Dynamics simulations. *System preparation.* The crystal structure of the LBCA TrpS complex (LBCA TrpA + LBCA TrpB), with PDB accession code 5ey5 was used as starting structure. The missing X-Ray regions were added using Modeller web server. The ANC3 TrpS complex was constructed by homology modelling. The ANC2 TrpB, the SPM3, SPM6 and SPM8 variants were generated from the ANC3 TrpB template using the mutagenesis tool included in Pymol (<http://www.pymol.org/>). Isolated TrpB enzymes were generated by manually removing its corresponding TrpA subunit. MD parameters for the reaction intermediates for TrpA (IGP, GP3) and TrpB (Ain and Q₂) were generated with the antechamber and parmchk modules of AMBER16⁴⁶ using the general amber force-field (GAFF). The partial charges (RESP model) were set to fit the electrostatic potential generated at the HF/6-31G(d) level of theory using the Gaussian09 software package. The different reaction intermediates were placed in the TrpA and TrpB subunits by alignment to available X-ray structures. For the simulations of TrpS complexes, two different combinations of bound substrates were used: in the first simulations IGP was introduced in TrpA subunit, while Ain intermediate was bound in TrpB; in the second set, G3P was introduced in TrpA, while Q₂ intermediate was placed in TrpB. A total of 13 systems were generated: 4 wild-type TrpS complexes (LBCA TrpS and ANC3 TrpS with IGP-Ain and GP3-Q₂ intermediates), 6 isolated TrpB enzymes (LBCA TrpB, ANC3 TrpB and ANC2 TrpB at Ain and Q₂ intermediates) and 3 TrpB enzyme variants (SPM3, SPM6 and SPM8 at Q₂ intermediate).

Molecular dynamics simulation details. All enzyme structures were filled with buffer in a pre-equilibrated cubic box of 10 Angstrom using the TIP3P water model and neutralized by the addition of explicit counterions (Na⁺ and Cl⁻) using the AMBER 16 leap module. All subsequent calculations were performed using a modification of the amber99 force field (ff14SB). A two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on solute by a harmonic potential with a force constant of 500 kcal mol⁻¹Å⁻², and the second stage is an unrestrained minimization of all the atoms in the simulation cell. All systems were gently heated using seven 50 ps steps, incrementing the temperature 50 K each step under constant-volume and periodic boundary conditions. Hydrogen bonds were set to fixed lengths using the SHAKE algorithm. Long-range electrostatic effects were modeled using the particle-mesh-Ewald method. An 8 Å cutoff was applied to Lennard-Jones and electrostatic interactions. Decreasing harmonic restraints were applied to the protein (210, 165, 125, 85, 45, 10 kcal mol⁻¹Å⁻²) during the thermal equilibration, with the Langevin scheme used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was equilibrated without restrains for 20 ns with a 2fs timestep at a constant pressure of 1 atm. After the systems were equilibrated in the NPT ensemble, a production run MD simulation was performed for each system in the NVT ensemble and periodic-boundary conditions. Production runs were performed with the AMBER16 software. For the systems subjected to the fast MD-based screening (ANC3 TrpS, ANC3 TrpB, ANC2 TrpB, SPM3, SPM6, SPM8) a production run of 3

independent replicas 500 ns each (i.e. 1.5 μ s accumulated time for each system) were performed at 333 K.

Metadynamics simulations. Path of collective variables. For the LBCA systems we were interested in obtaining the Free energy landscape as a function of the COMM domain Open-to-Closed transition. To that end we followed the same protocol employed in our previous work where we applied well-tempered metadynamics simulations with a path of collective variables to construct the FEL for the *p*fTrpS complex, the isolated *p*fTrpB and the evolved variant *p*fTrpB^{OB2}.³¹ The LBCA FEL obtained here are therefore directly comparable to the previous FEL explorations of *p*fTrpS. Metadynamics⁴⁷ is an enhancing sampling technique that consists in the addition of external energy potentials at regular number of MD steps in order to encourage the system to escape from prior stable conformations overcoming energy barriers and visiting other energy minima. In particular, the external potentials are added to a selected degree of freedom, often referred to as collective variables (CVs). After sufficient simulation time, metadynamics provides a reliable estimation of the underlying free energy as a function of the CVs by summing the external potentials added along the simulation. Here we used a path of collective variables approach that describes the process under study (path of conformations from open to closed states obtained by linear interpolation between available X-ray data, see Supplementary Fig. 1). Specifically, the x axis represents the progression along the path, encompassing 15 conformations from an open (x value = 1) to closed state (x value =15), while y axis measures the mean square deviation (MSD) from the reference path provided. Guided by structural information we restricted the path of structures to the α -carbons of the COMM domain (residues 97-184) and a loop region located at the base of the COMM domain (residues 282-305). Given the high 3D structural similarity between LBCA and the modern enzymes as shown by the available X-ray data, the path of conformations previously generated for the modern *p*fTrpB enzyme perfectly matches that obtained for LBCA. The λ parameter was computed as 2.3 multiplied by the inverse of the mean square displacement between successive frames, 80.

Well-tempered and Multiple walkers. We used the PLUMED2 software package⁴⁸ and GROMACS 5.1.2 code⁴⁹ to perform the metadynamics simulations. First, we carried out a metadynamics simulation for the two systems targeted (LBCA TrpS and LBCA TrpB) at the A_{in} and Q₂ reactant intermediates starting from the preequilibrated structures through conventional MD simulations (see above). Initial Gaussian potentials of height 0.15 kcal mol⁻¹, deposited every 2 ps of MD simulation at 350 K, were gradually decreased on the basis of the well-tempered adaptive bias with a bias factor of 10. The well-tempered approach allows for a smooth convergence of the FEL reconstruction avoiding the risk of overfilling. Besides, the adaptive Gaussian width scheme, in which hills variance is adapted to local properties of the free-energy surface, was used. Second, we extracted ten snapshots from the initial metadynamics exploration mostly covering the conformational space sampled by each system. These ten snapshots were used as the starting structures for the multiple-walkers metadynamics simulations. This approach allows to increase the sampling of the conformational space and to reach convergence of individual-energy profiles. The ten replicas (walkers) are run in parallel and each walker reads the energy

quantities (external potentials) deposited by the others during the simulation time. In this context, the ten walkers collaborate together to reconstruct the FEL. For this case each replica was run for 50-60 ns, giving a total of 500-600 ns per system (i.e. accumulated simulation time of ca. 2.3 μ s).

Convergence. The convergence of the recovered FEL was assessed by monitoring the energy difference ($\Delta\Delta G$) between selected regions of the conformational surface along the simulation time (see Supplementary Fig. 7). The selected regions correspond to the open, partially closed and closed energy minimum found and also open and closed regions that exhibited higher in energy free energy values. For instance, in the LBCA TrpS system at Q₂ intermediate, the energy differences were computed between the partially closed local energy minimum found and the higher in energy closed and open regions (i.e. PC-C and PC-O energy differences).

Structural analysis. A set of structures from each local energy minimum were clustered to obtain representative metastable conformations (see Supplementary Fig. 8). For consistency with our previous work, the local energy minima and the associated representative structures were labeled as open (O), partially closed (PC) and closed (C) accordingly with the path of CV values (x axis); (O)=1-5, (PC)=5-10 and (C)=10-15.

Caver analysis. The analysis of the available tunnels for the entrance of L-Serine was performed with the CAVER 3.0 software.⁵⁰ In this study we analyze the LBCA TrpS and LBCA TrpB systems at Ain intermediate. A total of 200 snapshots for the selected local energy minima from the metadynamic trajectories were subjected to Caver analysis. In particular, we used the structures from LBCA TrpS (Ain) open and the LBCA TrpB (Ain) partially closed energy minima. The starting point for the calculations was chosen at the L-Ser active site coordinates by alignment of the LBCA metastable structures at Ain intermediate with the X-ray structure (PDB ID 5DW0), which contains the Aex1 intermediate co-crystallized (intermediate formed after L-Serine covalent attachment with PLP). According to the parameters used in this study, a spherical probe of 0.9 Å radius was selected with a weighting coefficient of 1, and clustering threshold of 12.0.

Shortest Path Map analysis. The Shortest Path Map (SPM) analysis was performed using the metadynamics simulation of LBCA TrpB, and the MD simulation of the ANC3 variant. The first step for SPM construction relies on the calculation of the inter-residue mean distances and correlation values observed along the MD simulations and the conversion of this information into a simplified graph. For each residue of the enzyme a node is created. Each pair of nodes that display a mean distance of less than 6 Å along the MD simulation time is connected through a line. The length of the connecting line between residues *i* and *j* is weighted according to their correlation value ($d_{ij} = -\log |C_{ij}|$). In this way, those pairs of residues exhibiting large correlation values (i.e. highly correlated, values closer to 1 or -1) will be connected through shorter lines, whereas long lines will be drawn for those presenting lower correlation values (i.e. non-correlated, values closer to 0). At this point, the generated graph is further simplified to identify the shortest path lengths. The algorithm goes through all nodes of the initial graph

and detects the shortest paths to go from the first until the last protein residue. Following this strategy, those lines in the graph that are shorter, i.e. the connecting residues are more correlated, and that play a substantial role in the enzyme conformational dynamics are detected. The generated graph that we called SPM is then drawn on the 3D structure of the enzyme. More details about our SPM tool can be found in the recent publications ²⁴ and ²⁰.

Bacterial Strains and Chemicals. The proteins that were analyzed in this study were expressed in *E. coli* strain BL21 Gold (DE3) (purchased from Agilent Technologies). All chemicals used herein were purchased from commercial sources and were of analytical grade or higher.

Cloning. The genes for SPM3, SPM6, and SPM8 were codon optimized for expression in *E. coli* and purchased from Thermo Fisher Scientific (GeneArt Gene Synthesis). The genes were then cloned into a pET21a vector in a coupled digestion/ligation reaction using *Bsal* and T4 DNA ligase,⁵¹ which allowed for an isopropyl- β -thiogalactopyranoside (IPTG) inducible expression with a C-terminal His₆-tag.

Gene Expression and Protein Purification. The *E. coli* expression strain BL21 (DE3) Gold was transformed with plasmids harboring the genes for the TrpB variants SPM3, SPM6 and SPM8 and grown in 4 L lysogeny broth (LB) medium supplemented with 150 mg/mL ampicillin and 40 μ M PLP at 37°C. When an OD₆₀₀ of 0.6 was reached, expression was induced by addition of 0.5 mM IPTG and the cultures were further incubated over night at 20°C. Cells were then harvested by centrifugation and suspended in 50 mM KP (pH 7.5), 300 mM KCl, 10 mM imidazole, and 20 mM PLP. Cells were disrupted by sonication (Branson Sonifier W-250D, 30 % amplitude, 3 min, 2 s pulse, 2 s pause) and cell debris and insoluble aggregates were removed by centrifugation. The target proteins were purified from the supernatant by nickel-affinity chromatography (HisTrapTM FF crude or HisTrap excel, 5 mL, GE Healthcare) applying a linear imidazole gradient (10 mM to 500 mM). This was followed by size exclusion chromatography (Superdex 75 HiLoad 26/600, GE Healthcare) using 50 mM potassium phosphate (pH 7.5), 300 mM KCl, and 500 mM imidazole. The purified proteins were then dripped into liquid nitrogen and stored at -80°C.

The proteins LBCA TrpB, LBCA TrpA, Anc3 TrpB, and Anc3 TrpA were taken from previous work.³⁹

Steady-State Enzyme Kinetics. In order to monitor TrpB activity, the difference in absorbance between indole and L-Trp was used ($\Delta\epsilon_{290} = 1890 \text{ M}^{-1}\text{cm}^{-1}$). The reactions were performed at 30°C and changes in absorption were monitored with a spectrophotometer (JASCO V-750). The experimental conditions included 50 mM potassium phosphate (pH 7.5), 180 mM KCl, 40 μ M PLP, saturating concentrations of one substrate (either L-Ser or indole) and varying concentrations of the second substrate. When a constant baseline absorption was reached, reactions were initiated by addition of TrpB or the TrpS complex. In the case of TrpS, TrpA was added in molar excess to ensure complete complex formation. Initial slopes were determined and divided by $\Delta\epsilon_{290}$ to give the initial velocities (V_i). The values obtained for V_i were divided by the total concentration of TrpB enzyme ($[E]_t$) and plotted

against the substrate concentration [S]. The Michaelis constant K_M and the turnover number k_{cat} were obtained by fitting to the Michaelis-Menten equation (1) using Origin 2019 (Origin Lab).

$$\frac{V_i}{[E]_t} = \frac{k_{cat} [S]}{K_M + [S]} \quad (1)$$

ACKNOWLEDGMENTS

We thank the Generalitat de Catalunya for the emerging group CompBioLab (2017 SGR-1707) and Spanish MINECO for project PGC2018-102192-B-I00. M. A. M. S. was supported by the Spanish MINECO for a PhD fellowship (BES-2015-074964), J. I. F. was supported by the European Community for Marie Curie fellowship (H2020-MSCA-IF-2016-753045) and Juan de la Cierva-Incorporación fellowship (IJCI-2017-34129). S.O. is grateful to the funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC-2015-StG-679001), and the Human Frontier Science Program (HFSP) for project grant RGP0054/2020. We thank Sonja Fuchs, Sabine Laberer, Christiane Endres and Jeannette Ueckert for excellent technical assistance.

REFERENCES

1. Benkovic, S.J. & Hammes-Schiffer, S. A Perspective on Enzyme Catalysis. *Science* **301**, 1196-1202 (2003).
2. Hammes, G.G., Benkovic, S.J. & Hammes-Schiffer, S. Flexibility, Diversity, and Cooperativity: Pillars of Enzyme Catalysis. *Biochemistry* **50**, 10422-10430 (2011).
3. Marti, S. *et al.* Theoretical insights in enzyme catalysis. *Chem. Soc. Rev.* **33**, 98-107 (2004).
4. Maria-Solano, M.A., Serrano-Hervas, E., Romero-Rivera, A., Iglesias-Fernandez, J. & Osuna, S. Role of conformational dynamics in the evolution of novel enzyme function. *Chem Commun (Camb)* **54**, 6622-6634 (2018).
5. Boehr, D.D., Nussinov, R. & Wright, P.E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789-796 (2009).
6. Petrović, D., Risso, V.A., Kamerlin, S.C.L. & Sanchez-Ruiz, J.M. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* **15** (2018).
7. Tokuriki, N. & Tawfik, D.S. Protein Dynamism and Evolvability. *Science* **324**, 203-207 (2009).
8. Gunasekaran, K., Ma, B. & Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins* **57**, 433-443 (2004).
9. Lisi, G.P. & Loria, J.P. Allostery in enzyme catalysis. *Curr Opin Struct Biol* **47**, 123-130 (2017).
10. Campbell, E.C. *et al.* Laboratory evolution of protein conformational dynamics. *Curr. Opin. Struct. Biol.* **50**, 49-57 (2018).
11. Jiménez-Osés, G. *et al.* The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat. Chem. Biol.* **10**, 431-436 (2014).
12. Nussinov, R., Tsai, C.J. & Ma, B. The underappreciated role of allostery in the cellular network. *Annu Rev Biophys* **42**, 169-189 (2013).
13. Lee, J. & Goodey, N.M. Catalytic contributions from remote regions of enzyme structure. *Chem Rev* **111**, 7595-7624 (2011).
14. Qu, G., Li, A., Acevedo-Rocha, C.G., Sun, Z. & Reetz, M.T. The Crucial Role of Methodology Development in Directed Evolution of Selective Enzymes. *Angew. Chem. -Int. Ed. Engl.* **59**, 13204-13231 (2020).
15. Zeymer, C. & Hilvert, D. Directed Evolution of Protein Catalysts. *Annu. Rev. Biochem.* **87**, 131-157 (2018).
16. Arnold, F.H. Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). *Angew. Chem. -Int. Ed. Engl.* **58**, 14420-14426 (2019).
17. Hauer, B. Embracing Nature's Catalysts: A Viewpoint on the Future of Biocatalysis. *ACS Catal.* **10**, 8418-8427 (2020).
18. Damborsky, J. & Brezovsky, J. Computational tools for designing and engineering enzymes. *Curr. Opin. Chem. Biol.* **19**, 8-16 (2014).
19. Ebert, M.C. & Pelletier, J.N. Computational tools for enzyme improvement: why everyone can – and should – use them. *Curr. Opin. Chem. Biol.* **37**, 89-96 (2017).
20. Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* DOI: **10.1002/wcms.1502**, e1502.
21. Świderek, K., Tuñón, I. & Moliner, V. Predicting enzymatic reactivity: from theory to design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 407-421 (2014).
22. Currin, A., Swainston, N., Day, P.J. & Kell, D.B. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev* **44**, 1172-1239 (2015).
23. Morley, K.L. & Kazlauskas, R.J. Improving enzyme properties: when are closer mutations better? *Trends Biotechnol.* **23**, 231-237.
24. Romero-Rivera, A., Garcia-Borràs, M. & Osuna, S. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catal.* **7**, 8524-8532 (2017).
25. Hyde, C.C., Ahmed, S.A., Padlan, E.A., Miles, E.W. & Davies, D.R. Three-dimensional structure of the tryptophan synthase alpha 2 beta 2 multienzyme complex from *Salmonella typhimurium*. *J Biol Chem* **263**, 17857-17871 (1988).
26. Hioki, Y. *et al.* The crystal structure of the tryptophan synthase beta subunit from the hyperthermophile *Pyrococcus furiosus*. Investigation of stabilization factors. *Eur J Biochem* **271**, 2624-2635 (2004).
27. Fleming, J.R. *et al.* Evolutionary Morphing of Tryptophan Synthase: Functional Mechanisms for the Enzymatic Channeling of Indole. *J. Mol. Biol.* **430**, 5066-5079 (2018).

28. Barends, T.R.M. *et al.* Structure and mechanistic implications of a tryptophan synthase quinonoid intermediate. *Chembiochem* **9**, 1024-1028 (2008).
29. Dunn, M.F. Allosteric regulation of substrate channeling and catalysis in the tryptophan synthase holoenzyme complex. *Archives of Biochemistry and Biophysics* **519**, 154-166 (2012).
30. Niks, D. *et al.* Allostery and Substrate Channeling in the Tryptophan Synthase Holoenzyme Complex: Evidence for Two Subunit Conformations and Four Quaternary States. *Biochemistry* **52**, 6396-6411 (2013).
31. Maria-Solano, M.A., Iglesias-Fernández, J. & Osuna, S. Deciphering the Allosterically Driven Conformational Ensemble in Tryptophan Synthase Evolution. *J. Am. Chem. Soc.* **141**, 13049-13056 (2019).
32. Lee, S.J. *et al.* Conformational changes in the tryptophan synthase from a hyperthermophile upon alpha(2)beta(2) complex formation: Crystal structure of the complex. *Biochemistry* **44**, 11417-11427 (2005).
33. Buller, A.R. *et al.* Directed evolution of the tryptophan synthase beta-subunit for stand-alone function recapitulates allosteric activation. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 14599-14604 (2015).
34. Buller, A.R. *et al.* Directed Evolution Mimics Allosteric Activation by Stepwise Tuning of the Conformational Ensemble. *J. Am. Chem. Soc.* **140**, 7256-7266 (2018).
35. Romney, D.K., Murciano-Calles, J., Wehrmuller, J.E. & Arnold, F.H. Unlocking Reactivity of TrpB: A General Biocatalytic Platform for Synthesis of Tryptophan Analogues. *Journal of the American Chemical Society* **139**, 10769-10776 (2017).
36. Buller, A.R., van Roye, P., Murciano-Calles, J. & Arnold, F.H. Tryptophan Synthase Uses an Atypical Mechanism To Achieve Substrate Specificity. *Biochemistry* **55**, 7043-7046 (2016).
37. Herger, M. *et al.* Synthesis of beta-Branched Tryptophan Analogues Using an Engineered Subunit of Tryptophan Synthase. *Journal of the American Chemical Society* **138**, 8388-8391 (2016).
38. Murciano-Calles, J., Romney, D.K., Brinkmann-Chen, S., Buller, A.R. & Arnold, F.H. A Panel of TrpB Biocatalysts Derived from Tryptophan Synthase through the Transfer of Mutations that Mimic Allosteric Activation. *Angewandte Chemie-International Edition* **55**, 11577-11581 (2016).
39. Schupfner, M., Straub, K., Busch, F., Merkl, R. & Sterner, R. Analysis of allosteric communication in a multienzyme complex by ancestral sequence reconstruction. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 346-354 (2020).
40. Hochberg, G.K.A. & Thornton, J.W. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu. Rev. Biohys.* **46**, 247-269 (2017).
41. Merkl, R. & Sterner, R. Ancestral protein reconstruction: techniques and applications. *Biol. Chem.* **397**, 1 (2016).
42. Gardner, J.M., Biler, M., Risso, V.A., Sanchez-Ruiz, J.M. & Kamerlin, S.C.L. Manipulating Conformational Dynamics To Repurpose Ancient Proteins for Modern Catalytic Functions. *ACS Catal.* **10**, 4863-4870 (2020).
43. Busch, F. *et al.* Ancestral Tryptophan Synthase Reveals Functional Sophistication of Primordial Enzyme Complexes. *Cell Chem. Biol.* **23**, 709-715 (2016).
44. Motlagh, H.N., Wrabl, J.O., Li, J. & Hilser, V.J. The ensemble nature of allostery. *Nature* **508**, 331-339 (2014).
45. Crean, R.M., Gardner, J.M. & Kamerlin, S.C.L. Harnessing Conformational Plasticity to Generate Designer Enzymes. *J. Am. Chem. Soc.* (2020).
46. Case, D.A. *et al.* *AMBER 16*, University of California, San Francisco, 2016.
47. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **1**, 826-843 (2011).
48. Tribello, G.A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun* **185**, 604-613 (2014).
49. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845-854 (2013).
50. Chovancova, E. *et al.* CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol* **8**, e1002708 (2012).
51. Rohweder, B., Semmelmann, F., Endres, C. & Sterner, R. Standardized cloning vectors for protein production and generation of large gene libraries in Escherichia coli. *BioTechniques* **64**, 24-26 (2018).

Rational prediction of distal activity-enhancing mutations in tryptophan synthase

Miguel A. Maria-Solano,^{[a]#} Thomas Kinader,^{[b]#} Javier Iglesias-Fernández,^[a] Reinhard Sterner,^{[b]*} and Sílvia Osuna^{[a,c]*}

[a] CompBioLab group, Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, Girona, Spain

[b] Institute of Biophysics and Physical Biochemistry, Regensburg Center for Biochemistry, University of Regensburg, Universitätsstrasse 31, 93053 Regensburg, Germany

[c] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

*: These authors contributed equally to the work

Supplementary Information

Supplementary Table 1. Steady state enzyme kinetic parameters of LBCA TrpB and the LBCA TrpS complex.

Kinetic parameters	LBCA TrpB	LBCA TrpS
k_{cat} [s^{-1}]	0.51 ± 0.04	0.059 ± 0.004
$K_{\text{M}}^{\text{Ind}}$ [μM]	7.5 ± 1.6	9.1 ± 2.0
$k_{\text{cat}}/K_{\text{M}}^{\text{Ind}}$ [$\text{M}^{-1} \text{s}^{-1}$]	$6.6 \cdot 10^4$	$6.3 \cdot 10^3$
k_{cat} [s^{-1}]	0.48 ± 0.01	0.057 ± 0.002
$K_{\text{M}}^{\text{Ser}}$ [mM]	1.82 ± 0.13	0.28 ± 0.04
$k_{\text{cat}}/K_{\text{M}}^{\text{Ser}}$ [$\text{M}^{-1} \text{s}^{-1}$]	272	209

Experimental conditions: 50 mM potassium phosphate pH 7.5, 180 mM KCl, 40 μM PLP, 30°C. The concentration of the substrate added in excess was at least five times its respective K_{M} . For the measurements of TrpB, 0.2 μM TrpB LBCA were used, while for the LBCA TrpS complex 2.0 μM LBCA TrpA and 1.0 μM LBCA TrpB were employed. Under these conditions approximately 100% of LBCA TrpB was complexed with LBCA TrpA.

Supplementary Table 2. Steady state enzyme kinetic parameters of ANC3 TrpB and the ANC3 TrpS complex.

Kinetic parameters	ANC3 TrpB	ANC3 TrpS
k_{cat} [s^{-1}]	0.028 ± 0.001	0.91 ± 0.01
$K_{\text{M}}^{\text{Ind}}$ [μM]	9.8 ± 1.2	27.7 ± 1.5
$k_{\text{cat}}/K_{\text{M}}^{\text{Ind}}$ [$\text{M}^{-1} \text{s}^{-1}$]	$2.9 \cdot 10^3$	$30.7 \cdot 10^3$
k_{cat} [s^{-1}]	0.029 ± 0.001	0.81 ± 0.03
$K_{\text{M}}^{\text{Ser}}$ [mM]	2.4 ± 0.5	0.62 ± 0.07
$k_{\text{cat}}/K_{\text{M}}^{\text{Ser}}$ [$\text{M}^{-1} \text{s}^{-1}$]	12.1	$1.4 \cdot 10^3$

Experimental conditions: 50 mM potassium phosphate pH 7.5, 180 mM KCl, 40 μ M PLP, 30°C. The concentration of the substrate added in excess was at least seven times its respective K_M . For the measurements of TrpB, 2.0 μ M ANC3 TrpB were used, while for the ANC3 TrpS complex 5.0 μ M ANC3 TrpA and 0.2 μ M ANC3 TrpB were employed. Under these conditions TrpB was completely saturated with TrpA.

Supplementary Table 3. Steady state enzyme kinetic parameters of SPM6 TrpB and the SPM6 TrpS complex.

Kinetic constants	SPM6 TrpB	SPM6 TrpS
k_{cat} [s^{-1}]	0.23 ± 0.01	1.14 ± 0.07
K_M^{Ind} [μ M]	19.8 ± 1.9	56.8 ± 9.6
k_{cat}/K_M^{Ind} [$M^{-1} s^{-1}$]	$1.2 \cdot 10^4$	$1.9 \cdot 10^4$
k_{cat} [s^{-1}]	0.17 ± 0.004	1.06 ± 0.04
K_M^{Ser} [mM]	1.9 ± 0.1	1.02 ± 0.16
k_{cat}/K_M^{Ser} [$M^{-1} s^{-1}$]	89	$1.1 \cdot 10^3$

Experimental conditions: 50 mM potassium phosphate pH 7.5, 180 mM KCl, 40 μ M PLP, 30°C. The concentration of the substrate added in excess was at least five times its respective K_M . For the measurements of TrpB, 0.5 μ M TrpB SPM6 were used, while for the SPM6 TrpS complex 0.4 μ M ANC3 TrpA and 0.2 μ M TrpB SPM6 were employed. Under these conditions TrpB was completely saturated with TrpA.

Supplementary Table 4. Steady state enzyme kinetic parameters of SPM8 TrpB and the SPM8 TrpS complex.

Kinetic parameters	SPM8 TrpB	SPM8 TrpS
k_{cat} [s^{-1}]	0.15 ± 0.004	0.32 ± 0.016
K_M^{Ind} [μ M]	12.4 ± 1.1	19.0 ± 2.5
k_{cat}/K_M^{Ind} [$M^{-1} s^{-1}$]	$1.25 \cdot 10^4$	$1.55 \cdot 10^4$
k_{cat} [s^{-1}]	0.16 ± 0.005	0.27 ± 0.01
K_M^{Ser} [mM]	3.75 ± 0.3	0.51 ± 0.06
k_{cat}/K_M^{Ser} [$M^{-1} s^{-1}$]	41	578

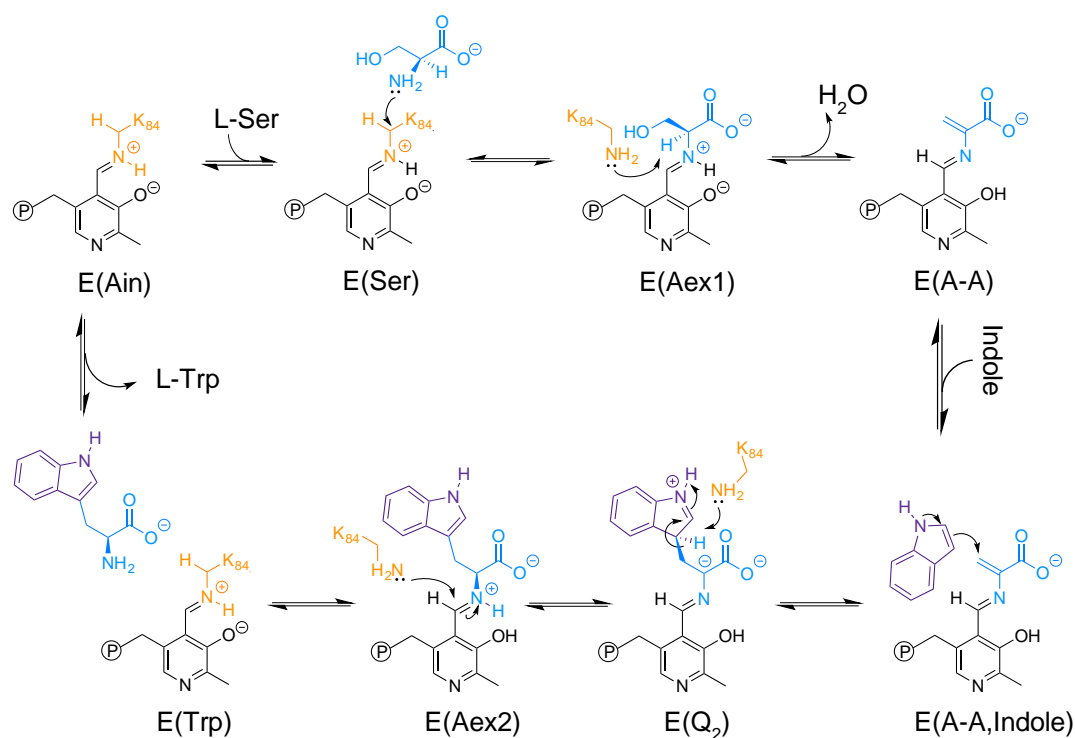
Experimental conditions: 50 mM potassium phosphate pH 7.5, 180 mM KCl, 40 μ M PLP, 30°C. The concentration of the substrate added in excess was at least six times its respective K_M . For the measurements of TrpB, 1.0 μ M TrpB SPM8 were used, while for the SPM8 TrpS complex 0.75 μ M ANC3 TrpA and 0.5 μ M TrpB SPM8 were employed. Under these conditions TrpB was completely saturated with TrpA.

Supplementary Table 5. Steady state enzyme kinetic parameters of SPM3 TrpB and the SPM3 TrpS complex.

Kinetic parameters	TrpB SPM3	SPM3 TrpS
k_{cat} [s^{-1}]	0.033 ± 0.002	0.83 ± 0.04
$K_{\text{M}}^{\text{Ind}}$ [μM]	16.3 ± 2.4	44.2 ± 5.3
$k_{\text{cat}}/K_{\text{M}}^{\text{Ind}}$ [$\text{M}^{-1} \text{s}^{-1}$]	$2.0 \cdot 10^3$	$1.6 \cdot 10^4$
k_{cat} [s^{-1}]	0.032 ± 0.001	0.88 ± 0.01
$K_{\text{M}}^{\text{Ser}}$ [mM]	0.38 ± 0.03	0.55 ± 0.03
$k_{\text{cat}}/K_{\text{M}}^{\text{Ser}}$ [$\text{M}^{-1} \text{s}^{-1}$]	85	$1.5 \cdot 10^3$

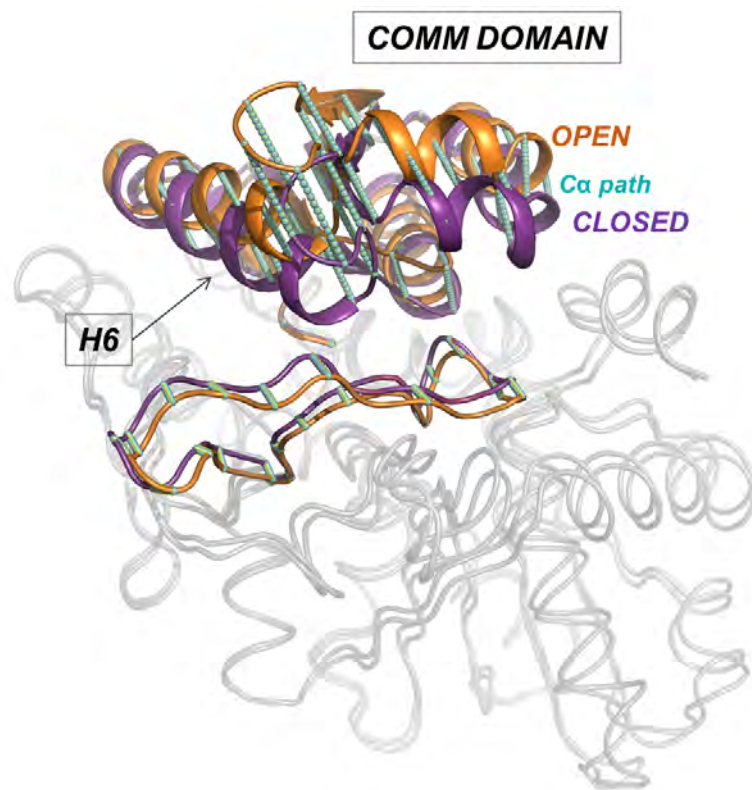
Experimental conditions: 50 mM potassium phosphate pH 7.5, 180 mM KCl, 40 μM PLP, 30°C. The concentration of the substrate added in excess was at least five times its respective K_{M} . For the measurements of TrpB, 1.5 μM TrpB SPM3 were used, while for the SPM3 TrpS complex 1.0 μM ANC3 TrpA and 0.5 μM TrpB SPM3 were employed. Under these conditions TrpB was completely saturated with TrpA.

Supplementary Scheme 1



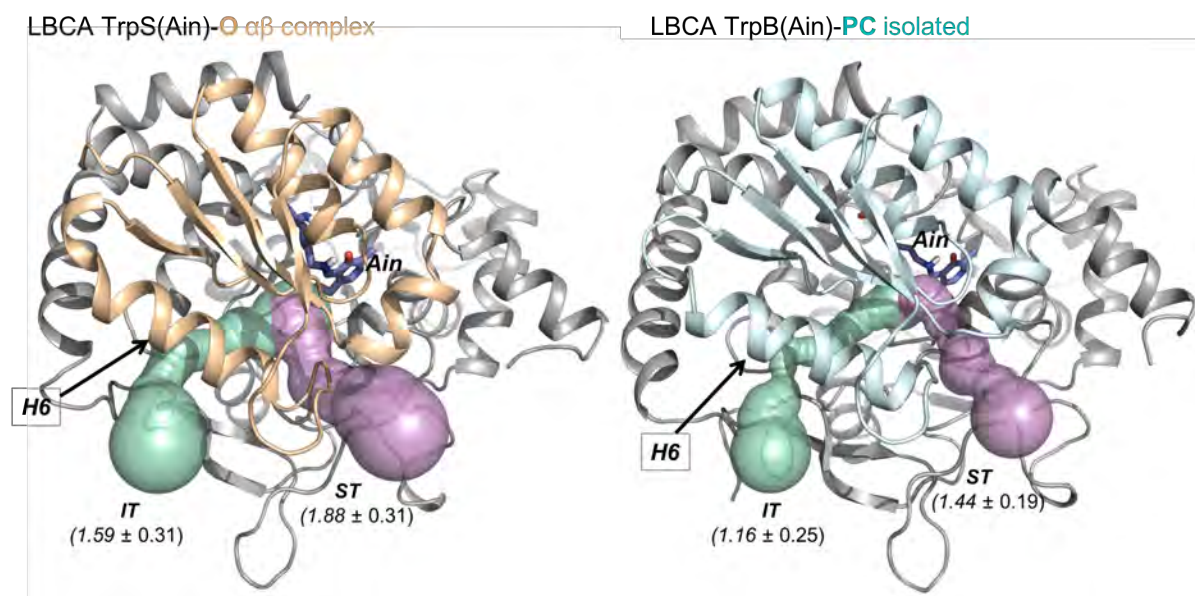
Supplementary Scheme 1. Tryptophan synthase reaction mechanism of the TrpB subunit. The catalytic K86 residue is colored in orange, while the L-Ser and indole substrates are blue and purple, respectively. In the resting state, a pyridoxal phosphate (PLP)-cofactor is covalently linked to the K84 active site residue, forming a Schiff base intermediate (E(Ain)). After transamination with L-serine E(Ser), an external aldimine intermediate E(Aex1) is formed. This intermediate undergoes deprotonation at C α , aided by K84, which is followed by a rapid elimination of the Aex1 hydroxyl group to form an electrophilic amino acrylate intermediate E(A-A). At this point, indole formed in TrpA reaches the TrpB active site and reacts with E(A-A) to form a quinonoid intermediate E(Q₂), which after proton extraction generates E(Q₃) (not shown). The reaction mechanism follows with the protonation at C α of Q₃ by K84 to form the E(Aex2) intermediate, which undergoes a second transamination reaction to finally release the L-tryptophan E(Trp) product and recover the resting state of the enzyme.

Supplementary Figure 1



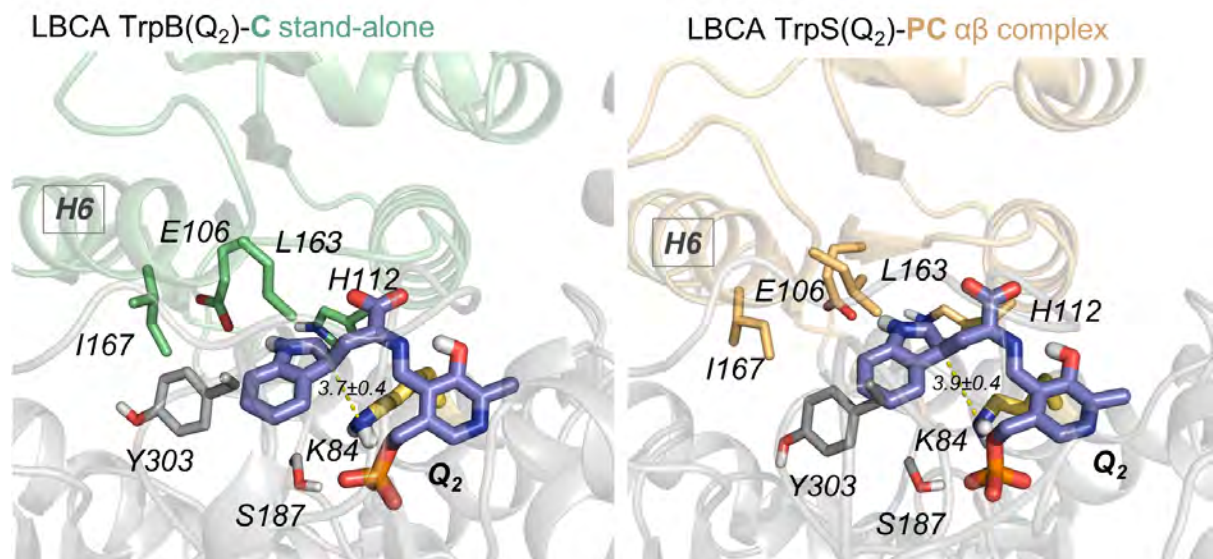
Supplementary Figure 1. Representation of the Open-to-Closed path of collective variables. The α carbon atoms encompassing the path (shown as cyan spheres) corresponds to the COMM domain residues (97-184) and the loop residues (282-305) going from the Open (PDB ID: 1WDW, in orange) to the Closed (PDB ID: 3CEP, in purple) X-ray structures.

Supplementary Figure 2



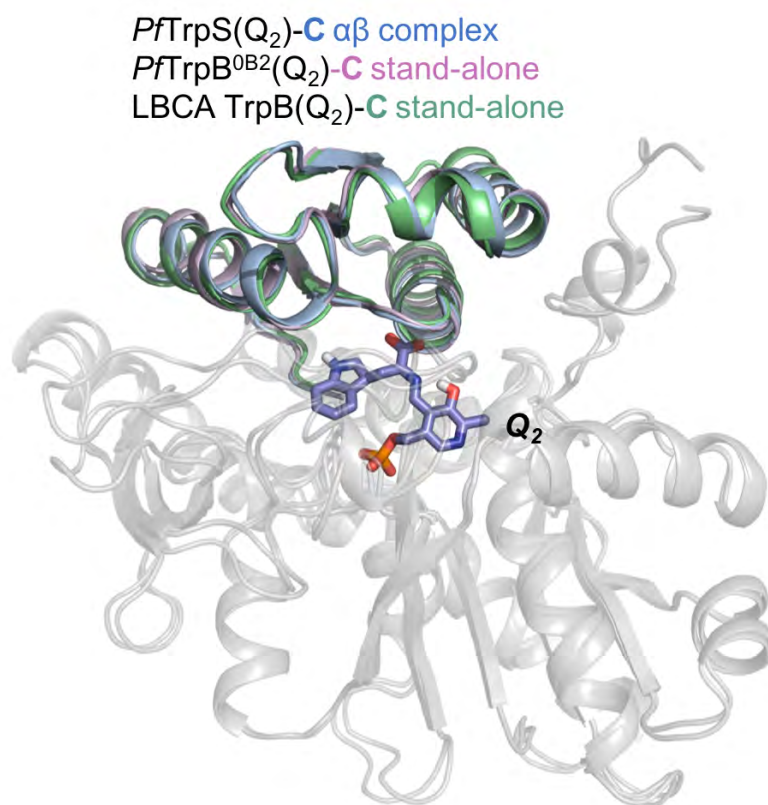
Supplementary Figure 2. Representation of the tunnel access at Ain reaction intermediate for the L-Ser substrate. The substrate access tunnels that were computed with CAVER 3.0 of the O and PC conformational ensembles corresponding to the LBCA TrpS complex and LBCA TrpB isolated respectively, revealed two different entry paths: The previously reported TrpA-TrpB internal tunnel (IT, green) and a secondary tunnel (ST, violet) that our group recently characterized.¹ The averaged bottleneck radii values (in Å) are also shown. We previously hypothesized that the secondary tunnel may play a role in the L-Ser entry but also in the L-Trp release.¹ The larger bottleneck radius of the ST in comparison with the IT highlights that the ST must be preferred for the L-Ser entry in both systems. The LBCA TrpS(Ain)-O state exhibited a larger bottleneck radius than the LBCA TrpB(Ain)-PC state in both tunnels, which indicates that the open state stabilization driven by the TrpA allosteric communication improves substrate accessibility.

Supplementary Figure 3



Supplementary Figure 3. Detailed active site view. Shown are the metastable conformations of the C state of LBCA TrpB (left) and PC state of LBCA TrpS (right) at the Q₂ reaction intermediate (violet) present in both structures. Active site residues are depicted in stick representation, the COMM domain is colored green (LBCA TrpB) or orange (LBCA TrpS), and catalytic K84 is highlighted in yellow. The number next to the yellow dashed line indicates the catalytic proton transfer distance (in Å) between K84 and the reaction intermediate

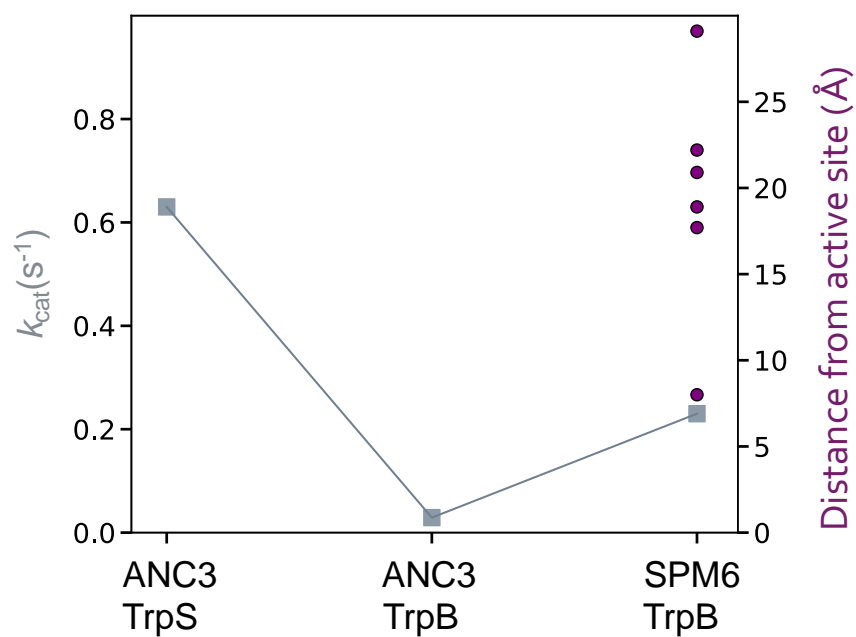
Supplementary Figure 4



Supplementary Figure 4. Overlays of the metastable conformations of the closed (C) states at Q₂ intermediate. The COMM domain is highlighted for *pf*TrpS (blue), the *pf*TrpB^{OB2} (pink) and LBCA TrpB (green). The *pf*TrpS and *pf*TrpB^{OB2} metastable structures were obtained from reference ¹.

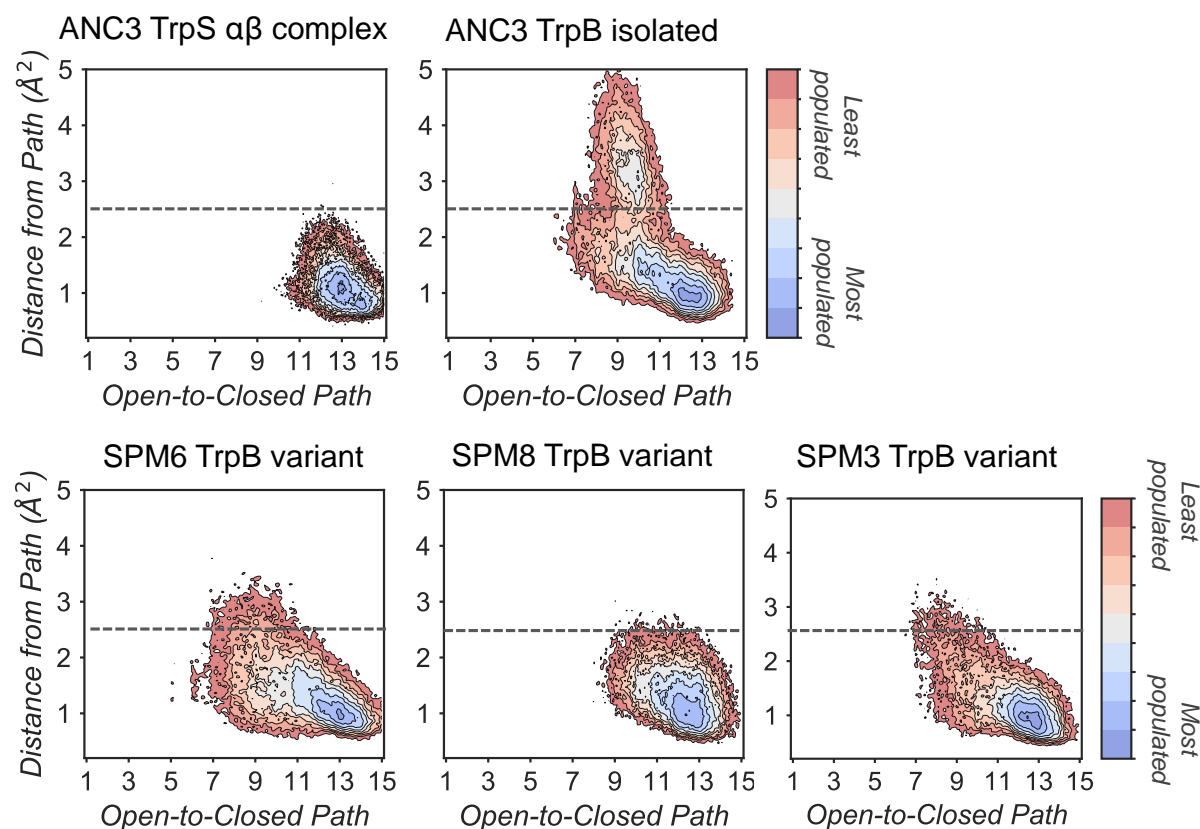
The FEL of the LBCA TrpB(Q₂) shows an energy minimum in the region that corresponds to a closed COMM conformations with large and low MSD distances equal in energy (Fig. 2a). In this context, we previously identified an energy minimum corresponding to low deviated closed conformations for the allosterically regulated *pf*TrpS complex and also for the stand-alone evolved variant *pf*TrpB^{OB2}.¹ These closed conformational ensembles displayed efficient active site preorganization by means of optimized non-covalent interactions networks and short catalytic distance between the Q₂ intermediate and the catalytic K84 that acts as proton acceptor. In particular the H6 was found to play an important role in the closure of the COMM domain and to form non-covalent interactions with the indole moiety of Q₂. Structural comparison between the metastable structure from the *pf*TrpS(Q₂)-C state, the *pf*TrpB^{OB2}(Q₂)-C and the LBCA TrpB(Q₂)-C from the low deviated closed ensemble obtained here showed that LBCA TrpB displays a highly similar degree of closure and also similar K84-Q₂ proton transfer distances. This indicates that LBCA TrpB has stand-alone properties due to the fact that it explores a stable catalytically competent closed conformation in the absence of TrpA.

Supplementary Figure 5



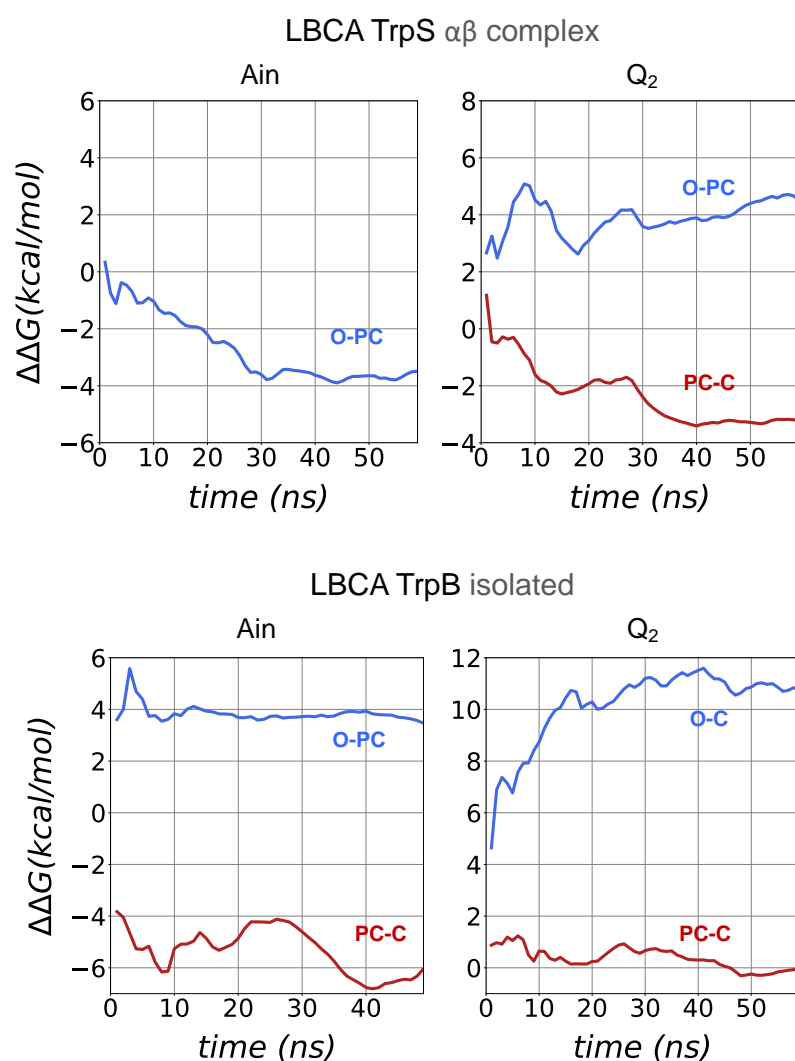
Supplementary Figure 5. Representation of the activity loss of ANC3 TrpB isolated and the activity recovery exerted by the SPM6 mutations. The catalytic activity (right y-axis) is represented in gray squares, while the distances between the α carbon atoms of the residues that were mutated in SPM6 with respect to the A_{in} reaction intermediate (C4A atom) are depicted in purple spheres.

Supplementary Figure 6



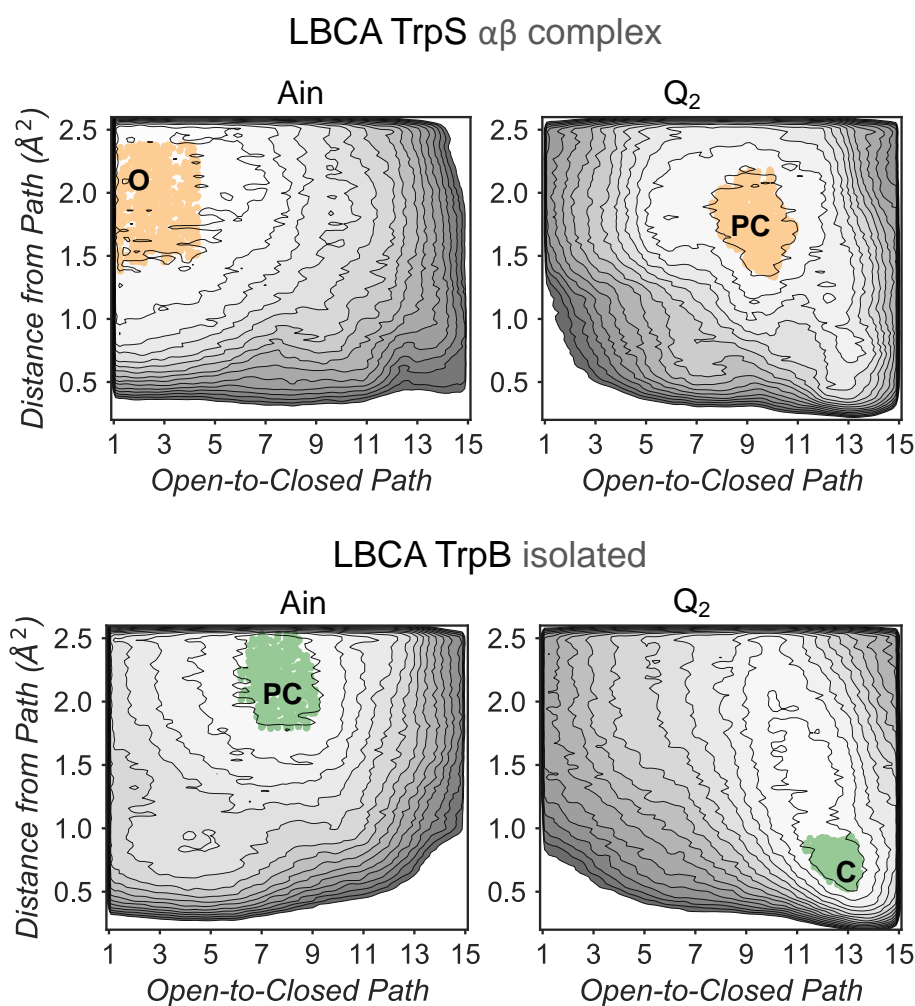
Supplementary Figure 6. Population analysis of the conventional molecular dynamics simulations. The MD data is plotted as a function of the same collective variables used for the metadynamics simulations (i.e. Open-to-Closed conformational transition from the X-ray data (reference path) and the MSD distance from the reference path). The dashed gray line indicates the y-axis top deviation value monitored in the metadynamics calculations (i.e. 2.5 Å²). Note that the inefficient ANC3 TrpB isolated samples non-productive COMM domain conformations (above the dashed line, high MSD distances), while the ANC3 TrpS in complex and the designed SPM enzyme variants better retain catalytically productive closed conformations (below the dashed line, low MSD distances). All the systems were simulated at the Q₂ intermediate.

Supplementary Figure 7



Supplementary Figure 7. Estimate of the differences in energy between selected regions of the FEL surface along the metadynamics simulations. The lines represent the mean $\Delta\Delta G$ value of the 10 walker replicas along the simulation time for the LBCA trpS complex and LBCA TrpB isolated at A_{in} and Q_2 intermediates. The line labels indicate the regions of the energy landscape that have been computed. With increasing simulation time, all lines tend to flatten, which is indicative of FEL convergence.

Supplementary Figure 8



Supplementary Figure 8. Projection of the conformations that correspond to the local energy minima coordinates from the metadynamics simulations on the FEL. The conformations projected are depicted as orange (LBCA TrpS complex) and green (LBCA TrpB isolated) dots. The representative metastable conformations of each local energy minimum presented in the main text were obtained clustering these set of structures.

REFERENCES

1. Maria-Solano, M.A., Iglesias-Fernández, J. & Osuna, S. Deciphering the Allosterically Driven Conformational Ensemble in Tryptophan Synthase Evolution. *J. Am. Chem. Soc.* **141**, 13049-13056 (2019).

Chapter 6. Results and discussion

In this chapter the main goals achieved in this thesis will be briefly discussed. To summarize the results, the chapter is divided in three sections. In the first place, the discussion is focused on the stereoselectivity and thermostability studies of alcohol dehydrogenase ADH enzymes and secondly, on the allosteric and stand-alone function studies of Tryptophan Synthase (TrpS) enzymes, which corresponds to the analysis of the main results gathered from **Chapter 4** and **Chapter 5**, respectively. The chapter finishes with a brief discussion about the link between the chemical step and the conformational dynamics of enzymes.

6.1 Alcohol dehydrogenase (ADH): enantioselectivity and thermostability studies

One of the most targeted enzymes for engineering enantioselectivity are alcohol dehydrogenases (ADH). ADHs are zinc-dependent enzymes that use NAD(P)H as cofactor, which delivers its hydride ion to the carbonyl group on the *Re* or *Si*-face of the *pro*-chiral ketone substrate yielding the corresponding (*S*) or (*R*)-alcohol. In an inspiring study from Lamed and coworkers, the active site shape of a thermophilic ADH enzyme from *Thermoethanolicus brockii* (TbADH) was speculated.^[176] They suggested that its structure would consist of two differently-sized active site pockets, one being larger than the other to accommodate the bulkier alkyl group of the *pro*-chiral ketone substituent.^[176] Interestingly, this hypothesis was later confirmed with the resolution of the crystal structure.^[177] Phillips rational site-specific mutagenesis studies indeed reported that by changing the size of the active site pockets the enantioselectivity and the substrate scope of the enzyme can be modulated.^[178] Reetz and coworkers successfully engineered the enantioselectivity of TbSADH on a rich array of substrates by applying CASTing, guided by the available crystal structure and Phillips studies.^[179]

In most experimental studies published, W110 and I86 positions located at the active site have been found to be key for enhancing the activity and reversing the enantioselectivity towards diverse bulky ketones.^[178-180] We hypothesized that these single point mutations might induce a significant shift on the conformations sampled by the enzyme, which may enable the accommodation of non-natural substrates and preferentially favors the formation of one enantiomer over the other. To shed further light on the enhanced enantioselectivity contribution of these two mutations, we decided to evaluate the conformational dynamics of wild-type TbSADH, and the singly-mutated TbSADH^{W110T} and TbSADH^{I86A} variants in the presence of

the *pro*-chiral ketone 4-alkediene cyclohexanone (**1a**), studied by Reetz and coworkers.^[179a] Experimentally, it was found that TbSADH is able to produce the corresponding (*R*)-alcohol but only with modest enantioselectivity (66 (*R*) % *ee*). In contrast, TbSADH^{W110T} exhibited (*R*)-enantioselectivity with 97 (*R*) % *ee*, whereas TbSADH^{I86A} reversed enantioselectivity with 98 (*S*) % *ee*.^[179a] Our MD simulations constrained the substrate bound to the Zn metal ion by imposing a force constant within the bonded model.^[84b, 124a, 181] This approach allows us to rationalize the preferences of the accommodation of the substrate in the active site along the simulation time. In this study, we coupled the MD simulations to active site volume calculations with POVME^[182] and the analysis of the most relevant non-covalent interactions with NCIplot^[183] in order to rationalize how favorable are the *pro*-(*R*) and *pro*-(*S*) conformations.^[84b]

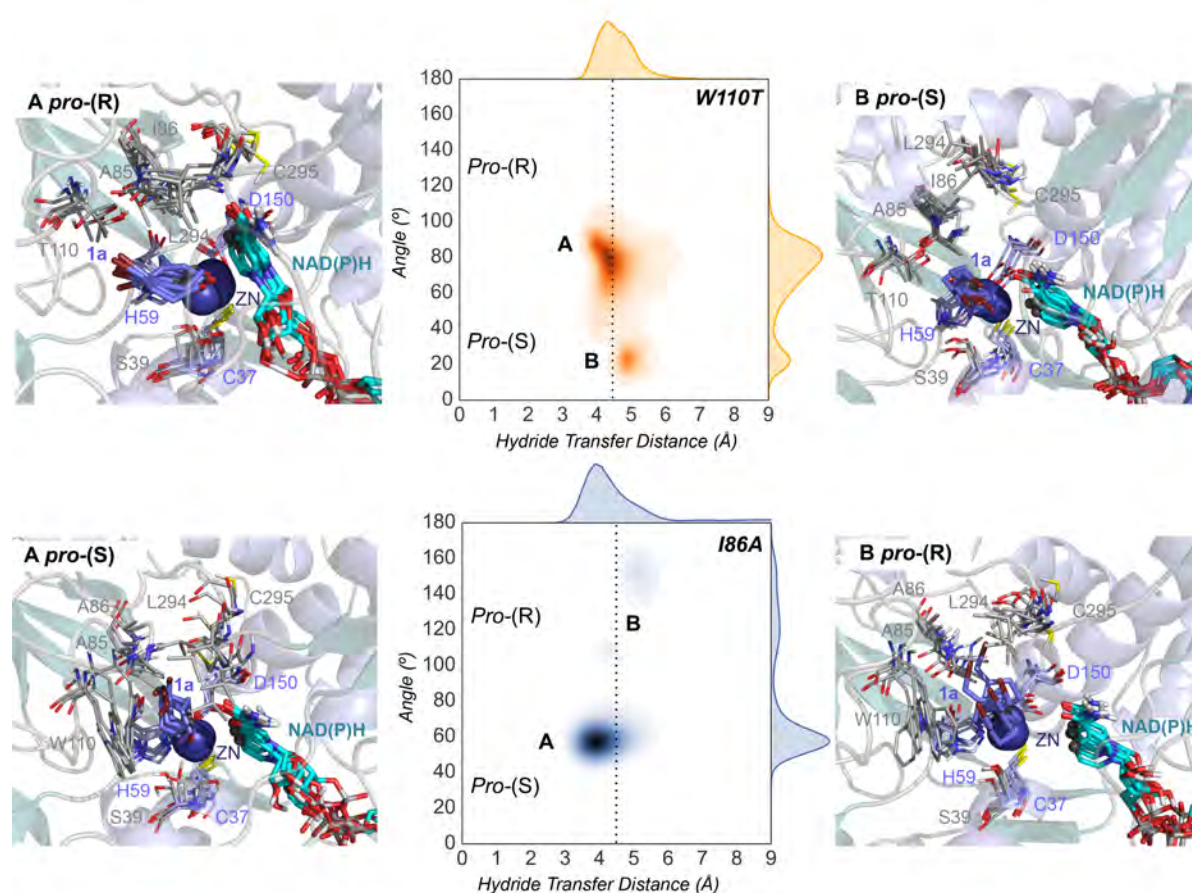


Figure 6.1. Representation of some representative snapshots of the different conformational states sampled along the MD simulations for the TbSADH^{W110T} and TbSADH^{I86A} starting from the *pro*-(*R*) (in orange) and *pro*-(*S*) (in blue) orientations of **1a**, respectively. The histogram of the hydride transfer distance together with the *pro*-(*R*)/*pro*-(*S*) angle between 1^a and an active site residue is displayed for both variants. High and low angle (in degrees) values represent *pro*-(*R*) and *pro*-(*S*) conformations, respectively. Short hydride transfer distances (in Å) values above the dashed line indicate catalytically productive orientations.

The conformational states sampled by the wild-type enzyme can position the **1a** substrate in a catalytically competent orientation for both *pro*-(R) and *pro*-(S) hydride transfer. However, the (R)- alcohol is substantially preferred, which is in line with the 66% ee observed in the experimental assays. The substitution of W110 by threonine alters the large binding pocket of the conformational states sampled, becoming even wider. The computed volume is *ca.* 165 Å³, whereas for the TbSADH was 100 Å³. The extra space released after this mutation allows the substrate to position the bromide into the large binding pocket stabilizing those conformations that adopt the catalytically productive *pro*-(R) orientation most of the simulation time (**Fig. 6.1** W110T *pro*- (R)) In contrast, the substitution of I86 by alanine enlarges the small binding pocket (from *ca.* 70 to 90 Å³). The additional space in the small binding pocket is occupied by the indole ring of W110, which favors the population of those conformations that better accommodate the *pro*-(S) productive orientation (**Fig. 6.1** I86A *pro*- (S)). The analysis of the non-covalent interactions occurring on the most populated conformational states sampled revealed how the active site pocket is remodeled to better stabilize the *pro*-(S) or *pro*-(R) orientations. These recent advances highlight the feasibility of MD simulations coupled with other computational tools such as POVME^[182] and NCIplot^[183] calculations for the engineering of natural enzyme active sites for enhanced enantioselectivity towards non-natural substrates.

In a later work, we addressed the engineering of the *TbADH* in the context of high activity at room temperatures towards the non-natural acetophenone substrate. Combining the virtues of pronounced enzyme robustness with high activity at ambient temperatures is of great interest for the production of chiral pharmaceuticals.^[184] As mentioned in the introduction (**section 1.5.3**), how catalytic efficiency adapts to temperature changes is currently poorly understood and the optimization of enzyme activity with little trade-off of thermostability and vice versa is a challenging task. The present study focuses on increasing activity of the *TbADH* (hyper)thermally stable enzyme at room temperatures while maintaining robustness utilizing a structure-guided directed evolution approach. Our goal is opposite to that of conventional thermostabilization of mesophilic enzymes by directed evolution, the usual alternative that is generally accompanied by a tradeoff in activity.^[185]

The best mutant screened, TbSADH-1 (A85G/I86A) was tested in upscaled reactions showing an excellent performance. At 30 °C ensured 96% conversion within 1.5 hour with complete enantio- selectivity (99% ee (R)). In contrast, at the same temperature WT TbSADH required 20 hours for a mere 4% conversion and 17% ee (S). The kinetic parameters show that TbSADH-1 variant displayed higher activity than WT at 30 °C (58-fold in k_{cat}) and 45 °C (51-fold in k_{cat}). Interestingly the robustness of the TbSADH-1 evolved was only lowered respect to the WT by 6 °C.

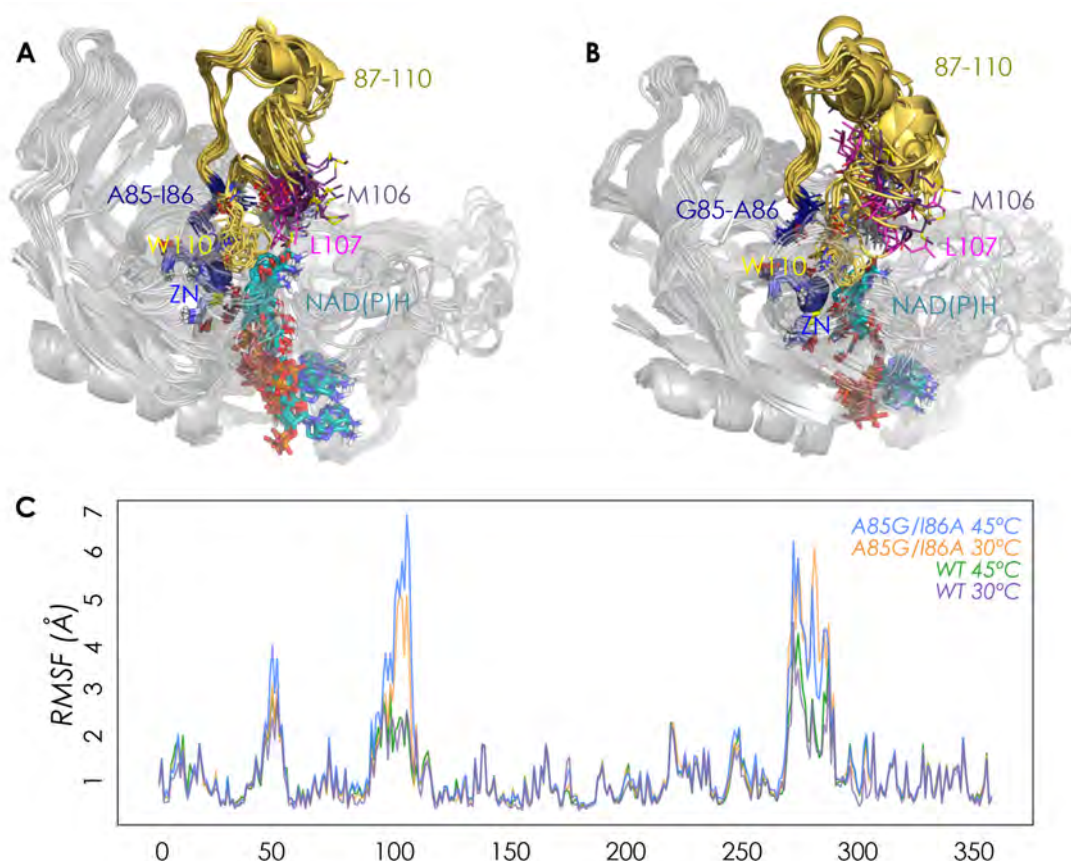


Figure 6.2 Overlay of representative snapshots for WT (A) and A85G/I86A variant (B) in the *apo* state at 30 °C. Root Mean Square Fluctuation (RMSF) values of all residues computed from the MD simulations in the *apo* state (C).

We then performed Molecular Dynamics (MD) simulations on the WT enzyme and the A85G/I86A variant, firstly for explaining the origin of dramatically enhanced activity at ambient temperature, and secondly to understand the reversed enantioselectivity. The analysis of RMSF in the *apo* state allows us to identify the most flexible regions of the enzyme structure, and rationalize the effect of the A85G/I86A mutations on the TbSADH conformational dynamics (see **Fig. 5.2**). The RMSF analysis reveals that the introduction of the A85G/I86A mutations increases the flexibility of a region composed by residues 87-110 that is partially

covering the active site (represented in yellow in **Fig. 6.2**). POVME^[182] calculations show that the A85G/I86A has a larger active site volume. The higher flexibility of the A85G/I86A variant, especially in the active site loop, confers the enzyme the ability to change the shape of the active site easily and to adapt to the new non-natural substrate, thus leading to higher activity at low temperatures.

We also performed MD simulations in presence of the acetophenone substrate in order to elucidate the origin of reversed enantioselectivity. The analysis of the non-covalent interactions using NCIplot^[183] calculations revealed how the higher flexibility of the active site loop 87–110 in the A85G/I86A variant plays a key role in dictating the enantioselectivity. This higher flexibility allows the acetophenone substrate to position the phenyl ring in the small binding pocket, thus stabilizing the *pro*-(R) orientation. Our results demonstrate the feasibility to evolve high activity of the thermophilic alcohol dehydrogenase TbSADH at ambient temperatures with excellent enantioselectivity and little tradeoff in thermostability. On the practical side, the present mutagenesis approach needs to be generalized by including other (hyper)thermostable enzymes. It will be interesting to see if flexibilization around the binding pocket is a general phenomenon characteristic of such laboratory-evolved enzymes.

6.2 Tryptophan synthase (TrpS): allostery and stand-alone function studies

Tryptophan Synthase (TrpS) is a heterodimeric enzyme complex composed by α (TrpA) - β (TrpB) subunits characterized by a tight allosteric coupling between them. In this context, the TrpA and TrpB reactions are synchronized, making them inefficient when isolated.^[94, 186] TrpS catalyzes the L-Tryptophan production in an interesting manner. TrpA catalyzes the retro-aldol cleavage of indole-3-glycerol phosphate (IGP) producing glyceraldehyde-3-phosphate (G3P) and indole; the latter being able to diffuse through an internal TrpA–TrpB tunnel to reach the TrpB active site. TrpB reaction is pyridoxal phosphate (PLP)-cofactor dependent and follows a multistep mechanism that involves many proton donor/abstraction reactions aided by K82. The TrpB resting state is characterized by a pyridoxal phosphate (PLP)-cofactor covalently linked to the K82 active site residue, forming a Schiff base intermediate (E(Ain)). After transamination with L-serine E(Ser), an external aldimine intermediate E(Aex1) is formed. This intermediate undergoes deprotonation at C α , assisted by K82, which is followed by a rapid elimination of the Aex1 hydroxyl group to form an electrophilic amino acrylate intermediate E(A–A). In the dimeric complex, indole formed in TrpA reaches the TrpB active site and reacts

with E(A–A) to form a quinonoid intermediate E(Q₂), which after proton extraction generates E(Q₃). At this point, protonation at C α of Q₃ by K82 forms the E(Aex2) intermediate, which undergoes a second transamination reaction to finally release the L-tryptophan E(Trp) product and restore the enzyme resting state. Available X-ray data show that both subunits explore an active site open-to-closed (O-to-C) transition along the catalytic cycle from open states at resting state to closed states as the reaction progresses. For the TrpB, the motion of a rigid COMM domain that is part of the active site cavity defines the O-to-C transition.

The use of TrpS for industrial purposes is hampered by its multimeric structure and its low activity of TrpB when isolated. In an insightful work, Arnold and coworkers applied DE to the TrpB subunit from *Pyrococcus furiosus*, resulting in the addition of 6 distal mutations to yield an efficient stand-alone catalyst (*Pf*TrpB^{OB2}).^[95] It is worth mentioning that the COMM domain structure is almost identical among different organisms (e.g., *Salmonella typhimurium* and *Pyrococcus furiosus*), isolated *Pf*TrpB enzyme, and *Pf*TrpB stand-alone variants, although all of them diverge in functionality. These observations suggest that the origin behind their different catalytic efficiencies could be attributed to alterations in the enzyme conformational ensemble induced by distal active site mutations (i.e. allosteric effects). Such effects had not been explored yet, although they are crucial to understand how the stand-alone functionality was achieved. Intrigued by the recovery of *Pf*TrpB^{OB2} activity in absence of the protein partner TrpA, we studied the conformational ensemble of the *Pf*TrpS complex, the *Pf*TrpB isolated and the *Pf*TrpB^{OB2} stand-alone variant employing enhanced sampling techniques. In particular, we applied metadynamics simulations^[167, 187] using a path of conformations that describes the allosteric COMM domain O-to-C transition found by X-ray data as collective variables. This approach allowed us to reconstruct the free energy landscape (FEL) associated with the O-to-C conformational dynamics. Several intermediates along the catalytic cycle were modeled. In particular, we selected the resting state E(Ain), E(Aex1), E(A–A), and E(Q₂) to evaluate the conformational exchange of the COMM domain found in X-ray data, but also to reproduce the multistep mechanism under study.

To elucidate the allosterically driven conformational ensemble of *Pf*TrpS complex, we reconstructed the FEL associated with the conformational dynamics of the COMM domain for a few selected reaction intermediates. The *Pf*TrpS complex FEL analysis shows that for the resting state the most favorable conformations are in the open state region, which agrees with its functional role in L-Serine substrate binding. As expected, as the reaction progresses a

population shift occurs towards partially closed and the catalytically active closed conformations. Interestingly, at quinonoid Q_2 intermediate (generated after indole coupling), *PfTrpS* samples all possible conformations of the COMM domain: O and PC states are almost equally stabilized, while the C state is higher in energy (**Fig. 5.3 A**, on the left). The C state of *PfTrpS*(Q_2) shows a highly preorganized active site with the catalytic residue K82 (proton acceptor) properly positioned for catalysis together with the indole moiety establishing many noncovalent interactions with the active site pocket. Comparison of *PfTrpS*(Q_2)-O and -C metastable structures show that the helix H6 closure is needed for forming key active site noncovalent interactions with the indole moiety (**Fig. 5.3 A**, on the right).

Experimental data showed that, in the absence of the allosteric partner *PfTrpA*, *PfTrpB* activity decreases 3.2-fold.^[95] The analysis of the FEL of *PfTrpB* isolated indicates that the absence of the TrpA allosteric communication restricts the COMM domain conformational heterogeneity along the cycle. In fact, only a single energy minimum is found at the Aex1 and Q_2 intermediates. In this way, the COMM domain is not able to escape from O states at Aex1, PC at Aex1, and C at Q_2 intermediates, as the other states are inaccessible (**Fig. 6.3 B**, on the left). In addition, the closed state sampled is highly deviated from the reference O-to-C conformational path (i.e. deviation larger than 1.5 Å). A detailed structural analysis of the isolated wild-type *PfTrpB*(Q_2) as compared to the *PfTrpS*(Q_2) complex in C states indicates that the isolated *PfTrpB* enzyme cannot efficiently sample catalytically competent C conformations; this is particularly true for the key COMM H6 closure and for the catalytic proton transfer distance K82- Q_2 , which is longer than in *PfTrpS* (**Fig. 6.3 B**, on the right). Our simulations have therefore shown that the isolated *PfTrpB* lacks the ability to easily access O, PC, and catalytic C states existing in the allosterically driven conformational ensemble of *PfTrpS*. Given the restricted conformational dynamics of the COMM domain in the absence of *PfTrpA* we decided to analyze whether distal mutations introduced in laboratory evolution were able to recover the allosterically driven conformational ensemble of *PfTrpS*.

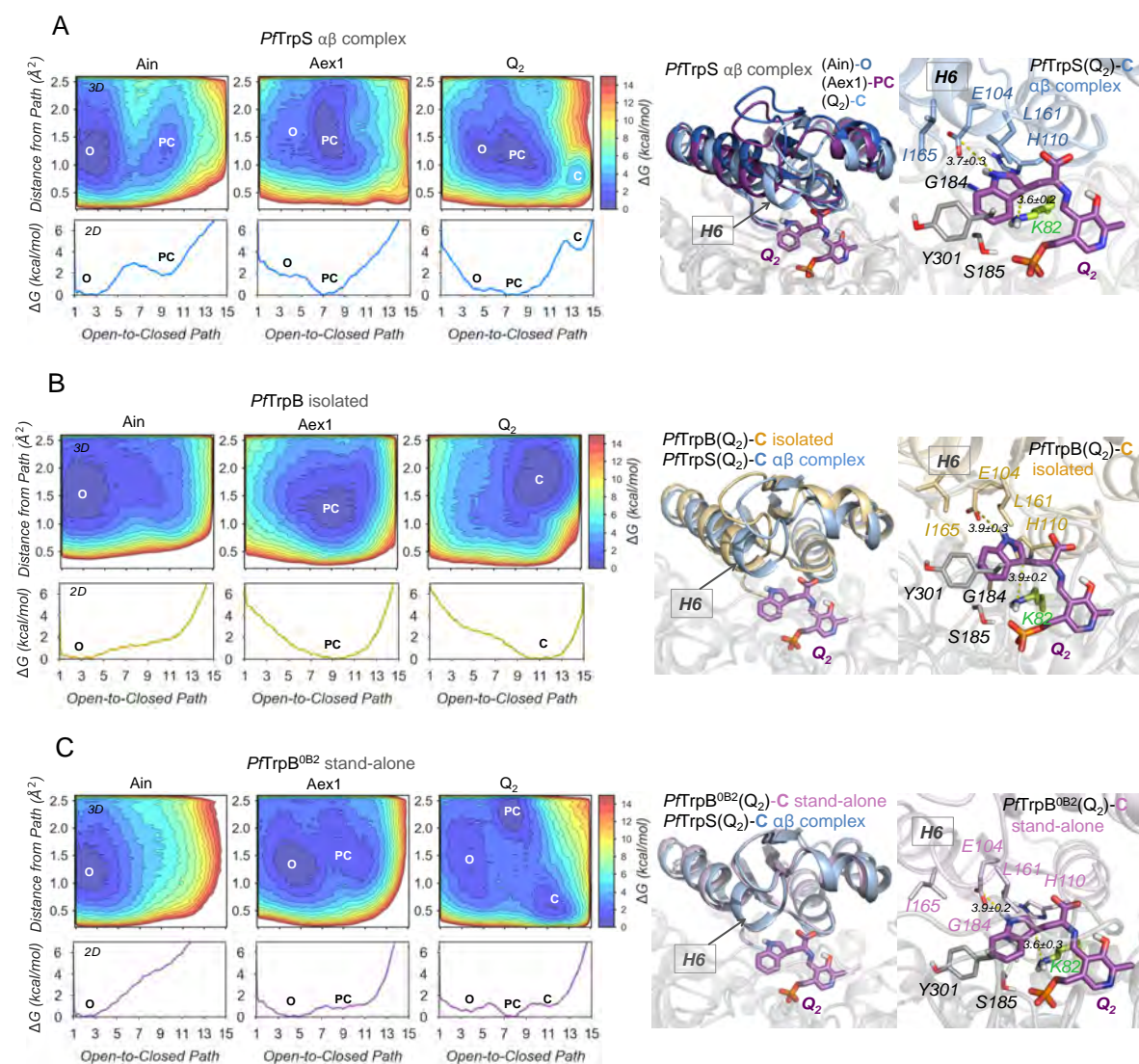


Figure 6.3 On the left, Free energy landscape (FEL) associated with the COMM domain open-to-closed (O-to-C) conformational exchange of the *PfTrpS* complex (A), *PfTrpB* isolated (B) and *PfTrpB*^{OB2} stand-alone (C) enzymes at Ain, Aex1, and Q₂ reaction intermediates. The x-axis corresponds to the progression along the reference O-to-C path generated from X-Ray data, while the y-axis to the mean square deviation (MSD) distance from the reference path. On the right, overlay of the *PfTrpS* metastable conformations of the open state at Ain intermediate, partially closed at Aex1, and closed at Q₂, respectively (A left panel), the metastable conformations of the closed states at Q₂ intermediate for *PfTrpB* and *PfTrpS* (B left) and the metastable conformations of the closed states at Q₂ intermediate for *PfTrpB*^{OB2} and *PfTrpS* (C left). The Detailed active site view of the metastable closed conformations at Q₂ intermediate for the *PfTrpS* complex (A right panel), *PfTrpB* isolated (B right) and *PfTrpB*^{OB2} stand-alone (C right) is also shown together with the catalytic distances (in \AA) between charge-charge stabilization E104-Q₂ and proton transfer K82-Q₂.

Experimentally, it was found that *PfTrpB*^{OB2} displays a considerably improved catalytic constant with respect to both the isolated wild-type *PfTrpB* (9.4-fold) and the *PfTrpS* complex (2.9-fold).^[95] By comparing the reconstructed FELs for stand-alone *PfTrpB*^{OB2} and *PfTrpS* complex along the different reaction intermediates, it becomes evident that the *PfTrpB*^{OB2} variant recovers the conformational heterogeneity of the COMM domain, characteristic of the allosterically regulated enzyme. In similarity with the *PfTrpS* allosteric complex, *PfTrpB*^{OB2} at Q₂ intermediate is able to sample O, PC and C states (**Fig. 5.3 C**, on the left). The *PfTrpB*^{OB2} O, PC and C states are substantially stable separated by low energy barriers (*ca.* 2 kcal/mol). This allows the COMM domain to easily adopt the catalytically competent C conformations. The high stability of the catalytically relevant C state contrasts with the *PfTrpS* system where the closed state is *ca.* 5 kcal/mol higher in energy. This difference in the C state stabilization explains the improved catalytic activity of the evolved stand-alone variant. The C state of stand-alone *PfTrpB*^{OB2}(Q₂) has an almost identical degree of closure of the COMM domain as the *PfTrpS* catalytically competent conformation, and a similar catalytic K82-Q₂ proton transfer distance (**Fig. 5.3 C**, on the right). This indicates that the C state of the stand-alone *PfTrpB*^{OB2} exhibits a proper preorganization for the reaction.

Finally, we were intrigued by the possibility of predicting distal positions for stand-alone function. To that end we relied on residue-by-residue correlation analysis. In particular we applied our Shortest Path Map (SPM) method.^[188] This computational tool identifies those enzyme pathways that have a higher contribution to the conformational dynamics of the enzyme in terms of correlated motions. We focused our analysis on the *PfTrpS*(Q₂) metadynamics trajectory because of the complete O-to-C conformational exchange sampled in it. *PfTrpB*^{OB2} presents six mutations: P12L, E17G, I68V, T292S, F274S, and T321A, from which two were directly predicted by the SPM tool, three were persistently interacting with a SPM positions and only one showed a minor role in the COMM domain conformational dynamics, making negligible interactions with SPM residues.

The present study demonstrates that fine-tuned control of the Open-to-Closed COMM domain conformational ensemble plays a key role along the TrpB catalytic cycle. Our new proposed methodology makes use of metadynamics simulations to enforce the sampling of the allosterically regulated O-to-C transition, and identifies which residues present a higher contribution to the O-to-C COMM domain conformational exchange through inter-residue

correlation calculations. With this new computational approach, distal positions involved in the allosteric transition can be identified, thus providing a set of key positions for the generation of smart libraries for stand-alone function. However, multiple positions are identified and there is a lack of information on which specific amino-acid substitution should be introduced for achieving an efficient conformational ensemble for stand-alone function. Indeed, the prediction of key conformationally relevant mutations for novel functionality, especially those located at remote positions from the active site, is an extremely difficult task in the enzyme design field.

In a following-up work, we focused on the development of a computational strategy for stand-alone function engineering targeting remote mutations using the SPM method. To that end, we were inspired by a previous work performed by our experimental collaborators. They were able to reconstruct the phylogenetic tree of TrpS, in particular they focused on the phylogenetic lineage that connects the last bacterial common ancestor (LBCA) to the modern *Neptuniibacter caesariensis* TrpS enzyme, which involves 6 intermediate nodes (ANC1-6).^[189] Careful analysis of the steady state kinetics revealed how TrpA exerts an allosteric inhibition in LBCA, which progressively shifts towards allosteric activation along the phylogenetic tree. ANC3 was the first TrpS enzyme that shows TrpA allosteric dependence. Indeed, the ANC3 TrpB activity decays dramatically in the absence of its allosteric activator ANC3 TrpA (30.2-fold decrease). Given the poor activity of ANC3 TrpB isolated we focus our computational design on the ANC3 TrpB scaffold.

The FEL associated to the COMM domain O-to-C transition for LBCA TrpB of the LBCA confirmed the LBCA TrpB stand-alone properties and the LBCA TrpA allosteric inhibition of LBCA TrpB activity found experimentally. LBCA TrpB adopts a stable catalytically productive COMM domain closure, which is hampered with the presence of the LBCA TrpA protein partner that limits the COMM domain ability for completing the O-to-C transition and achieving catalytically productive closed states. This exploration together with the findings of our previous work, elucidates the conformational ensemble that a stand-alone catalyst has to display for being efficient. This information is relevant for the designer in order to rationally progress towards the targeted enzyme design goal. In this study we exploited the ability of LBCA TrpB to stabilize catalytically competent COMM domain closed conformations when isolated (inherent stand-alone properties), and develop a novel computational enzyme design approach for achieving stand-alone function based on the LBCA TrpB conformational ensemble analysis through our SPM method. We hypothesized that the identification of the

conformationally-relevant SPM positions could be potential hotspots for tuning the conformational ensemble of TrpA-dependent TrpB enzymes, such as the targeted ANC3 TrpB enzyme. The SPM analysis reduced the sequence space from 20^{393} to 20^{74} possible mutations. However, this still leads to a massive amount of enzyme variants to screen. We solved this sequence dimensionality problem by analyzing the sequence conservation between LBCA TrpB and the targeted ANC3 TrpB system for the 74 SPM positions (see the workflow followed in **Fig. 6.4**). This sequence comparison of the identified SPM positions allowed us to decrease the SPM library to only 6 mutations yielding to the SPM6 enzyme variant. The nature of the specific amino acid transfer to the ANC3 template was then restricted to the natural amino acid found in LBCA TrpB. Interestingly, 5 out of 6 positions were located far away from the active site, and none is included in the COMM domain. This approach assumes that the transfer of the non-conserved conformationally relevant SPM mutations from the LBCA to the targeted ANC3 TrpB template will tune the ANC3 TrpB conformational ensemble towards a better stabilization of catalytically productive closed state through allosteric effects.

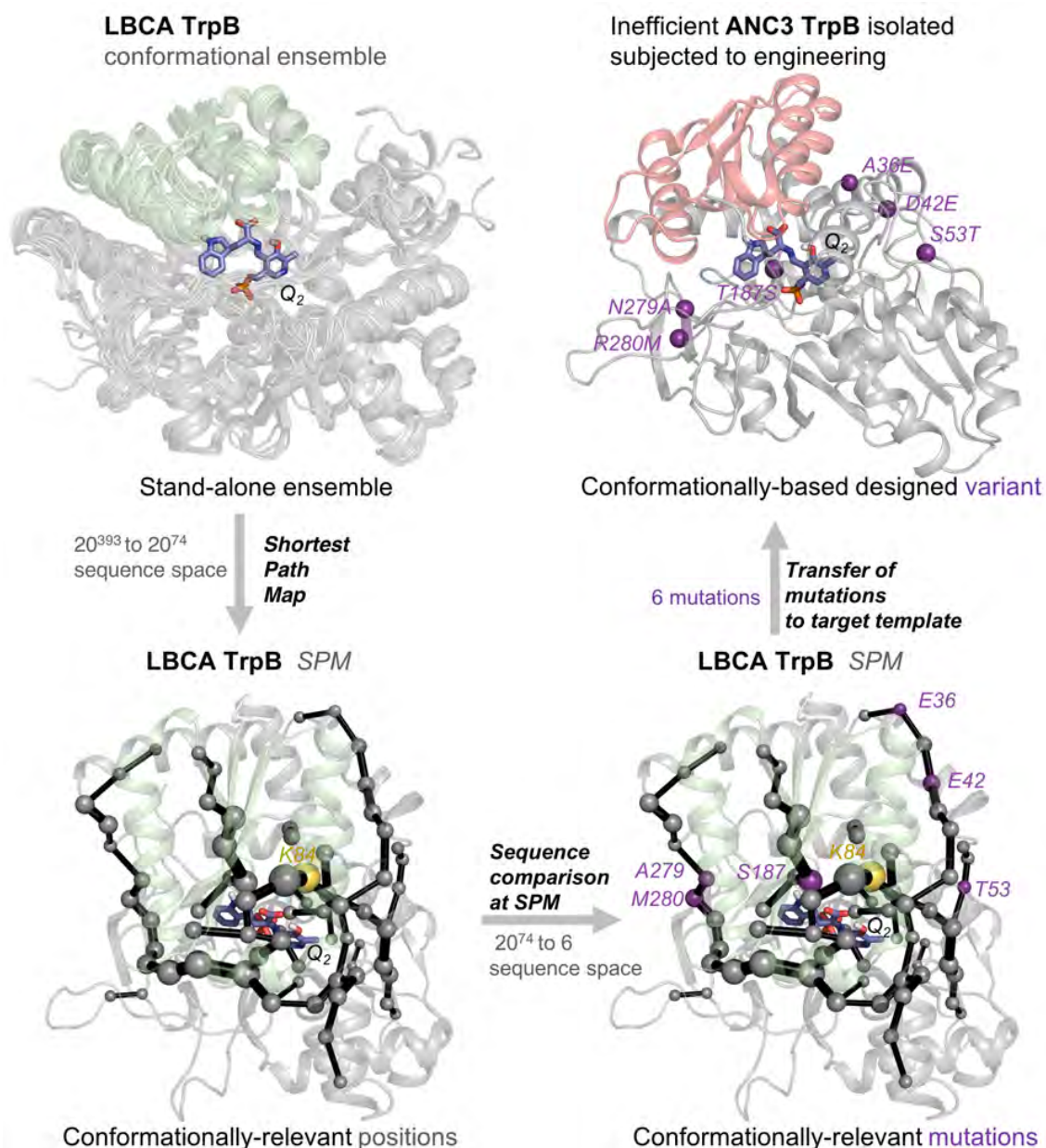


Figure 6.4 SPM-based computational workflow for the rational design of SPM6 TrpB enzyme variant. By analyzing the conformational ensemble of the stand-alone LBCA TrpB with high catalytic activity (upper left ensemble) through the SPM, we identified positions (grey spheres, lower left structure) within allosteric pathways (black edges) in the enzyme that most contribute to the LBCA TrpB conformational dynamics in the Q_2 intermediate. Thereby the size of each edge and node corresponds to the relevance for conformational dynamics; catalytic K84 is highlighted in yellow. Excluding residues that do not participate in an allosteric pathway reduces the sequence space from 20^{393} to 20^{74} possible activity enhancing substitutions. Sequence comparison at the SPM positions between stand-alone LBCA TrpB and inefficient ANC3 TrpB reduces the sequence space to 6 mutations with respect to LBCA TrpB (lower right structure, purple residues), that were introduced into ANC3 TrpB (upper right structure, purple residues) and tested *in-vitro*.

The experimental evaluation of the SPM6 enzyme variant indicated that the 6 introduced mutations successfully enhances the catalytic activity with respect to ANC3 TrpB *ca.* one order of magnitude (7-fold increase in k_{cat}). This enhancement obtained by solely testing a single variant is comparable to that observed for the laboratory evolved *PfTrpB*^{OB2} after three rounds of Directed Evolution that involved the screening of *ca.* 3,080 variants. The observed enhancement of ANC3 TrpB stand-alone activity still does not completely recover the activity displayed by the ANC3 TrpS complex. The new SPM6 designed variant enhances the low initial 3% activity displayed by ANC3 TrpB up to *ca.* 23%. It should be also mentioned that the SPM6 design is based on the template scaffold LBCA-TrpB, whose catalytic activity is lower than that of ANC3 TrpS complex (LBCA TrpB activity is *ca.* 58% that of ANC3 TrpS). In the case of the DE *pfTrpB*^{OB2} enzyme variant, a 300 % of activity recovery was observed.

The partial recovery observed for SPM6, is in part due to the dramatic loss of activity displayed by ANC3 TrpB in the absence of TrpA (97% of activity loss), which is more moderate in *pfTrpB* (69%). These numbers indicate that the total recovery of ANC3 activity is more demanding from an engineering point of view, and suggest that the new generated SPM6 variant still presents some predisposition towards TrpA regulation. This evidences that the distal mutations introduced in SPM6 variant successfully enhanced the stand-alone activity of ANC3 TrpB activity through inter-unit allosteric effects, however, they did not completely free TrpB from the intra-unit allosteric regulation exerted by TrpA. To our surprise, SPM6 in complex with TrpA showed the most efficient turnover tested in this work, which indicates that the combination of intra- and inter-allosteric effects can operate synergistically to successfully tune the O-to-C conformational ensemble and achieve high catalytic efficiencies. These findings indicate that our computational strategy could be exploited for the rational engineering of TrpB enzyme variants either for improved stand-alone or in complex function.

The approach presented in this work highlights that the exploration of the enzyme conformational ensemble is essential for successful computational enzyme design. The detection of the key conformationally-relevant positions and the combined analysis of its conservation along ancestral phylogenetic trees harbors meaningful information for solving the current challenge in computational enzyme design of distal active site prediction for enhanced function.

6.3 Ending Thoughts

The coupling between enzyme conformational dynamics and the chemical steps of the enzymatic reactions remains under debate.^[190] Along the catalytic turnover, enzymes have to search in the conformational space for those conformations more appropriate for binding the substrate, subsequently stabilizing the transition state(s) and finally releasing the product. Enzymatic motions take place at varying time scales (from femtosecond to second), but not all of them participate or play a role along the catalytic cycle. If a given a structural rearrangement is key for the turnover, such exchange should not have a time-scale longer than the enzymatic reaction. Otherwise, the rearrangement is not involved in the catalytic cycle. In this context, rearrangements that take place in time scales equal or shorter than the turnover may play an important role.

One could consider a hypothetical enzyme that presents a conformational exchange key for the turnover ranging in the same time scale to the one observed in the enzymatic reaction. If the introduction of a mutation accelerates such conformational transition, and as a consequence the enzymatic reaction rate increases, then one could think that the conformational exchange was rate limiting before mutation, and afterwards is the chemical step that determines the velocity of the enzymatic reaction. Thus, in these cases the efficiency of enzymes can be optimized by tuning the conformational dynamics up to a certain extend. However, considering the population distribution concept, there is still room for improvement. For those cases where the conformational exchange involved in the enzyme mechanism is accessible during the turnover, redistributing the relative stabilities of the enzyme conformational states that are interconnected by the conformational exchange may affect the enzymatic reaction rate. In this case, the introduction of a mutation that efficiently stabilizes the active conformational state that for instance triggers catalysis or aids the product release, increases the enzymatic reaction rate proportionally.

Regarding the results obtained in **Chapter 5.1** for the computational study of tryptophan synthase, the free energy calculations on the conformational exchange of the COMM domain indicate that upon substrate binding the rate for the open-to-closed conformational transition is relatively fast (in the nanosecond to microsecond time scale) in comparison with the reaction steps and turnover time scale (millisecond to second). This is a good example that illustrates

how by altering the relative stabilities of open, partially closed, and closed conformational states of the COMM domain through allosteric effects, each reaction step along the catalytic pathway can be efficiently optimized yielding to higher enzymatic reaction rates. A similar behavior was observed experimentally for adenylate kinase, where the substrate binding increases dramatically a domain open-to-closed rates of interconversion and yet exert influence on the much-slower turnover. They propose a paradigm for the mode of action of enzymes, in which numerous cycles of conformational rearrangement are required to find a mutual orientation of substrates that is optimal for the chemical reaction.^[191] In another insightful study of kemp eliminases, it has been shown experimentally how directed evolution gradually alters the enzyme conformational ensemble to populate a highly active conformational state that dramatically accelerates the enzymatic reaction.^[192]

Although the naturally occurring enzyme motions plays a crucial role in many enzymes, theoretically is very complicated to quantitatively relate enzyme reaction rates (i.e. catalytic constants) enhancements with the role of conformational dynamics. In part because it is difficult to define the boundaries of catalytically active regions in the conformational energy landscape and also because the quantitative characterization of the catalytic efficiency of the explored conformational states need to be estimated, like for instance with the Near Attack conformation analysis or theozyme-like conformations of the catalytic residues or much more accurately by means of EVB or QM/MM calculations at a higher computational cost. All together points out that a profound knowledge of both, chemical reaction and conformational dynamics events together with their relationship is crucial to completely understand enzyme reaction mechanisms. As shown in **Chapter 5.2**, a detailed analysis of the enzyme conformational ensembles harbors meaningful information that can be used in order to enhance enzyme catalytic efficiencies.

Chapter 7. Conclusions

In this thesis, we have explored the protein conformational energy landscape and its link with enzyme properties such as enantioselectivity, thermoadaptation and allostery. In particular, we have evaluated how the introduction of mutations alter the enzyme population distributions in laboratory-evolved enzyme variants. Following the conformational exploration, cross-correlation tools have been employed to characterize the enzyme pathways that most contribute to their conformational dynamism. This information has been used to identify potential hotspots and develop computational enzyme design strategies. In general, the studies of this thesis emphasize the importance of considering the enzyme conformational dynamics in computational design processes and highlights the concept that allostery is an intrinsic property of enzymes that can be exploited for enzyme evolution. This thesis also supports that the computational enzyme design can be adressed as a population shift problem.

The **main conclusions** regarding the different projects carried out are divided in **two blocks**:

I. Enantioselectivity and thermoadaptation studies:

- In **Chapter 4.1**, the bonded model protocol was applied to computationally explore the *pro*-(R) and *pro*-(S) conformations of a thermophilic zinc dependent ADH *Thermoethanolicus brockii* (*TbADH*) towards a non-natural substrate. The population distribution analysis associated with the *pro*-(R)/*pro*-(S) poses derived from the MD simulations evidences the poor enantioselectivity of the wild type *TbADH* enzyme and the enantio-preferences controlled by the DE mutations. Further in-depth structural analysis by means of volume and non-covalent interactions calculations reveal how the active site pocket is remodeled to better stabilize *pro*-(R) or *pro*-(S) conformations in the different laboratory-evolved enantioselective enzymes studied. Thus, the origin of enantioselectivity observed experimentally is explained. These advances show that the combination of MD simulations with other computational tools serves as a potential strategy to rationalize the enantioselectivity control and supports their feasibility to be implemented in engineering processes for enhanced enantioselectivity towards non-natural substrates.
- In **Chapter 4.2**, the enhanced activity and stereoselectivity towards a non-natural substrate of a *TbADH* laboratory-evolved variant by our experimental collaborators

was evaluated. Interestingly, the engineering strategy followed was the contrary to the usual thermostabilization of mesophilic enzymes, generally accompanied by a tradeoff in activity. The engineering approach carried out focuses in the thermoadaptation of a (hyper)thermally stable enzyme at room temperatures while maintaining its robustness. The RMSF analysis and active site volume calculations from MD simulations uncover the origin for thermoadaptation. The *TbADH* evolved variant shows a larger active site volume and higher flexibility of an active site loop, which confers the enzyme the ability to adapt to the bulky non-natural substrate at ambient temperatures. Further structural analysis through non-covalent interactions calculations reveal the molecular basis of the flexible active site loop role controlling enantioselectivity. This study points out that the fine-tuned flexibilization around the binding pocket can be a potential feature to target thermoadaptation engineering towards bulky non-natural substrates.

II. Allostery and stand-alone function studies:

- In **Chapter 5.1**, the conformational energy landscape of an allosteric complex enzyme from *Pyrococcus furiosus* TrpS (TrpA-TrpB), the isolated TrpB and a laboratory-evolved TrpB stand-alone variant was evaluated. The study reveals that the allosterically TrpA-driven conformational ensemble exhibits a fine-tuned control that plays a key role along the catalytic cycle. This catalytic conformational ensemble is hampered in the isolated TrpB enzyme. The lack of allosteric effects exerted by TrpA leads to a restricted conformational heterogeneity and the stabilization of unproductive catalytic states. In contrast, the DE introduced mutations recovered the allosterically driven conformational ensemble and better stabilize productive catalytic states, which explains its superior catalytic activity. Careful analysis of the *PfTrpS* conformational ensemble through Shortest Path Map (SPM) correlation-based tool elucidates the enzyme pathways most contributing to the TrpS conformational dynamics, which interestingly included some important DE positions. This study evidences that allosteric regulation and enzyme evolution obey the same natural laws and highlights that the mutations introduced along an evolutionary pathway may promote the stabilization of only those conformations preactivated for novel function.

- In **Chapter 5.2**, the promising use of SPM-based strategies for achieving stand-alone function was assessed. The insights gained in **Chapter 5.1** together with the previous work of our experimental collaborators where they reconstructed the phylogenetic tree of TrpS, were key for the successful computational design. The exploration of the conformational energy landscape of the Last bacterial common ancestor (LBCA) corroborates its stand-alone function properties tested experimentally. Careful inspection of the LBCA conformational ensemble through the SPM correlation-based tool identifies conformationally-relevant positions that can potentially alter the TrpB conformational ensemble. Sequence conservation analysis of the conformationally relevant SPM positions identified 6 mutations, where 5/6 were located distal from the active site, that could potentially enhance TrpB stand-alone activity. The experimental validation of the TrpB SPM6 rationally designed variant indicated a boost in the catalytic activity of near one order of magnitude, which evidences how the mutations introduced successfully achieved stand-alone properties by tuning the TrpB conformational ensemble through allosteric effects. This study corresponds to the proof of concept that the exploration of the conformational energy landscape is an essential factor in computational enzyme design. The rational prediction of key conformationally-relevant positions and the combined analysis of its conservation along ancestral phylogenetic trees, harbors meaningful information to address the current challenge of remote mutations prediction for enhanced function in computational enzyme design.

Bibliography

- [1] D. L. Nelson, M. M. Cox, *Lehninger Principles of Biochemistry*, Fifth ed., W. H. Freeman and Company.
- [2] a) J. H. Northrop, *Science* **1929**, *69*, 580; b) J. H. Northrop, M. Kunitz, *Science* **1931**, *73*, 262-263.
- [3] a) L. Michaelis, M. Menten, *Biochemistry* **1913**, 333-369; b) L. Michaelis, *et al.*, *Biochemistry* **2011**, *50*, 8264-8269.
- [4] J. C. Kendrew, *et al.*, *Nature* **1958**, *181*, 662-666.
- [5] J. D. Watson, F. H. Crick, *Cold Spring Harb Symp Quant Biol* **1953**, *18*, 123-131.
- [6] A. Fersht, *Enzyme structure and mechanism*, Second ed.
- [7] G. E. Briggs, J. B. Haldane, *Biochem J* **1925**, *19*, 338-339.
- [8] D. L. Purich, *Enzyme Kinetics: Catalysis & Control*, Elsevier, **2010**.
- [9] R. Chang, *Physical Chemistry for the Chemical and Biological Science*, Third ed., University Science Books.
- [10] R. Clausius, *The mechanical theory of heat*, Macmillan and Co., London, **1879**.
- [11] K. Sharp, F. Matschinsky, *Entropy* **2015**, *17*, 1971-2009.
- [12] W. Nernst, *Nachr. Kgl. Ges. Wiss. Gott* **1906**, *1*, 1-40.
- [13] Y. N. Harari, *Sapiens: A brief history of humankind* **2014**.
- [14] J. W. Gibbs, *On the equilibrium of heterogeneous substances*, the Academy, **1874**.
- [15] K. M. Polizzi, *et al.*, *Curr Opin Chem Biol* **2007**, *11*, 220-225.
- [16] D. M. Zuckerman, *Statistical Physics of Biomolecules*, CRC PressTaylor & Francis Group, **2010**.
- [17] D. M. Zuckerman, *Annu Rev Biophys* **2011**, *40*, 41-62.
- [18] J. Monod, *et al.*, *J Mol Biol* **1965**, *12*, 88-118.
- [19] M. Kovermann, *et al.*, *Proc Natl Acad Sci USA* **2017**, *114*, 6298-6303.
- [20] N. Tokuriki, D. S. Tawfik, *Science* **2009**, *324*, 203-207.
- [21] a) A. Romero-Rivera, *et al.*, *ACS Catalysis* **2017**, *7*, 8524-8532; b) E. Campbell, *et al.*, *Nat Chem Biol* **2016**, *12*, 944-950; c) B. Ma, R. Nussinov, *Nat. Chem. Biol.* **2016**, *12*, 890-891.
- [22] Ben E. Clifton, Colin J. Jackson, *Cell Chem Biol*, *23*, 236-245.
- [23] Sophie M. C. Gobeil, *et al.*, *Chem Biol* **2014**, *21*, 1330-1340.
- [24] A. Warshel, *et al.*, *Chem Rev* **2006**, *106*, 3210-3235.
- [25] M. I. Page, W. P. Jencks, *Proc Natl Acad Sci USA* **1971**, *68*, 1678-1683.
- [26] a) S. J. Benkovic, S. Hammes-Schiffer, *Science* **2003**, *301*, 1196-1202; b) G. Bhabha, *et al.*, *Science* **2011**, *332*, 234-238; c) R. G. Silva, *et al.*, *Proc Natl Acad Sci USA* **2011**, *108*, 18661-18665; d) D. R. Glowacki, *et al.*, *Nat Chem* **2012**, *4*, 169-176; e) S. C. L. Kamerlin, A. Warshel, *Proteins* **2010**, *78*, 1339-1375.
- [27] a) S. Hammes-Schiffer, S. J. Benkovic, *Annu Rev Biochem* **2006**, *75*, 519-541; b) K. A. Henzler-Wildman, *et al.*, *Nature* **2007**, *450*, 913-916; c) S. Osuna, *et al.*, *Acc Chem Res* **2015**, *48*, 1080-1089.

- [28] a) D. Kern, E. R. Zuiderweg, *Curr Opin Struct Biol* **2003**, *13*, 748-757; b) L. D. Handley, *et al.*, *Sci Rep* **2017**, *7*, 39575.
- [29] a) K. A. Henzler-Wildman, *et al.*, *Nature* **2007**, *450*, 838-844; b) A. Neu, *et al.*, *Nat Chem Biol* **2015**, *11*, 697-704; c) N. Tokuriki, C. J. Jackson, *Chem Biol* **2014**, *21*, 1259-1260.
- [30] B. M. Nestl, B. Hauer, *ACS Catalysis* **2014**, *4*, 3201-3211.
- [31] J. M. Axe, *et al.*, *J Am Chem Soc* **2014**, *136*, 6818-6821.
- [32] a) M. Orozco, *Chem Soc Rev* **2014**, *43*, 5051-5066; b) A. Romero-Rivera, *et al.*, *Chem Commun* **2017**, *53*, 284-297; c) G. Jiménez-Osés, *et al.*, *Nat Chem Biol* **2014**, *10*, 431-436.
- [33] W. S. Mak, J. B. Siegel, *Curr Opin Struct Biol* **2014**, *27*, 87-94.
- [34] M. D. Truppo, *ACS Med Chem Lett* **2017**, *8*, 476-480.
- [35] a) F. H. Arnold, *Q Rev Biophys* **2015**, *48*, 404-410; b) J. M. Woodley, *Curr Opin Chem Biol* **2013**, *17*, 310-316.
- [36] R. A. Chica, *et al.*, *Curr Opin Biotech* **2005**, *16*, 378-384.
- [37] a) M. T. Reetz, *Angew Chem Int Ed* **2011**, *50*, 138-174; b) M. T. Reetz, *et al.*, *Angew Chem Int Ed Engl* **2006**, *45*, 1236-1241.
- [38] A. Pavelka, *et al.*, *Nucleic Acids Research* **2009**, *37*, W376-W383.
- [39] R. J. Fox, *et al.*, *Nature Biotechnology* **2007**, *25*, 338.
- [40] T. Davids, *et al.*, *Curr Opin Chem Biol* **2013**, *17*, 215-220.
- [41] M. C. Ebert, J. N. Pelletier, *Curr Opin Chem Biol* **2017**, *37*, 89-96.
- [42] a) G. Kiss, *et al.*, *Angew Chem Int Ed* **2013**, *52*, 5700-5725; b) D. N. Bolon, S. L. Mayo, *Proc Natl Acad Sci U S A* **2001**, *98*, 14274-14279.
- [43] a) P. A. Romero, F. H. Arnold, *Nat Rev Mol Cell Biol* **2009**, *10*, 866-876; b) C. Jaeckel, *et al.*, *Annual Review of Biophysics* **2008**, *37*, 153-173; c) H. Renata, *et al.*, *Angew Chem Int Ed* **2015**, *54*, 3351-3367.
- [44] a) L. Jiang, *et al.*, *Science* **2008**, *319*, 1387-1391; b) D. Röthlisberger, *et al.*, *Nature* **2008**, *453*, 190-195; c) L. Giger, *et al.*, *Nat Chem Biol* **2013**, *9*, 494-498; d) O. Khersonsky, *et al.*, *Proc Natl Acad Sci USA* **2012**, *109*, 10358-10363; e) R. Blomberg, *et al.*, *Nature* **2013**, *503*, 418-421; f) E. A. Althoff, *et al.*, *Protein science : a publication of the Protein Society* **2012**, *21*, 717-726; g) J. K. Lassila, *et al.*, *Proc Natl Acad Sci USA* **2010**, *107*, 4937-4942; h) R. Obexer, *et al.*, *Chemcatchem* **2014**, *6*, 1043-1050; i) X. Garrabou, *et al.*, *Angew Chem Int Ed* **2015**, *54*, 5609-5612; j) X. Garrabou, *et al.*, *J Am Chem Soc* **2016**, *138*, 6972-6974.
- [45] Y. Kipnis, D. Baker, *Protein Science* **2012**, *21*, 1388-1395.
- [46] H. Kries, *et al.*, *Curr Opin Chem Biol* **2013**, *17*, 221-228.
- [47] D. D. Boehr, *et al.*, *Nat Chem Biol* **2009**, *5*, 789-796.
- [48] F. Richter, *et al.*, *PLoS ONE* **2011**, *6*, e19230.
- [49] D. J. Tantillo, *et al.*, *Curr Opin Chem Biol* **1998**, *2*, 743-750.
- [50] A. Zanghellini, *et al.*, *Protein Sci* **2006**, *15*, 2785-2794.
- [51] C. Malisi, *et al.*, *Proteins* **2009**, *77*, 74-83.
- [52] D. Hilvert, *Annu Rev Biochem* **2013**, *82*, 447-470.
- [53] H. K. Privett, *et al.*, *Proc Natl Acad Sci USA* **2012**, *109*, 3790-3795.
- [54] D. Rothlisberger, *et al.*, *Nature* **2008**, *453*, 190-195.

- [55] J. B. Siegel, *et al.*, *Science (New York, N.Y.)* **2010**, 329, 309-313.
- [56] F. Richter, *et al.*, *J Am Chem Soc* **2012**, 134, 16197-16206.
- [57] S. Bjelic, *et al.*, *ACS Chem Biol* **2013**, 8, 749-757.
- [58] D. Davidi, *et al.*, *Chem Rev* **2018**, 118, 8786-8797.
- [59] G. Kiss, *et al.*, *Protein Science* **2010**, 19, 1760-1773.
- [60] a) O. Khersonsky, *et al.*, *J Mol Biol* **2010**, 396, 1025-1042; b) O. Khersonsky, *et al.*, *J Mol Biol* **2011**, 407, 391-412; c) R. Blomberg, *et al.*, *Nature* **2013**, 503, 418-421.
- [61] A. Broom, *et al.*, *Nat Commun* **2020**, 11, 4808.
- [62] V. A. Risso, *et al.*, *Nat Commun* **2017**, 8, 16113.
- [63] V. A. Risso, *et al.*, *Chem Sci* **2020**, 11, 6134-6148.
- [64] R. Obexer, *et al.*, *Nat Chem* **2017**, 9, 50-56.
- [65] C. Zeymer, *et al.*, *J Am Chem Soc* **2017**, 139, 12541-12549.
- [66] a) D. De Raffe, *et al.*, *Acs Catalysis* **2020**, 10, 7871-7883; b) D. De Raffe, *et al.*, *Acs Catalysis* **2019**, 9, 2482-2492.
- [67] E. C. Campbell, *et al.*, *Curr Opin Struct Biol* **2018**, 50, 49-57.
- [68] M. A. Maria-Solano, *et al.*, *Chem Commun* **2018**.
- [69] a) N. S. Hong, *et al.*, *Nat Commun* **2018**, 9, 3900; b) S. C. Dodani, *et al.*, *Nat Chem* **2016**, 8, 419-425.
- [70] F. De Luca, *et al.*, *Proc Natl Acad Sci USA* **2011**, 108, 18424-18429.
- [71] a) H. M. Senn, W. Thiel, *Angew Chem Int Ed Engl* **2009**, 48, 1198-1229; b) M. W. van der Kamp, A. J. Mulholland, *Biochemistry* **2013**, 52, 2708-2728; c) D. T. Major, M. Weitman, *J Am Chem Soc* **2012**, 134, 19454-19462.
- [72] a) H. K. Privett, *et al.*, *Proc Natl Acad Sci USA* **2012**, 109, 3790-3795; b) G. Kiss, *et al.*, *Protein Sci.* **2010**, 19, 1760-1773.
- [73] a) J. A. Davey, R. A. Chica, *Protein Sci* **2012**, 21, 1241-1252; b) B. D. Allen, *et al.*, *Proc Natl Acad Sci USA* **2010**, 107, 19838-19843.
- [74] A. D. St-Jacques, *et al.*, *Acs Catalysis* **2019**, 9, 5480-5485.
- [75] a) H. J. Wijma, *et al.*, *Angew Chem Int Ed Engl* **2015**, 54, 3726-3730; b) R. Li, *et al.*, *Nature Chemical Biology* **2018**, 14, 664-670.
- [76] a) H. J. Wijma, *et al.*, *Protein Eng Des Sel* **2014**, 27, 49-58; b) R. J. Floor, *et al.*, *Proteins* **2015**, 83, 940-951.
- [77] S. Alonso, *et al.*, *Nat. Catal.* **2020**, 3, 319-328.
- [78] B. A. Amrein, *et al.*, *IUCrJ* **2017**, 4, 50-64.
- [79] A. Currin, *et al.*, *Chem Soc Rev* **2015**, 44, 1172-1239.
- [80] a) U. Doshi, *et al.*, *Proc Natl Acad Sci USA* **2016**, 113, 4735-4740; b) M. J. Holliday, *et al.*, *Structure* **2017**, 25, 276-286.
- [81] K. L. Morley, R. J. Kazlauskas, *Trends Biotechnol*, 23, 231-237.
- [82] K. Gunasekaran, *et al.*, *Proteins* **2004**, 57, 433-443.
- [83] a) Z. Sun, *et al.*, *Chemistry* **2016**, 22, 5046-5054; b) M. T. Reetz, *et al.*, *Angew Chem Int Ed Engl* **2005**, 44, 4192-4196.

- [84] a) H. J. Wijma, *et al.*, *J Chem Inf Model* **2014**, *54*, 2079-2092; b) M. A. Maria-Solano, *et al.*, *Org Biomol Chem* **2017**, *15*, 4122-4129; c) Z. Sun, *et al.*, *J Am Chem Soc* **2018**, *140*, 310-318; d) E. L. Noey, *et al.*, *Proc Natl Acad Sci USA* **2015**, *112*, E7065-E7072.
- [85] a) G. Li, *et al.*, *Chem Commun (Camb)* **2017**, *53*, 9454-9457; b) B. Yang, *et al.*, *Acs Catalysis* **2017**, *7*, 7593-7599.
- [86] V. G. Eijnsink, *et al.*, *J Biotechnol* **2004**, *113*, 105-120.
- [87] a) G. Vriend, V. Eijnsink, *J Comput Aided Mol Des* **1993**, *7*, 367-396; b) J. M. Finke, *et al.*, *Biochemistry* **2000**, *39*, 575-583; c) S. M. Doyle, *et al.*, *J Mol Biol* **2003**, *332*, 937-951.
- [88] a) F. H. Arnold, *et al.*, *Trends Biochem Sci* **2001**, *26*, 100-106; b) R. Sterner, W. Liebl, *Crit Rev Biochem Mol Biol* **2001**, *36*, 39-106.
- [89] a) F. D. Ciccarelli, *et al.*, *Science* **2006**, *311*, 1283-1287; b) V. Nguyen, *et al.*, *Science* **2017**, *355*, 289-294.
- [90] S. J. Kerns, *et al.*, *Nat Struct Mol Biol* **2015**, *22*, 124-131.
- [91] H. G. Saavedra, *et al.*, *Nature* **2018**, *558*, 324-328.
- [92] P. A. Fields, G. N. Somero, *Proc Natl Acad Sci USA* **1998**, *95*, 11476-11481.
- [93] a) R. B. Stockbridge, *et al.*, *Proc Natl Acad Sci USA* **2010**, *107*, 22102-22105; b) G. V. Isaksen, *et al.*, *Proc Natl Acad Sci U S A* **2016**, *113*, 7822-7827; c) J. Aqvist, *et al.*, *Nat. Rev. Chem.* **2017**, *1*, 14.
- [94] M. F. Dunn, *Archives of Biochemistry and Biophysics* **2012**, *519*, 154-166.
- [95] A. R. Buller, *et al.*, *Proc Natl Acad Sci USA* **2015**, *112*, 14599-14604.
- [96] A. R. Buller, *et al.*, *J. Am. Chem. Soc.* **2018**, *140*, 7256-7266.
- [97] B. F. Volkman, *et al.*, *Science* **2001**, *291*, 2429-2433.
- [98] a) M. W. Clarkson, *et al.*, *Biochemistry* **2006**, *45*, 7693-7699; b) J. Lee, N. M. Goodey, *Chem Rev* **2011**, *111*, 7595-7624.
- [99] J. D. Dunitz, *Chem Biol* **1995**, *2*, 709-712.
- [100] A. K. Bronowska, *Thermodynamics of Ligand-Protein Interactions: Implications for Molecular Design, Thermodynamics - Interaction Studies - Solids, Liquids and Gases*, InTech, **2011**.
- [101] C. A. MacRaild, *et al.*, *J Mol Biol* **2007**, *368*, 822-832.
- [102] a) D. Shukla, *et al.*, *Nat Commun* **2014**, *5*, 3397; b) Y. Meng, *et al.*, *J Mol Biol* **2018**, *430*, 4439; c) Y. Meng, *et al.*, *Proc Natl Acad Sci U S A* **2016**, *113*, 9193-9198.
- [103] a) L. G. Ahuja, *et al.*, *IUBMB Life* **2019**, *71*, 685-696; b) A. Cooper, D. T. Dryden, *Eur Biophys J* **1984**, *11*, 103-109.
- [104] A. J. Wand, *Science* **2001**, *293*, 1395.
- [105] N. Popovych, *et al.*, *Nat Struct Mol Biol* **2006**, *13*, 831-838.
- [106] J. Guo, *et al.*, *Structure* **2015**, *23*, 237-247.
- [107] E. M. Behiry, *et al.*, *Angew Chem Int Ed Engl* **2018**, *57*, 3128-3131.
- [108] D. Yang, L. E. Kay, *J Mol Biol* **1996**, *263*, 369-382.
- [109] J. Schlitter, *Chem Phys Lett* **1993**, *215*, 617-621.
- [110] a) I. Rivalta, *et al.*, *Proc Natl Acad Sci USA* **2012**, *109*, E1428-E1436; b) L. G. Ahuja, *et al.*, *Proc Natl Acad Sci U S A* **2019**, *116*, 15052-15061; c) A. Sethi, *et al.*, *Proc Natl Acad Sci USA* **2009**, *106*, 6620-6625.

- [111] a) P. M. Zakas, *et al.*, *Nat Biotechnol* **2017**, *35*, 35-37; b) A. Manteca, *et al.*, *Nat Struct Mol Biol* **2017**, *24*, 652-657.
- [112] V. A. Risso, *et al.*, *Curr Opin Struct Biol* **2018**, *51*, 106-115.
- [113] a) P. J. O'Brien, D. Herschlag, *Chemistry & Biology* **1999**, *6*, R91-R105; b) M. A. Siddiq, *et al.*, *Curr Opin Struct Biol* **2017**, *47*, 113-122.
- [114] P. Campitelli, *et al.*, *Annu Rev Biophys* **2020**, *49*, 267-288.
- [115] A. Romero-Rivera, *et al.*, *Chem Commun* **2017**, *53*, 284-297.
- [116] E. P. Barros, *et al.*, *J Chem Theory Comput* **2019**, *15*, 5703-5715.
- [117] P. A. M. Dirac, *Proc R Soc* **1929**, *123*, 714-733.
- [118] F. Himo, *J Am Chem Soc* **2017**, *139*, 6780-6786.
- [119] W. L. Jorgensen, *et al.*, *J Chem Phys* **1983**, *79*, 926-935.
- [120] S. Piana, *et al.*, *J Phys Chem B* **2015**, *119*, 5113-5123.
- [121] A. Halgren Thomas, *J Comput Chem* **1996**, *17*, 490-519.
- [122] C. I. Bayly, *et al.*, *J Phys Chem* **1993**, *97*, 10269-10280.
- [123] a) M. A. González, *Collection SFN* **2011**, *12*, 169-200; b) F. Jensen, *Introduction to computational chemistry*, **2017**; c) C. D. Berweger, *et al.*, *Chem Phys Lett* **1995**, *232*, 429-436.
- [124] a) J. M. Seminario, *International J Quantum Chem* **1996**, *60*, 1271-1277; b) L. Hu, U. Ryde, *J Chem Theory Comput* **2011**, *7*, 2452-2463; c) P. Li, *et al.*, *J Chem Theory Comput* **2013**, *9*, 2733-2748; d) F. Duarte, *et al.*, *J Phys Chem B* **2014**, *118*, 4351-4362.
- [125] a) P. Li, K. M. Merz, *J Chem Theory Comput* **2014**, *10*, 289-297; b) P. Li, *et al.*, *J Phys Chem B* **2015**, *119*, 883-895; c) P. Li, *et al.*, *J Chem Theory Computation* **2015**, *11*, 1645-1657.
- [126] P. Li, K. M. Merz, *J Chem Inf Model* **2016**, *56*, 599-604.
- [127] K. Nilsson, *et al.*, *Acta Cryst D* **2003**, *59*, 274-289.
- [128] J. Aqvist, A. Warshel, *J. Am. Chem. Soc.* **1990**, *112*, 2860-2868.
- [129] a) S. Y. Lu, *et al.*, *Proteins* **2012**, *81*, 740-753; b) Y. Jiang, *et al.*, *J Chem Inf Model* **2015**, *55*, 2575-2586.
- [130] L. Verlet, *Phys Rev* **1967**, *159*, 98-103.
- [131] M. P. Allen, D. J. Tildesley, *Computer simulation of liquids*, Clarendon Press, **1989**.
- [132] D. Beeman, *J Comput Phys* **1976**, *20*, 130-139.
- [133] E. Braun, *et al.*, *Living J Comput Mol Sci* **2019**, *1*.
- [134] T. Schlick, *et al.*, *Annu Rev Bioph Biom* **1997**, *26*, 181-222.
- [135] a) J.-P. Ryckaert, *et al.*, *J Comput Phy* **1977**, *23*, 327-341; b) D. J. Tobias, C. L. Brooks, *J Chem Phys* **1988**, *89*, 5115-5127.
- [136] B. Hess, *et al.*, *J Comput Chem* **1998**, *18*, 1463-1472.
- [137] M. Ringner, *Nat Biotechnol* **2008**, *26*, 303-304.
- [138] M. Ernst, *et al.*, *J Chem Phys* **2015**, *143*, 244114.
- [139] a) Y. Naritomi, S. Fuchigami, *J Chem Phys* **2013**, *139*, 215102; b) G. Pérez-Hernández, *et al.*, *J Chem Phys* **2013**, *139*, 015102.
- [140] R. R. Coifman, *et al.*, *Proc Natl Acad Sci USA* **2005**, *102*, 7426-7431.
- [141] F. Nuske, *et al.*, *J Chem Theory Comput* **2014**, *10*, 1739-1752.

- [142] M. Ceriotti, *et al.*, *Proc Natl Acad Sci USA* **2011**, *108*, 13023-13028.
- [143] G. Bussi, G. A. Tribello, *Methods Mol Biol* **2019**, *2022*, 529-578.
- [144] K. Henzler-Wildman, D. Kern, *Nature* **2007**, *450*, 964-972.
- [145] J. A. McCammon, *et al.*, *Nature* **1977**, *267*, 585-590.
- [146] P. L. Freddolino, *et al.*, *Structure* **2006**, *14*, 437-449.
- [147] D. E. Shaw, *et al.*, *Science* **2010**, *330*, 341-346.
- [148] K. Lindorff-Larsen, *et al.*, *Science* **2011**, *334*, 517-520.
- [149] R. O. Dror, *et al.*, *Nature* **2013**, *503*, 295-299.
- [150] A. C. Pan, *et al.*, *Drug Discov Today* **2013**, *18*, 667-673.
- [151] M. J. Harvey, *et al.*, *J Chem Theory Comput* **2009**, *5*, 1632-1639.
- [152] P. Eastman, *et al.*, *J Chem Theory Comput* **2013**, *9*, 461-469.
- [153] R. Salomon-Ferrer, *et al.*, *J Chem Theory Comput* **2013**, *9*, 3878-3888.
- [154] S. Pronk, *et al.*, *Bioinformatics* **2013**, *29*, 845-854.
- [155] J. C. Phillips, *et al.*, *J Comput Chem* **2005**, *26*, 1781-1802.
- [156] a) S. Olsson, F. Noé, *J Am Chem Soc* **2017**, *139*, 200-210; b) J. H. Prinz, *et al.*, *J Chem Phys* **2011**, *134*, 174105.
- [157] N. Plattner, F. Noé, *Nat Commun* **2015**, *6*, 7653.
- [158] M. M. Sultan, *et al.*, *Sci Rep* **2017**, *7*, 15604.
- [159] U. H. E. Hansmann, *Chemical Physics Letters* **1997**, *281*, 140-150.
- [160] a) P. H. Nguyen, *et al.*, *Proteins* **2005**, *61*, 795-808; b) J. W. Pitera, W. Swope, *Proc Natl Acad Sci USA* **2003**, *100*, 7587-7592.
- [161] G. M. Torrie, J. P. Valleau, *J Comput Phys* **1977**, *23*, 187-199.
- [162] S. Kumar, *et al.*, *J Comput Chem* **1992**, *13*, 1011-1021.
- [163] L. Maragliano, *et al.*, *J Chem Phys* **2006**, *125*, 24106.
- [164] B. Peters, B. L. Trout, *J Chem Phys* **2006**, *125*, 054108.
- [165] A. Laio, M. Parrinello, *Proc Natl Acad Sci USA* **2002**, *99*, 12562-12566.
- [166] G. Bussi, A. Laio, *Nat. Rev. Phys.* **2020**, *2*, 200-212.
- [167] A. Laio, F. L. Gervasio, *Reports on Progress in Physics* **2008**, *71*, 22.
- [168] D. Branduardi, *et al.*, *J Chem Phys* **2007**, *126*, 054103.
- [169] A. Barducci, *et al.*, *Phys Rev Lett* **2008**, *100*, 020603.
- [170] P. Raiteri, *et al.*, *J Phys Chem B* **2006**, *110*, 3533-3539.
- [171] a) D. Granata, *et al.*, *Proc Natl Acad Sci USA* **2013**, *110*, 6817-6822; b) G. Saladino, F. L. Gervasio, *Curr Opin Struct Biol* **2016**, *37*, 108-114.
- [172] D. Hamelberg, *et al.*, *J Chem Phys* **2004**, *120*, 11919-11929.
- [173] Y. Miao, *et al.*, *J Comput Chem* **2015**, *36*, 1536-1549.
- [174] P. R. Markwick, J. A. McCammon, *Phys Chem Chem Phys* **2011**, *13*, 20053-20065.
- [175] G. Csardi, T. Nepusz, *J Complex Syst* **2006**, *1695*, 1-9.
- [176] E. Keinan, *et al.*, *J Am Chem Soc* **1986**, *108*, 162-169.
- [177] a) O. Kleinfeld, *et al.*, *Biochemistry* **2000**, *39*, 7702-7711; b) C. Li, *et al.*, *Proteins* **1999**, *37*, 619-627; c) Y. Korkhin, *et al.*, *J Mol Biol* **1998**, *278*, 967-981.

- [178] a) M. M. Musa, *et al.*, *J Org Chem* **2007**, *72*, 30-34; b) M. M. Musa, *et al.*, *ChemCatChem* **2009**, *1*, 89-93; c) K. I. Ziegelmann-Fjeld, *et al.*, *Protein Eng Des Sel* **2007**, *20*, 47-55; d) M. M. Musa, *et al.*, *J Mol Catal B-Enzym* **2015**, *115*, 155-159.
- [179] a) R. Agudo, *et al.*, *J Am Chem Soc* **2013**, *135*, 1665-1668; b) Z. T. Sun, *et al.*, *Acs Catalysis* **2016**, *6*, 1598-1605; c) Z. T. Sun, *et al.*, *Tetrahedron Lett* **2016**, *57*, 3648-3651.
- [180] M. M. Musa, *et al.*, *Org Biomol Chem* **2008**, *6*, 887-892.
- [181] L. Hu, U. Ryde, *J Chem Theory Comput* **2011**, *7*, 2452-2463.
- [182] J. D. Durrant, *et al.*, *J Chem Theory Comput* **2014**, *10*, 5047-5056.
- [183] a) J. Contreras-Garcia, *et al.*, *J Chem Theory Comput* **2011**, *7*, 625-632; b) E. R. Johnson, *et al.*, *J Am Chem Soc* **2010**, *132*, 6498-6506.
- [184] *Organic Synthesis Using Biocatalysis*, Elsevier, **2015**.
- [185] a) A. S. Bommarius, M. F. Paye, *Chem Soc Rev* **2013**, *42*, 6534-6565; b) M. J. Liszka, *et al.*, *Annu Rev Chem Biomol Eng* **2012**, *3*, 77-102.
- [186] a) C. C. Hyde, *et al.*, *J Biol Chem* **1988**, *263*, 17857-17871; b) S. J. Lee, *et al.*, *Biochemistry* **2005**, *44*, 11417-11427.
- [187] A. Barducci, *et al.*, *Wiley Interdiscip Rev Comput Mol Sci* **2011**, *1*, 826-843.
- [188] A. Romero-Rivera, *et al.*, *Chem Commun* **2017**, *53*, 284-297.
- [189] a) F. Busch, *et al.*, *Cell Chem Biol* **2016**, *23*, 709-715; b) M. Schupfner, *et al.*, *Proc Natl Acad Sci USA* **2020**, *117*, 346-354.
- [190] a) A. Warshel, R. P. Bora, *J Chem Phys* **2016**, *144*, 180901; b) A. Kohen, *Acc Chem Res* **2015**, *48*, 466-473; c) K. Henzler-Wildman, D. Kern, *Nature* **2007**, *450*, 964-972; d) R. Callender, R. B. Dyer, *Acc Chem Res* **2015**, *48*, 407-413.
- [191] H. Y. Aviram, *et al.*, *Proc Natl Acad Sci USA* **2018**, *115*, 3243-3248.
- [192] R. Otten, *et al.*, *Science* **2020**, *370*, 1442-1446.

