

DEVELOPMENT OF NOVEL COMPUTATIONAL
PROTOCOLS FOR THE DESIGN OF EFFICIENT
ENZYMES

Guillem Casadevall i Franco



<http://creativecommons.org/licenses/by-nc-sa/4.0/deed.ca>

Aquesta obra està subjecta a una llicència Creative Commons Reconeixement-
NoComercial-CompartirIgual

Esta obra está bajo una licencia Creative Commons Reconocimiento-NoComercial-
CompartirIgual

This work is licensed under a Creative Commons Attribution-NonCommercial-
ShareAlike licence



DOCTORAL THESIS

Development of novel computational protocols
for the design of efficient enzymes

Guillem Casadevall i Franco

2024



DOCTORAL THESIS

**Development of novel computational protocols for the
design of efficient enzymes**

Guillem Casadevall i Franco

2024

Doctoral Programme in Chemistry

Supervisors:

Prof. Dr. Sílvia Osuna Oliveras

Dr. Javier Iglesias Fernandez

Tutor:

Prof. Marcel Swart

*Presented to obtain the degree of PhD at the
University of Girona*

List of Publications

This thesis is presented as a compendium of publications.

Published and submitted articles included in this Thesis:

1. Casadevall, G.; Duran, C.; Estévez-Gay, M.; Osuna, S. Estimating Conformational Heterogeneity of Tryptophan Synthase with a Template-based AlphaFold2 Approach. *Prot. Sci.*, **2022**, 31 (10), e4426. DOI:10.1002/pro.4426. [Biochemistry (Q1); Molecular Biology (Q1), JIQ: 8.1]
2. Casadevall, G.; Duran, C.; Osuna, S. AlphaFold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design. *JACS Au*, **2023**, 3 (6), 1554–1562. DOI:10.1021/jacsau.3c00188. [Chemistry (miscellaneous) (Q1), Organic Chemistry (Q1), Physical and Theoretical Chemistry (Q1), JIQ: 8.0]
3. Casadevall, G.; Casadevall, J.; Duran, C.; Osuna, S. The Shortest Path Method (SPM) Webserver for Computational Enzyme Design. *Protein Eng. Des. Sel.*, **2024**, 37, gzae005. DOI:10.1093/protein/gzae005. [Biochemistry (Q2); Bioengineering (Q2); Biotechnology (Q2); Molecular Biology (Q3), JIQ: 2.2]

Preprint articles included in this Thesis:

1. Casadevall, G.; Pierce, C.; Guan, B.; Iglesias-Fernandez, J.; Lim, H.-Y.; Greenberg, L. R.; Walsh, M. E.; Shi, K.; Gordon, W.; Aihara, H.; Evans, R. L.; Kazlauskas, R.; Osuna, S. Designing efficient enzymes: Eight predicted mutations convert a hydroxynitrile lyase into an efficient esterase. *BioRxiv*, **2023**. DOI:10.1101/2023.08.23.554512.

Published and submitted articles not included in this Thesis:

1. Palone, A.; Casadevall, G.; Ruiz-Barragan, S.; Call, A.; Osuna, S.; Bietti, M.; Costas, M. C–H Bonds as Functional Groups: Simultaneous Generation of Multiple Stereocenters by Enantioselective Hydroxylation at Unactivated Tertiary C–H Bonds. *J. Am. Chem. Soc.*, **2023**, 145 (29), 15742–15753. [Biochemistry (Q1); Catalysis (Q1); Chemistry (miscellaneous) (Q1), JIQ: 14.8]
2. Duran, C.; Casadevall, G.; Osuna, S. Harnessing Conformational Dynamics in Enzyme Catalysis to achieve Nature-like catalytic efficiencies: The Shortest Path Map tool for computational enzyme design. *Faraday Discuss.*, **2024**, Accepted Manuscript. [Physicial and Theoretical Chemistry (Q1), JIQ: 2.88]

List of Abbreviations

Abbreviation	Description
aaRSs	aminoacyl-tRNA Synthetases
AF2	AlphaFold2
BPTI	Bovine Pancreatic Trypsin Inhibitor
CASCO	CAlytic Selectivity by COmputational design
CFE	Consistent Force Field
COMM	COMMunication
DE	Directed Evolution
DFT	Density Functional Theory
DNN	Deep Neural Network
ELISA	Enzyme-Linked ImmunoSorbent Assay
EVB	Empirical Valence Bond
FEL	Free Energy Landscape
FF	Force Field
G	Gibbs Free Energy
GAFF	General AMBER Force Field
GDT	Global Distance Test
HF	Hartree-Fock
HNL	Hydroxynitrile Lyase
IDR	Intrinsically Disordered Regions
JIQ	Journal Impact Factor
KSI	Ketosteroid Isomerase
L-Trp	L-tryptophan
L-Ser	L-Serine
MAO-N	Monoamine Oxidase N
MD	Molecular Dynamics
ML	Machine Learning
MM	MOlecular Mechanics
MSD	Mean square deviation
MSA	Multiple Sequence Alignment
NACs	Near Attack Conformations

Abbreviation	Description
OPC	Optimal Point Charge
PCA	Principal Component Analysis
PDB	Protein Data Bank
PLP	Pyridoxal-5'-phosphate
PSSM	Position-Specific Scoring Matrices
PTE	Phosphotriesterase
QM	Quantum Mechanics
QM/MM	Quantum Mechanics/Molecular Mechanics
RESP	Restrained Electrostatic Potential
RF	RoseTTAFold
RFdiffusion	RoseTTAFold diffusion
SPM	Shortest Path Map
SPMweb	SPM webserver
TICA	Time-lagged Independent Component Analysis
TIP3P	Three-site Transferable Intermolecular Potential 3 Point
TrpA	Tryptophan Synthase Alpha Subunit
TrpB	Tryptophan Synthase Beta Subunit



La Prof. Dra. Sílvia Osuna Oliveras, de la Universitat de Girona, i el Dr. Javier Iglesias Fernandez de l'empresa Nostrum Biodiscovery,

DECLAREM:

Que el treball titulat "Development of novel computational protocols for the design of efficient enzymes", que presenta Guillem Casadevall i Franco per a l'obtenció del títol de doctor, ha estat realitzat sota la nostra direcció i que compleix els requisits per poder optar a Menció Internacional.

I, perquè així consti i tingui els efectes oportuns, signem aquest document.

Prof. Dra. Sílvia Osuna Oliveras

Dr. Javier Iglesias-Fernandez

Girona, 26/06/2024

Acknowledgements

The projects included in this thesis have been performed thanks to the financial support of the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC-2015-StG-679001, ERC-2022-POC-101112805) and the Human Frontier Science Program (HFSP, RGP0054/2020).

I would like to start by giving a great acknowledgment to my supervisor, Sílvia, and co-supervisor, Javi, for introducing me to the world of biocatalysis. I would also like to thank researchers, PhD students, and other coworkers I collaborated with in the elaboration of this thesis. Thank you for the opportunity to participate and take part in the discussion of all these fascinating projects. In particular, I want to thank Prof. Romas Kazlauskas and Colin Pierce for many extended and enriching discussions. Thanks also to the members and staff of the IQCC.

Also, I would like to thank Prof. Olsson and his team for hosting me during my research stay at the Chalmers University of Technology, Sweden. Finally, thanks also to Prof. Sterner and his lab team for such a nice stay at the University of Regensburg, Germany.

Contents

List of Publications	
List of Abbreviations	
Acknowledgements	
List of Figures	1
Summary	2
Resum	4
Resumen	6
1 Introduction	9
1.1 Enzymes: Nature's Catalysts and Beyond	10
1.1.1 Enzyme structure, function, and dynamics	12
1.1.2 Tryptophan synthase	16
1.1.3 α/β -hydrolase fold enzymes from Plant: Hydroxynitrile Lyase and Arylesterase	17
1.2 In silico methods for enzymes	20
1.2.1 Molecular Mechanics and Force Fields	21
1.2.2 Ligand Parametrization	22
1.2.3 Computational Enzyme Design Approaches	23
1.2.4 Shortest Path Map (SPM) Tool	24
1.2.5 Protein folding and the success of AlphaFold2	26
2 Objectives	31
3 The shortest path method (SPM) webserver for computational enzyme design	33
4 AlphaFold2 and Deep Learning for Estimating Conformational Heterogeneity and Designing Proteins: The Case of Tryptophan Synthase	43

5	Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase	67
6	Results and Discussion	97
6.1	SPM Webserver for Computational Enzyme Design	98
6.2	AlphaFold2 and Deep Learning for Estimating Conformational Heterogeneity and Designing Proteins: The Case of Tryptophan Synthase	100
6.3	Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase	105
7	Conclusions	111
	References	115
	Appendix	125
	Supporting Information of Chapter 4	125
	Supporting Information of Chapter 5	143

List of Figures

1.1	Representation of the reaction coordinate with and without a catalyst.	12
1.2	Timescale representation of different protein motions.	14
1.3	Representation of free energy landscape concept and the population shift	15
1.4	Overlay of the different tryptophan synthase B (TrpB) X-rays.	16
1.5	X-rays of (A) <i>HbHNL</i> and (B) SABP2 active sites.	18
1.6	Illustration of force field components in molecular dynamics.	21
1.7	Shortest Path Map (SPM) construction workflow.	25
1.8	Overview of different strategies developed for predicting different conformational states with AlphaFold2.	29
6.1	Main interface of the SPM Webserver.	99
6.2	Template-based AF2 approach diagram.	102
6.3	Representation of the previously reconstructed FEL of the 0B2- <i>pf</i> TrpB variant.	104
6.4	Shortest Path Maps of HNL3V and SABP2.	106
6.5	Oxyanion hole regeneration histogram and catalytic triad conformational states.	108
6.6	Cooperative interactions among mutations that enhance catalytic activity.	109

Summary

Enzymes, as biocatalysts, have evolved to catalyze biochemical reactions efficiently under mild conditions, yet their utility in industry is often limited by their natural substrate specificity and reaction scope. The enzyme modification for achieving a function of interest in specific conditions is not a straightforward task. Over the last decades, experimental techniques such as Directed Evolution (DE) have given hope to start obtaining industrially-relevant enzymes. However, DE results lack a rational explanation of why a variant improved towards a certain trait. This together with the cost and the time required, are considered the major drawbacks of DE techniques.

Computational techniques emerged as a new avenue of possibilities to accelerate the end-to-end process of enzyme design. Previous studies have computationally shown the importance of rationally understanding enzyme catalytic mechanisms and dynamics to surpass the limitations of non-rational procedures. These include static concepts like transition state stabilization, to understand how enzymes accelerate reactions, or dynamic concepts like the recovery of the Free Energy Landscape (FEL) to gain insights into the conformational heterogeneity of enzymes. In this regard, recent advances in Molecular Dynamics (MD) simulations allow the estimation of conformational ensembles of enzymes. However, escaping from the global energy conformation is sometimes difficult and expensive. Enhanced sampling techniques like accelerated MD or metadynamics are great alternatives to achieve a better exploration of conformational heterogeneity. Nonetheless, these methods often start from a predefined structure, which can make it difficult to fully explore significantly different conformational states that are separated by large energy barriers. Once enough sampling is obtained, translating these dynamics to meaningful low-dimension components, which can be geometric information between relevant positions, is not an easy task. Some dimensionality reduction techniques, like Principal Component Analysis (PCA) or Time-lagged Independent Component Analysis (TICA), have successfully helped to solve the intricate protein dynamics, yet they do not identify which positions are connected to the identified conformational change.

In this thesis, we develop new computational strategies for the exploration of enzymes' conformational landscape and the identification of mutational hotspots for the design of enzymes with new functions. Herein, in **Chapter 3**, we present the *Shortest Path Map* (SPM) webserver to open access to the academic community of our in-house tool for assessing the relevant dynamic connections of the system. The user-friendly interface allows easy modification of the only two parameters that can modify the results. We present in **Chapter 4** a complete review of advances in Deep Learning (DL)

methods and AlphaFold2 (AF2) for protein design followed by a new pipeline for conformation structure prediction named template-based AF2 approach, where multiple conformations can be predicted by altering the amount of co-evolutionary information coupled with a specific 3D structure used as a template. In this regard, these structures can accelerate the sampling of FEL and recover protein conformational heterogeneity. In **Chapter 5**, we show the conversion of hydroxynitrile lyase (HNL) to an efficient esterase (EST) enzyme through targeted mutations suggested by the SPM results, followed by a rational explanation of the improved variants. Finally, a brief discussion of the findings from each article is given in **Chapter 6**, and the primary conclusions derived from this thesis are provided in **Chapter 7**.

Resum

Els enzims, com a biocatalitzadors, han evolucionat per catalitzar reaccions bioquímiques eficientment sota condicions suaus, però la seva utilitat en la indústria sovint es veu limitada per la seva especificitat de substrat natural i rang de reaccions possibles. La modificació enzimàtica per aconseguir una funció d'interès en condicions específiques no és una tasca directa. Durant les últimes dècades, tècniques experimentals com l'Evolució Dirigida (DE, per les sigles en anglès) han donat esperança per començar a obtenir enzims rellevants per a la indústria. No obstant això, els resultats de DE manquen d'una explicació racional de per què una variant millora cap a un determinat tret. Això, juntament amb el cost i el temps requerits, es consideren els principals inconvenients de les tècniques de DE.

Les tècniques computacionals han sorgit com una nova via de possibilitats per accelerar el procés de disseny d'enzims de principi a fi. Estudis previs han demostrat computacionalment la importància de comprendre racionalment els mecanismes catalítics i la dinàmica dels enzims per superar les limitacions dels procediments no racionals. Aquests inclouen conceptes estàtics com l'estabilització de l'estat de transició, per comprendre com els enzims acceleren les reaccions, o conceptes dinàmics com la recuperació de la Superfície d'Energia Lliure (FEL, per les sigles en anglès) per aconseguir informació sobre l'heterogeneïtat conformacional dels enzims. En aquest sentit, els recents avanços en simulacions de Dinàmica Molecular (MD, per les sigles en anglès) permeten l'estimació de conjunts conformacionals d'estructures d'enzims. No obstant això, escapar de la conformació d'energia global a vegades és difícil i costós. Tècniques de mostreig millorades com la MD accelerada o *metadynamics* són excel·lents alternatives per assolir una millor exploració de l'heterogeneïtat conformacional. Tanmateix, aquests mètodes sovint parteixen d'una estructura predefinida, la qual cosa pot dificultar l'exploració completa d'estats conformacionals significativament diferents que estan separats per grans barreres energètiques. Un cop s'obté prou mostreig, traduir aquesta dinàmica en components amb un reduït nombre de dimensions significatius, que poden ser informació geomètrica entre posicions rellevants, no és una tasca fàcil. Algunes tècniques de reducció de dimensionalitat, com l'Anàlisi de Components Principals (PCA, per les sigles en anglès) o l'Anàlisi de Components Independents amb Retard en el Temps (TICA, per les sigles en anglès), han ajudat amb èxit a resoldre la intrincada dinàmica de les proteïnes, no obstant això, no identifiquen quines posicions estan connectades al canvi conformacional identificat.

En aquesta tesi, desenvolupem noves estratègies computacionals per a l'exploració del superfície conformacional dels enzims i la identificació de posicions interessants per a ser mutades per al disseny d'enzims amb noves

funcions. En aquesta tesi, al **Capítol 3**, presentem el servidor web *Shortest Path Map* (SPM) per donar accés obert a la comunitat acadèmica de la nostra eina per avaluar les connexions dinàmiques rellevants del sistema. La interfície d'usuari permet la modificació d'una manera fàcil dels únics dos paràmetres que poden modificar els resultats. En el **Capítol 4** presentem una revisió complerta dels avenços en els mètodes d'aprenentatge profund (DL, per les sigles en anglès) y Alphafold2 (AF2) pel disseny de proteïnes seguit d'un nou mètode per a la predicció de conformacions de proteïnes anomenat *template-based AF2 approach*, on es poden predir múltiples conformacions modificant la quantitat d'informació de coevolució juntament amb una estructura 3D específica utilitzada com a plantilla. En aquest sentit, aquestes estructures poden accelerar el mostreig de la FEL i recuperar l'heterogeneïtat conformacional de la proteïna. Al **Capítol 5**, mostrem la conversió de la hidroxinitril liasa (HNL) en una enzim esterasa (EST) eficient a través de mutacions dirigides suggerides pels resultats de SPM, seguida d'una explicació racional de les variants millorades. Finalment al **Capítol 6**, es presenta una breu discussió dels resultats de cada article, i es proporcionen les principals conclusions derivades d'aquesta tesi al **Capítol 7**.

Resumen

Las enzimas, como biocatalizadores, han evolucionado para catalizar reacciones bioquímicas eficientemente bajo condiciones suaves, sin embargo, su utilidad en la industria a menudo se ve limitada por su especificidad por el sustrato natural y rango de reacciones posible. La modificación de enzimas para lograr una función de interés en condiciones específicas no es una tarea fácil. En las últimas décadas, técnicas experimentales como la Evolución Dirigida (DE, por sus siglas en inglés) han dado esperanza para comenzar a obtener enzimas relevantes para la industria. Sin embargo, los resultados de la DE carecen de una explicación racional de por qué una variante mejora hacia un cierto rasgo. Esto, junto con el coste y el tiempo requeridos, se consideran los principales inconvenientes de las técnicas de DE.

Las técnicas computacionales han surgido como una nueva vía de posibilidades para acelerar el proceso de diseño de enzimas de principio a fin. Estudios previos han demostrado computacionalmente la importancia de comprender racionalmente los mecanismos catalíticos y la dinámica de las enzimas para superar las limitaciones de los procedimientos no racionales. Estos incluyen conceptos estáticos como la estabilización del estado de transición, para entender cómo las enzimas aceleran las reacciones, o conceptos dinámicos como la recuperación de la Superficie de Energía Libre (FEL, por sus siglas en inglés) para obtener información sobre la heterogeneidad conformacional de las enzimas. En este sentido, los avances recientes en simulaciones de Dinámica Molecular (MD, por sus siglas en inglés) permiten la estimación de conjuntos conformacionales de estructuras de enzimas. Sin embargo, escapar de la conformación de energía global a veces es difícil y costoso. Técnicas de muestreo mejorado como la MD acelerada o *metadynamics* son excelentes alternativas para lograr una mejor exploración de la heterogeneidad conformacional. Sin embargo, estos métodos a menudo parten de una estructura predefinida, lo que puede dificultar la exploración completa de estados conformacionales significativamente diferentes que están separados por grandes barreras energéticas. Una vez que se obtiene suficiente muestreo, traducir esta dinámica en componentes con un reducido número de dimensiones significativas, que pueden ser información geométrica entre posiciones relevantes, no es una tarea fácil. Algunas técnicas de reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA, por sus siglas en inglés) o el Análisis de Componentes Independientes con Retardo en el Tiempo (TICA, por sus siglas en inglés), han ayudado con éxito a resolver la intrincada dinámica de las proteínas, sin embargo, no identifican qué posiciones están conectadas al cambio conformacional identificado.

En esta tesis, desarrollamos nuevas estrategias computacionales para la exploración de la superficie conformacional de las enzimas y la identificación

de posiciones interesantes para ser mutadas para el diseño de enzimas con nuevas funciones. En esta tesis, en el **Capítulo 3**, presentamos el servidor web *Shortest Path Map* (SPM) para dar acceso abierto a la comunidad académica de nuestra herramienta para evaluar las conexiones dinámicas relevantes del sistema. La interfaz de usuario permite la modificación de una manera fácil de los únicos dos parámetros que pueden modificar los resultados. En el **Capítulo 4** presentamos una revisión completa de los avances en los métodos de aprendizaje profundo (DL, por sus siglas en inglés) y AlphaFold2 (AF2) para el diseño de proteínas seguido de un nuevo método para la predicción de conformaciones de proteínas llamado *template-based AF2 approach*, donde múltiples conformaciones pueden ser predichas modificando la cantidad de información de coevolución junto con una estructura 3D específica utilizada como plantilla. En este sentido, estas estructuras pueden acelerar el muestreo de la FEL y recuperar la heterogeneidad conformacional de la proteína. En el **Capítulo 5**, mostramos la conversión de la hidroxinitrilo liasa (HNL) en una enzima esterasa (EST) eficiente a través de mutaciones dirigidas sugeridas por los resultados de SPM, seguida de una explicación racional de las variantes mejoradas. Finalmente, en el **Capítulo 6**, se presenta una breve discusión de los resultados de cada artículo, y se proporcionan las principales conclusiones derivadas de esta tesis en el **Capítulo 7**.

Chapter 1:

Introduction

1.1 Enzymes: Nature's Catalysts and Beyond

Enzymes are the proteins chosen as catalysts for our existence. These biological catalysts are capable of performing a certain reaction efficiently; some can accelerate chemical reaction rates by more than a million times, while others perform in a highly enantioselective manner. These features, among others, make them essential for life. Unlike non-biological catalysts, enzymes can operate under mild conditions of temperature, pH, pressure, and solvent (*i.e.*, aqueous), making them the most environmentally friendly catalysts. Enzymes display incredible specificity, rooted in their ability to recognize and bind to specific substrates. This specificity contributes to their ability to minimize cross-reactivity, which is often a challenge with synthetic catalysts, leading to cleaner reactions.

Enzymes are crucial for regulating metabolic pathways such as glycolysis and the citric acid cycle, which are essential for cell energy production. For example, the enzyme hexokinase catalyzes the rate-limiting step of glycolysis and is regulated by its product, glucose-6-phosphate.¹

The study of enzymes has facilitated a deeper understanding of evolutionary biology. Enzymes evolve to meet the organism's needs, which can be seen in the variety of enzymes in different species and environments. This adaptability is a testament to the evolutionary success of enzymes and provides a rich source of information for studying the principles of natural selection and adaptation. For example, the work of Pinney et al. demonstrated that thermal adaptability can be achieved simply by altering the catalytic hydrogen bond donor (HBD) from D103 to S103 plus water in mesophilic to thermophilic variants of the enzyme ketosteroid isomerase (KSI). They highlighted the trade-off in both activity and stability of having a strong HBD and a lower pK_a of one of the key catalytic amino acids in mesophilic KSI for high activity and low stability. The opposite was found for thermophilic variants.²

The importance of enzymes is further highlighted by their application across numerous fields, ranging from industrial processes to medicine. Within industry, enzymes have transformed processes by offering greener, more efficient, and safer alternatives to traditional chemical catalysts. Their use in synthesizing biofuels, processing food and beverages, improving coated paper quality, or waste management exemplifies their utility in green chemistry, where the focus is on reducing environmental impact.³⁻⁸ Moreover, enzymes' ability to function in aqueous environments minimizes the dependence on harmful organic solvents and aligns with the principles of green chemistry, contributing to sustainable industrial processes. In medicine, enzymes are utilized for diagnostic purposes, such as the early pregnancy detection methods employing Enzyme-Linked ImmunoSorbent Assay (ELISA),

where horseradish peroxidase functions as the enzyme label for detecting human chorionic gonadotropin.^{9–11} Another example is being used as treatment options in the form of drug targets or therapeutic agents is enzyme replacement therapy in Pompe disease (*i.e.*, glycogen storage disease type II), an inherited lysosomal disease caused by a deficiency of the enzyme acid alpha-glucosidase. Since EU approval in 2023, patients can take an intravenous infusion of the cipaglucosidase alfa enzyme (PombilitiTM).¹²

The research and innovation surrounding enzymes continue to broaden their applications beyond natural biochemical pathways, enabling reactions and processing of previously untouchable substrates.¹³ Recent work by Sarai et al. demonstrates the ability of an evolved enzyme to cleave silicon-carbon bonds in volatile methyl siloxanes, exemplifying the potential of enzymes to address non-natural reactions for non-natural substrates.¹⁴

This advancement aligns with the rapidly growing field of protein design, which benefits greatly from novel protein engineering methods. On the computational side, breakthroughs such as AlphaFold2 (AF2), RoseTTAFold, or trRosetta, which predicts protein structures with high accuracy using deep learning techniques, have revolutionized the field.^{15–17} Concurrently, the laboratory method of directed evolution (DE), awarded a Nobel Prize in 2018, allowed the creation of novel proteins with specific functions by emulating natural selection. This method iteratively selects the best variants across multiple rounds, optimizing the desired properties to achieve the fittest protein.¹⁸ These developments have enabled the creation of tailor-made enzymes that can efficiently catalyze a wide range of chemical reactions, pushing the boundaries of traditional enzyme applications in organic synthesis and offering new avenues for biocatalysis.^{19,20}

The amazing source of catalysis-related knowledge extracted from enzymes has helped us understand biological processes better and opened the door for bioinspired catalysis. By creating synthetic catalysts that closely resemble the active site and the confined space of the enzyme, bioinspired catalysis aims to imitate the high selectivity and efficiency of enzymatic processes under synthetically relevant conditions. Inspired by the mechanisms of metalloenzymes, the development and application of catalysts that are made from earth-abundant metals, such as manganese (Mn) and Iron (Fe), in asymmetric oxidation reactions using benign oxidants such as hydrogen peroxide (H₂O₂) and molecular oxygen (O₂) represent significant progress toward achieving sustainable chemical catalysis in the bioinspired field.^{21–23}

This Ph.D. thesis explores innovative computational methods to unravel the complexities of confined space catalysis inherent in enzymatic catalysis. Thus aiming to speed up the development of new environmentally friendly biocatalysts.

1.1.1 Enzyme structure, function, and dynamics

As proteins, the structure of enzymes is composed of a linear chain of amino acids. Each amino acid, the building block of proteins, consists of a central carbon atom ($C\alpha$) bonded to an amino group (NH_2), a carboxyl group ($COOH$), a hydrogen atom, and an R-group (*i.e.*, side chain). That defines the characteristics and role within the protein of the 20 canonical amino acids encoded by the genetic code. The sequential arrangement of these building blocks is the primary structure. The local spatial arrangement of the sequence of amino acids by their backbone, excluding the R-groups, through hydrogen bonds is the secondary structure, where the most frequent elements are the α -helices and β -sheets and less frequent local foldings like loops or coils. Further folding brings these elements into a three-dimensional tertiary structure, where the side chains' interactions determine the protein's final shape. For some proteins, a higher-order assembly between multiple polypeptide chains (subunits) is necessary for their function, which forms the quaternary structure.²⁴

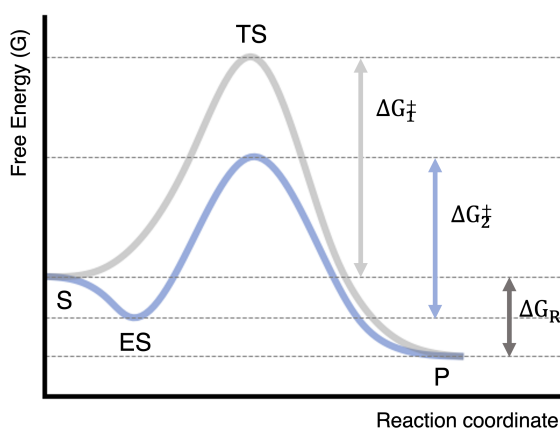


Figure 1.1: **Representation of a reaction coordinate diagram in the absence (grey) and in the presence (blue) of an enzyme as a catalyst.** The plot exemplifies the free energy as a function of the chemical reaction course, which starts from the substrate (S) and involves a transition state (TS) to form the final product (P). The enzyme-catalyzed reaction also involves the enzyme-substrate (ES) complex. The activation energy in the absence (ΔG_1^\ddagger) and in the presence (ΔG_2^\ddagger) of enzymes is also shown, as well as the energy difference between the S and P (ΔG_R).

The precise folding of enzymes is what determines their function. Their activity will be performed in a specific confined space of the protein, the active site. The amino acid architecture of the active site is exquisitely

arranged to complement the substrate’s structure, allowing for a highly selective interaction. This structural and chemical complementarity creates an environment optimized to recognize, bind, and catalyze their respective reactions with remarkable specificity and efficiency. These residues can serve as donors or acceptors of electrons, participate in hydrogen bonding, or contribute to the hydrophobic environment, guiding the substrate into a transition state (TS) to lower the Gibbs free energy (G) barrier required for the reaction (ΔG^\ddagger), which is what defines how a catalyst acts (*i.e.*, reducing ΔG^\ddagger compared to the reaction occurring without the catalyst). The ΔG^\ddagger is defined as the difference in Gibbs free energy between the TS and the reactants. Therefore, catalysts do not change the free-energy difference between reactants and products (ΔG_R), therefore their role in reducing the TS Gibbs free energy (Fig. 1.1).

In this regard, in the mid-20th century, Linus Pauling proposed that enzymes accelerate reactions based on the high specificity of the TS structure compared to the Michaelis complex (MC).²⁵ Continuing this idea, Warshell and co-workers then established that enzymes’ mechanism to lower ΔG^\ddagger is with their highly preorganized environment. These active site pockets are precisely reassembled, orienting the catalytic residues in an optimal arrangement for transition state stabilization (TSS). The concept of enzyme active site preorganization involves not only geometrical descriptors but also electrostatic complementarity to the transition state through strategic alignment of electric fields within the enzyme structure, even in the absence of the reactant. This preorganized structural and electronic configuration mirrors the TS’s properties, minimizing the reorganization energy required upon substrate binding and thus reducing the enthalpy and entropy of the ΔG^\ddagger for the reaction.^{26–29} To exemplify this electric field reorganization, Boxer and co-workers demonstrated through vibrational spectroscopy on KSI and its mutants that an activation barrier’s increase was linearly correlated with a decrease in electric field strength.^{30–32} In this regard, as stated by Wolfenden, TS analogs must bind more tightly to the enzyme’s active site than the natural substrates, proportionate to the level of catalytic rate enhancement observed with the enzyme.³³ For example, this is validated for the human enzyme hypoxanthine-guanine phosphoribosyltransferases (HGPRTs), which are strongly inhibited using TS analogs, showing more than a 1,000-fold tighter binding compared to the binding of the natural nucleotide substrates.^{34–36}

While the concept of enzyme preorganization suggests a certain rigidity or static character that can match in some sense the lock-and-key model (*i.e.*, precise and static shape complementarity between enzyme and substrate)³⁷ adding the ingredient of the preorganized state for the TS, enzymes are in fact, dynamic entities. They exist in a landscape of microstates (*i.e.*, conformations) of different stabilities influenced by conditions such as temperature,

pH, and substrate presence, among others. The enzyme's dynamic nature fosters its versatility and adaptability to facilitate a wide range of biochemical processes. This flexibility enables enzymes to accept different substrates and catalyze diverse reaction types apart from those specifically evolved. This is termed substrate and catalytic promiscuity, a property highly present in ancestral enzymes, which makes them generalist enzymes.^{38,39} On the contrary, enzymes with high substrate specificity are termed specialists.⁴⁰ Due to this inherent flexibility, the lock-and-key view evolved to new models of interaction, starting with the induced-fit model (*i.e.*, substrate binding alters enzyme shape for catalysis),⁴¹ and finally evolving to the conformational selection model. This last model illustrates how enzymes exist in various conformations before interacting with a substrate. This model suggests the substrate binds to more energetically favorable conformations, thereby selecting the enzyme state that best facilitates catalysis.⁴²

This conformational dynamism can be reflected in the so-called free energy landscape (FEL), where all thermally accessible conformations are represented, and thus, the thermodynamic differences and kinetic barriers that separate them are shown. The FEL is based on the free energy calculation obtained from the negative logarithm of the conformation population's distribution in $k_B T$ units, where k_B is the Boltzmann constant and T is the temperature value. So, highly populated microstates will be the most stable energy minima and small energy barriers will be defined by fast interconversions. The timescales for the conformational transitions can range from bond vibrations at the femtosecond timescale, to side chain rotation or loop motions in the range of picosecond (ps) to nanosecond (ns) timescale, followed by protein folding in the range of microsecond (μs) to seconds (Fig. 1.2).

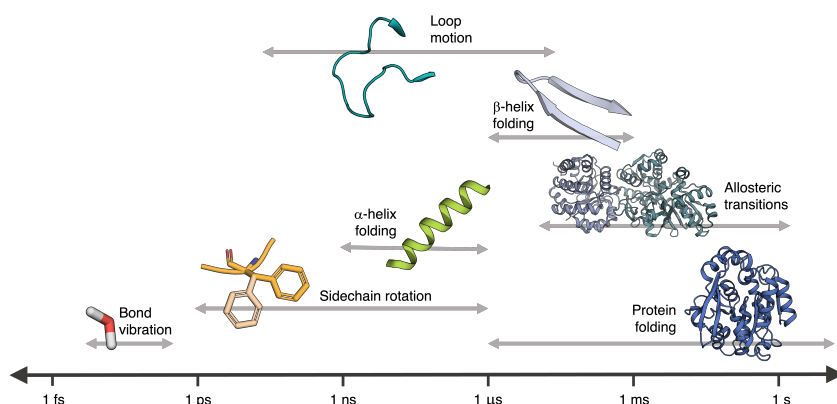


Figure 1.2: **Timescale representation of different protein motions.** These go from bond vibration up to protein folding.

For some enzymes, the activity can be controlled through various mechanisms, including post-translational modifications, allosteric modulation, and interactions with cofactors or other proteins. These regulatory mechanisms allow the enzyme to respond dynamically to changes in the cellular environment, adjusting its activity to meet the cell’s metabolic needs. For instance, allosteric modulation involves binding a molecule at a site other than the active site, inducing a conformational change that affects enzyme activity, or, in the case of cofactors, organic molecules like vitamins or metal ions, that are essential for the catalytic activity of some enzymes, assisting in the reaction by stabilizing the transition state or acting as electron donors or acceptors. All these mechanisms will fine-tune the corresponding enzyme FEL, making more or less thermodynamically accessible some conformations compared to others.¹⁹ In this regard, the cell can regulate its enzyme activities and, thus, its available resources.

Similarly, enzyme designers (*i.e.*, scientists specialized in the creation and modification of enzymes for specific needs) can leverage the FEL concept to selectively mutate and choose those enzyme variants that make microstates more thermodynamically accessible, reassembling to an optimized, preorganized enzyme state. In this regard, by incorporating insights from ancestral enzymes, known for their high substrate and catalytic promiscuity, structural flexibility, and stability, enzyme designers can benefit from these enzymes and the FEL concept to guide and create variants improving the stabilization of the biochemically relevant microstates.^{19,43} This is represented in Fig. 1.3 where mutations make the populations shift towards other conformational states. It is worth mentioning that although reconstructed ancestral enzymes offer an excellent starting structure for enzyme design campaigns, their advantages come at the cost of low catalytic efficiency compared to modern enzymes.⁴⁰

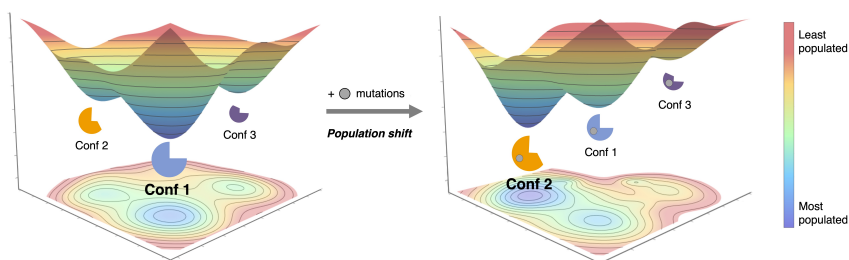


Figure 1.3: Representation of an enzyme’s free energy landscape (FEL) and the population shift induced by mutations. The relative populations of each conformation are represented by its size. Mutations are indicated as grey circles.

Leveraging the introduced concepts, such as the FEL theory and insights from generalist enzymes, the following subsections outline the specific enzymes this thesis investigates. These systems are introduced, highlighting their application and importance.

1.1.2 Tryptophan synthase

Tryptophan synthase is a heterodimeric enzyme complex comprising two α -subunits (TrpA) and two β -subunits (TrpB). This enzyme catalyzes the final production of L-tryptophan (L-Trp) and is governed by an allosteric communication pathway between subunits. The overall reaction starts at TrpA, where the retro-aldol cleavage of indole glycerol phosphate produces glyceraldehyde 3-phosphate and indole. The latter is diffused to the active site of TrpB through an inter-subunit channel, highlighting the key role of allosteric communication between subunits. TrpB is a pyridoxal-5'-phosphate (PLP) (*i.e.*, the metabolically active form of Vitamin B6 that acts as a cofactor in the TrpB enzyme)-dependent enzyme that catalyzes the condensation of L-Serine (L-Ser) with indole through a multistep reaction involving many intermediates to produce the final L-Trp.

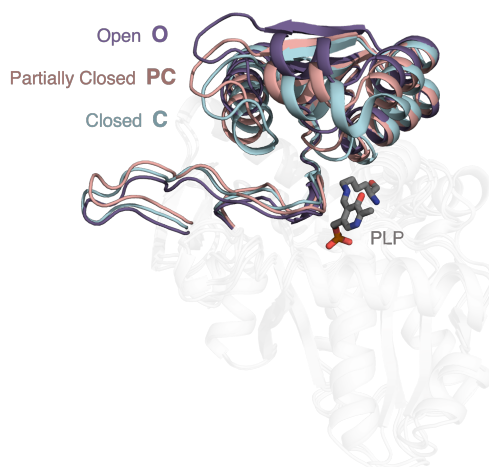


Figure 1.4: **Overlay of the different tryptophan synthase B (TrpB) X-rays.** The different COMM domain conformational states are highlighted: O highlighted in violet (PDB ID: 1WDW), PC in brown (PDB ID: 5DW0), and C in blue (PDB ID: 3CEP). Pyridoxal-5'-phosphate (PLP) cofactor is shown in grey.

The communication (COMM) domain covers the active site of TrpB, whose conformational dynamics, described by exposing the active site in three different degrees (Fig. 1.4)), have been studied and are related to the

catalytic efficiency of TrpB. Arnold and coworkers applied DE to the wild-type (WT) *Pyrococcus furiosus* TrpB (*Pf*TrpB) that operates inefficiently in the absence of its partner, the TrpA subunit, due to restricted conformational heterogeneity. It was found that the laboratory-evolved 0B2-*Pf*TrpB can explore closed, partially closed, and open conformations of the COMM domain, which translates to an increase in its stand-alone activity.⁴⁴ In this line, Osuna, Sterner, and coworkers rationally designed SPM6-TrpB, an ancestral ANC3-TrpB enzyme variant that also improved their stand-alone activity.^{45,46} It is worth mentioning that this design was based on the already stand-alone Last Bacterial Common Ancestor (LBCA) TrpB, a reconstructed ancestral enzyme with high catalytic efficiency, something unexpected for an ancestral enzyme, therefore a generalist enzyme, lower specific activities compared to specialist enzymes may be expected.^{45,47}

The interest in TrpB enzymes can be endorsed for having been deeply studied for more than 6 decades.⁴⁸ Due to their highly selective carbon-carbon bond-forming reaction, many engineering efforts have been applied to the TrpB enzyme to perform new β -substitutions reactions with a wide array of C-nucleophiles, including indole derivatives, ketone-derived enolates, and nitroalkenes.^{49–51}

In the context of this thesis, the conformational dynamics of the COMM domain have been studied for the previously mentioned systems (*i.e.*, *Pf*TrpB, 0B2-*Pf*TrpB, LBCA TrpB, and SPM6-TrpB) using a new protocol based on AF2 (see chapter 4).

1.1.3 α/β -hydrolase fold enzymes from Plant: Hydroxynitrile Lyase and Arylesterase

α/β -hydrolase fold superfamily comprises one of the largest groups of structurally related enzymes that exhibit diverse catalytic functions. Some enzymes in this family include arylesterase, acetylcholinesterase, lipase, thioesterase, serine carboxypeptidase, haloalkane dehalogenase, epoxide hydrolase, amidase, and hydroxynitrile lyase, among others.⁵² They share a common structure characterized by a core α/β -fold composed of a central β -sheet flanked by α -helices. This fold contains a conserved catalytic triad, typically composed of a nucleophile, a histidine, an acid (either aspartate or glutamate), and an oxyanion hole formed by two or three amino acids. The nucleophile is located in a sharp turn between a β -strand and an α -helix named nucleophilic elbow, which is identified by the consensus sequence Sm-X-Nu-X-Sm (*i.e.*, Sm = small residue, X = any residue, and Nu = nucleophile)⁵³ In addition to this catalytic core, these enzymes often contain other structural domains, like a lid or a cap. In some publications, these extra domains are referred to as the enzyme’s lid, or in some cases, the

1.1 Enzymes: Nature's Catalysts and Beyond

flap. However, it is also accepted to differentiate them by the immobile or mobile module covering the active site, referring to the cap or lid domains, respectively.^{54–56}

The versatility of this superfamily has been shown to have great potential in a wide array of industrially relevant applications. Among these, microbial lipase is a prime example, with an estimated market of about USD 425.0 Million in 2018. Its industrial applications extend to biodiesel production, food and drink processing, leather and textile treatment, detergent formulation, pharmaceuticals, and medical applications.⁸ Additionally, members of this superfamily are promising candidates for industrial applications in the degradation of polyethylene terephthalate (PET), a common plastic. However, industrial-scale enzyme application for plastic recycling faces significant challenges, despite advancements in enzyme optimization and a need for continued research and development.⁵⁷

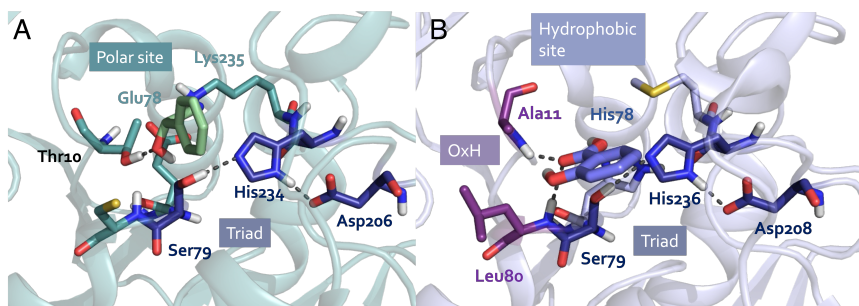


Figure 1.5: X-rays of (A) *HbHNL* and (B) *SABP2* active sites (PDB IDs: 1YB6 and 1Y7I, respectively). For both cases, important residues are shown in sticks and the shared catalytic triad is highlighted in deep blue. *SABP2* has a *HbHNL* has a hydrophobic site and an oxyanion hole (OxH, highlighted in purple), whereas *HbHNL* has a polar site and lacks an OxH. *HbHNL* substrate mandelonitrile is shown in green, and *SABP2* product salicylic acid is colored in lilac.

Within this thesis, significant attention will be given to the hydroxynitrile lyase enzyme from the rubber tree (*Hevea brasiliensis*, *HbHNL*) and the arylesterase enzyme salicylic acid binding protein 2 (*SABP2*) from the tobacco plant (*Nicotiana tabacum*).^{58–60} These enzymes play an important role in the organism's defense mechanism or metabolism. *HbHNL* enzyme catalyzes the elimination of hydrogen cyanide from acetone cyanohydrin to defend against herbivorous insects and microbial attack, or as a nitrogen source.^{61–63} *SABP2* enzyme catalyzes the arylesterase hydrolysis of the inactive methyl salicylate (MeSA) into the active salicylic acid (SA) and methanol, resulting in the activation of the SA-dependent defense signaling pathway.^{64,65}

In a recent study by Kazlauskas and coworkers, the phylogenetic tree that connects these two enzymes was reconstructed, and thus, the probable ancestral branched enzymes were located. They highlighted the ancestral enzymes’ generalist character, specifically the catalytic promiscuity of these ancestral enzymes compared to modern specialist enzymes. Those reconstructed ancient enzymes presented the main function of ester hydrolysis, showing how the defensive cyanogenesis produced from the hydroxynitrile lyase (HNL) reaction evolved from esterase (EST) enzymes. The different evolutionary pressures made SABP2 and *HbHNL* enzymes share a sequence identity of 44%, yet between both share a common fold and catalytic triad, the HNL and EST mechanisms are quite different.⁶⁶ SABP2’s reaction can be split into two parts described by the formation of the stable acyl intermediate, and deacylation. To accomplish this, the enzyme needs an oxyanion hole, created by two backbone amide groups, to stabilize the carbonyl oxygen’s anion formed in the two subsequent tetrahedral intermediates, and the catalytic triad serine acts as a nucleophile to attack the substrate’s carbonyl carbon. Compared to EST enzymes, *HbHNL*’s reaction proceeds without any stable intermediate and, therefore, can be considered a one-step reaction. In this HNL reaction, the EST oxyanion hole’s positions are not used, instead, a side-chain threonine hydroxyl group needs to activate and stabilize the substrate’s hydroxyl group, and a lysine, assisted with glutamate, must stabilize the formed cyanide anion in what is called the polar site of the active site. Contrary to this polar site, HNL enzymes have a hydrophobic or nonpolar site composed of histidine or phenylalanine (Fig. 1.5).⁶⁷

Based on the knowledge gained in the previous study, Kazlauskas and coworkers proposed to mutate the catalytic obvious positions from the *HbHNL* enzyme to the ones present in the specialized esterase SABP2 enzyme, thus creating a *HbHNL* variant (*HbHNL*-EST) with improved esterase activity. The mutations consisted of Thr11Gly, to regenerate the oxyanion hole in orientation and space, and Glu79His-Lys236Met, to regenerate from the polar to the hydrophobic site. Compared to *HbHNL*, this new variant improved the catalytic efficiency hydrolyzing p-nitrophenyl acetate from $110 \text{ M}^{-1} \text{ min}^{-1}$ up to $4200 \text{ M}^{-1} \text{ min}^{-1}$, or $10100 \text{ M}^{-1} \text{ min}^{-1}$ if instead K236M there is K236G mutation, being far away from the SABP2 value of $86000 \text{ M}^{-1} \text{ min}^{-1}$. It is worth mentioning that this new variant nearly loses HNL functionality.⁶⁸ In a continuation of this study, they also tried to apply the three obvious mutations (*i.e.*, Thr11Gly-Glu79His-Lys236Gly, not copying the SABP2 amino acid Met236) to the transitional functionality EST to HNL enzyme ancestor HNL1, creating the HNL1-EST variant. As the HNL1 ancestor enzyme presents promiscuous esterase activity, the HNL1-EST has even higher catalytic efficiency in hydrolyzing p-nitrophenyl acetate, up to $12000 \text{ M}^{-1} \text{ min}^{-1}$, and retains some HNL promiscuous activity.⁶⁹

Although there is a relevant improvement in these HNL variants for the

EST reaction, the activities are far from those of the specialized esterase SABP2. This thesis will tackle this problem and show how we create *HbHNL*-EST variants that surpass SABP2 activity.

1.2 In silico methods for enzymes

The field of computational chemistry has been profoundly shaped by advances in our understanding of atomic and molecular dynamics (MD) over the past several decades. Central to this evolution has been the pioneering work of 2013 Nobel laureates Martin Karplus, Michael Levitt, and Arieh Warshel, whose developments in MD simulations and quantum mechanics/molecular mechanics (QM/MM) methods have significantly influenced the study of enzyme conformational dynamics and reactivity.

A key advancement was the development of the Consistent Force Field (CFF) by Shneior Lifson and later enhanced by Warshel and Levitt at the Weizmann Institute.⁷⁰ They coded, using the Fortran programming language, a program named CFF that let them compute the energy and the corresponding first and second derivatives of any molecular system using a simple potential energy function. This allowed for the simulation of molecular systems based on simple potential energy functions, a revolutionary step toward studying biological macromolecules.⁷¹

Building upon the CFF, Karplus conducted groundbreaking MD simulations in the late 1970s. This includes his landmark 1977 study on bovine pancreatic trypsin inhibitor (BPTI) with 9.2 ps of MD simulation in vacuum, which highlighted the dynamic nature of proteins beyond what static X-ray crystallography could reveal and the potential of MD simulations for unveiling the conformational flexibility of biomolecules.⁷²

Concurrently, Warshel and Levitt introduced the QM/MM hybrid method, which combined quantum and classical mechanics to model electronic interactions at enzymatic active sites, while efficiently managing the larger molecular structure through molecular mechanics (MM).⁷³ This approach has significantly enhanced our understanding of enzymatic processes, especially in how enzymes manage complex biochemical reactions.

Further developments by Warshel created the Empirical Valence Bond (EVB) model, which improved simulation accuracy by offering a simple general framework to model reactive processes through the coupling of multiple FFs.⁷⁴ Levitt and Sharon later demonstrated how using explicit solvents in BPTI simulation could achieve more realistic dynamics behaviors.⁷⁵

Advances in computational power, particularly through the use of GPUs, have facilitated the study of larger and more complex systems through

longer nanosecond-timescale MD simulations. The development of modern FF for proteins (*e.g.*, AMBER, CHARMM, GROMOS, and OPLS) and water models (*e.g.*, as TIP3P, TIP4P, and OPC) has enabled more precise simulations.

1.2.1 Molecular Mechanics and Force Fields

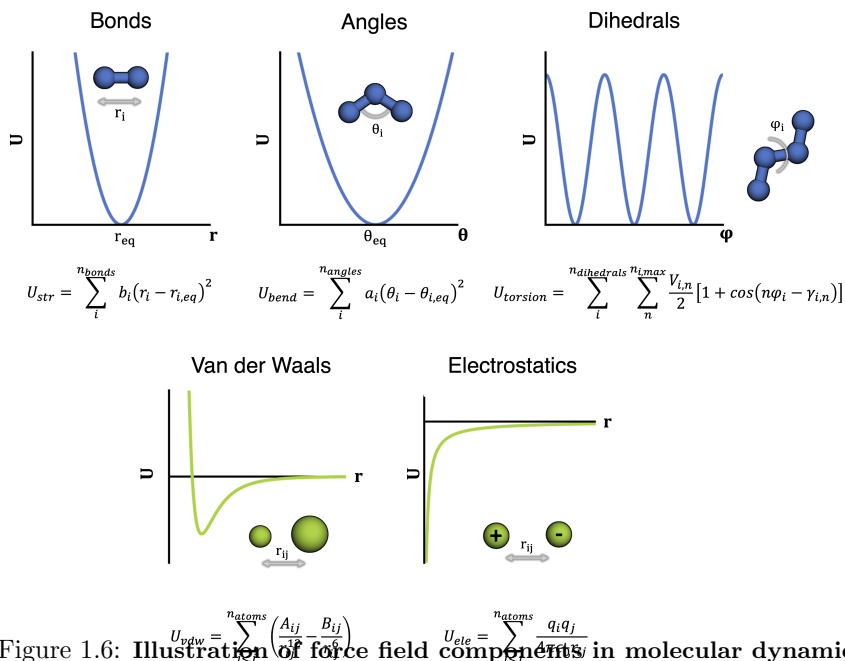


Figure 1.6: **Illustration of force field components in molecular dynamics.** The top diagrams in blue represent the bonded interactions, including bonds, angles, and dihedrals with their respective energy functions. The lower diagrams in green depict the nonbonded interactions, detailing Van der Waals forces and electrostatics, with their mathematical expressions.

Molecular mechanics (MM) simplifies the simulation of complex biological molecules such as proteins and enzymes, which may consist of thousands to millions of atoms. In MM, atoms are treated as classical particles, using a "ball and spring" model where each atom is depicted as a sphere, characterized by specific charges, masses, and radii. Bonds between atoms are modeled as springs, capturing the essence of molecular interactions without defining quantum mechanical details such as electron arrangement and bond formation.

In MM, the potential energy of molecules is derived from force fields (FFs), which are categorized based on complexity. Class 1 FFs, such as the

AMBER99 protein-specific FFs and updates like ff14SB and ff19SB,^{76,77} employ simpler, harmonic terms for bonded interactions (such as bonds, angles, and dihedrals) and Lennard-Jones potentials alongside Coulomb’s law for nonbonded interactions like van der Waals forces and electrostatics (Fig. 1.6). These FFs balance accuracy and computational efficiency, making them suitable for simulating the molecule’s FEL as it shifts from one state to another.

General FFs, like Generalized Amber Force Fields (GAFF)⁷⁸ and the updated version GAFF2, can cover a wider chemical space for accurately modeling small organic molecules that are frequently part of larger biomolecular systems. Class 2 and Class 3 FFs introduce advanced terms for anharmonic effects and polarization, respectively. For instance, Class 3 FFs like AMOEBA polarizable FF^{79,80} account for complex chemical effects such as polarization and are used for systems where a more detailed electron distribution model is necessary. Other FF alternatives are reactive FFs, like EVB created by Warshel,⁷⁴ or ML-FFs, that try to narrow the gap between the accuracy of ab initio methods and the efficiency of classical FFs.⁸¹

This thesis applies specifically parametrized protein FFs like ff14SB⁷⁶ with the TIP3P water model and its updated version ff19SB⁷⁷ paired with the OPC water model, to refine simulation accuracy. For small molecules, GAFF2 is employed. TIP3P⁸² is favored for its effective representation of bulk water properties, while OPC’s⁸³ four-point model improves the representation of hydrogen bonding and thermodynamic properties, thus providing a more refined approach with an extra computational cost.

Duran et al. demonstrated that the ff19SB+OPC combination better described the X-ray structures of the TrpB enzyme as compared to ff14SB+TIP3P, highlighting the importance of selecting appropriate force fields and water models to accurately depict biomolecular interactions.⁸⁴

1.2.2 Ligand Parametrization

Parameter computation is essential in MD simulations in which non-canonical amino acids and other molecules not covered in protein force fields like ff14SB or ff19SB are incorporated. The methodology that we followed in this thesis begins with a quantum mechanics (QM) optimization to determine the molecule’s ground state using the B3LYP (Becke, 3-parameter, Lee–Yang–Parr) functional,^{85,86} a hybrid generalized-gradient approximation (GGA) functional that incorporates 20% exact Hartree-Fock (HF) exchange. Additionally, Grimme’s dispersion correction with Becke-Johnson damping (D3-BJ) is applied to accurately model intramolecular dispersion effects. The polarizable conductor model (PCM),⁸⁷ utilizing dichloromethane, with

a dielectric constant of 8.9, as a solvent, estimates dielectric permittivity within the enzyme active site.⁸⁸ Following this, a single-point HF calculation captures the molecular electrostatic potential (ESP) grid. This grid, fitting to the QM electrostatic potential, eases computing partial charges via the two-stage restrained ESP (RESP) model using the antechamber package.⁸⁹⁻⁹³ All QM calculations performed employ the 6-31G(d) Pople basis set, chosen for its ability to deliver accurate coordinates and ESPs, computed with the Gaussian16 software package.⁹⁴ Lastly, the parmchk2 module from the antechamber package⁸⁹ assigns bond and angle parameters for unlisted parameters in the force field.

1.2.3 Computational Enzyme Design Approaches

In the rapidly evolving sector of enzyme design, two different methodologies have co-evolved to create new desired enzymes. From the experimental part, as previously highlighted, DE is the biggest breakthrough to close the gap to industrial needs. On the computational side, a wide range of methods have been developed to compete with the success of DE.

Focusing on computational approaches can be classified into three main categories based on their primary focus.¹⁹ The first category is composed of methods that utilize multiple sequence information or protein folds to obtain evolutionary-based insights into which positions have important roles in the desired feature (*i.e.*, function or stability). Examples include SigniSite, which uses MSA information to analyze the evolutionary conservation and variability of residues within a protein family,⁹⁵ or FuncLib, which obtains the evolution conservation scores creating position-specific scoring matrices (PSSM) coupled with Rosetta design calculations.^{96,97}

The second category focuses on the chemical steps of catalysis. One of the most successful methods in this category is the *inside-out* protocol, developed through the collaboration of the Baker and Houk research groups.⁹⁸ This method is based on the TSS concept, obtaining the ideal QM geometry of the minimum set of catalytic residues (*i.e.*, also known as *theozyme*).⁹⁹ This geometry is then transferred to an existing protein template using RosettaMatch, and further mutations are introduced in the active site with RosettaDesign.^{100,101} This method was later improved by including conformational ensembles generated through MD simulations of the designs, checking the deviation from the computed *theozyme* model.¹⁰² Building on this idea, the Janssen Lab, in collaboration with the Baker Lab, released the Catalytic Selectivity by Computational Design (CASCO) framework, which quantifies the conformations through MD simulation that matches a geometric criteria for near attack conformations (NACs) based on the TS structure.¹⁰³

The last category focuses on the conformational dynamics of the enzyme, not just the active site. This includes the keyhole-lock-key model, which tackles the importance of the substrate entrance and product release channels by understanding the effect on the tunnels that provide access from the surface to the active site.¹⁰⁴ Methods such as CAVER and AQUA-DUCT can be used to identify these tunnels and their positions.^{105,106} Additionally, loop engineering is known as a key factor for catalytic success, as modifications in loop regions can significantly enhance enzyme performance.¹⁰⁷

The next section presents our in-house approach that fits in the last category of methods, where the overall dynamics of the protein are key for finding the hot spots that affect catalysis.

1.2.4 Shortest Path Map (SPM) Tool

It could seem obvious that mutations around the catalytic cavity can have an important effect on enzyme performance, as was shown in the introduction for the evolutionary mutations in the KSI enzyme performance at different temperatures.² However, looking at this reduced space of the enzyme is insufficient to achieve high improvements in the desired activity. In this regard, distal mutations are of extreme importance for going beyond small gains in activity.¹⁰⁸ Distal mutations are defined as residues that go beyond the active site's first shell of residues. The relevance of distal mutations will increase if we go beyond similar substrates or functions, which our reference enzyme was not evolved to do. Thanks to techniques like DE, we can learn the implications of these distal mutations through each round of evolution for improving a desired function. For instance, DE enhanced the acyltransferase LovD enzyme for simvastatin production, a blockbuster cholesterol-lowering drug, achieving a 1000-fold increase in activity in the 9th DE round. Impressively, 18 of the 29 mutations introduced were located on the protein's surface.¹⁰⁹ Additionally, the significance of mutations in distal locations is noticeable in other DE applications, such as for broadening substrate scope in monoamine oxidase (MAO-N), converting *Pseudomonas diminuta* phosphotriesterase (PTE) into arylesterase (AE), and enhancing PET depolymerization capacities in *Ideonella sakaiensis*.^{19,110–113}

Rational identification of these hotspots, as DE does, has been a significant challenge in computational enzyme design. This rationalization needs to find what properties are these distal mutations affecting, which will be the driving force behind determining what positions to select for mutagenesis. These properties can range from conformational dynamism, stability, and solubility, among others.

To tackle this problem, our group developed the Shortest Path Map

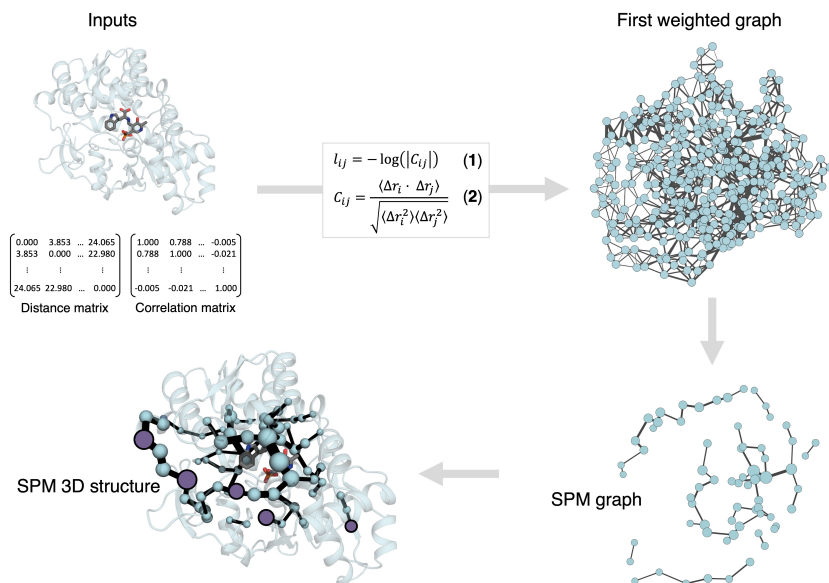


Figure 1.7: **Shortest Path Map (SPM) construction workflow.** Key equations 1 and 2 convert an enzyme into a weighted graph,¹¹⁴ each node representing a residue. The edges linking each pair of nodes are assigned weights following equations 1 and 2, where l_{ij} is the computed correlation value, and Δr_i and Δr_j are the displacement of the $C\alpha$ or $C\beta$ atoms of residues i and j observed in the MD simulation to a reference structure. This complex weighted graph is simplified to identify the shortest paths (SP) that have a higher contribution to the conformational dynamics. SPM can be drawn on the 3D structure to directly assess how different parts of the enzyme are connected. Hotspots residues identified in the SPM are colored in lilac.

(SPM), a powerful graph-based method for identifying key dynamic residues. This method utilizes MD simulations to generate correlation and distance matrices between atoms ($C\alpha$ or $C\beta$), generally processed through packages like cpptraj or pytraj in Python.^{115,116} A graph is then constructed, similarly as Sethi et al. for allosteric study of aminoacyl-tRNA synthetases (aaRSs),¹¹⁴ using the igraph package,¹¹⁷ where residues are nodes and edges between residues that are in less of 6\AA are weighted according to the correlation strength:

$$l_{ij} = -\log(|C_{ij}|) \quad (1)$$

Here, C_{ij} corresponds to the computed correlation value calculated as follows:

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle \langle \Delta r_j^2 \rangle}} \quad (2)$$

In this equation, Δr_i and Δr_j are the displacements of the $C\alpha$ or $C\beta$ atoms of residues i and j from their mean positions during the MD simulations to a reference structure.

This extensive graph is then analyzed to determine the shortest path (SP) between all nodes in the graph using Dijkstra's algorithm implemented in the `igraph` module.¹¹⁷ Each edge's usage is counted every time it forms part of the SP. All edges are normalized hereafter, and only those most frequently used are represented in the SPM. The SPM is then superimposed on the 3D structure of the enzyme for clearer analysis and interpretation. This map of connections enables the identification of potential hot-spot residues that can be key for enzyme design, and it provides a straightforward view of the dynamics and relationships between amino acids (Fig. 1.7).

The effectiveness of SPM has already been proven by detecting mutations introduced in DE campaigns. For example, SPM identified 13 mutations in the top-performing RA95.5-8F variant, directly including 7 and adjacent 4 to the SPM, enhancing retro-aldolase activity to match natural enzymes.^{19,118}

Based on different studied enzymes, the recommended approach for analyzing SPM results for enzyme (re)design is to first focus on the key regions of the enzyme, and their interconnectivity (*e.g.*, loops or domains controlling allostery, active site, regulatory pockets...). Still, many residues can be detected using the SPM tool, therefore information about co-evolution can be helpful, although it is not always needed. Those residues that interconnect key regions and are important regarding co-evolution, can be treated as hot spot positions and can be considered as mutation points. In the ideal case, site-directed mutagenesis could be applied if a small set of positions is found. Nevertheless, this experimental technique is not always an option so co-evolutionary information can be beneficial to decide which amino acid residue should be introduced in that particular position.

1.2.5 Protein folding and the success of AlphaFold2

In 1969, Cyrus Levinthal highlighted a fundamental issue in protein folding that would later be known as Levinthal's paradox: an unfolded protein cannot find its folded states by randomly trying every possible configuration, because the number of potential configurations is astronomically high so requiring a time longer than the age of the universe to reach its native folded state.^{119,120} Given that proteins fold within milliseconds to seconds, except when slowed by factors like proline isomerization,¹²¹ makes the paradox.¹²² Levinthal's problem makes us think about how vast the folding problem could be if it happens with the simplest model, a random search in a FEL of equally probable states (*i.e.*, without energy barriers between them) with

a single minimum corresponding to the most thermodynamically favorable state, also referred to as the *golf-course* potential surface.¹²³

The evolution of Levinthal's *golf-course* definition arises with the introduction of the funnel-shape energy landscape, which represents a broader thermodynamic principle applicable to the folding of proteins, RNA, or any polymers.^{124,125} This model illustrates how protein folding progresses energetically downhill, where the width of the funnel defines the configurational entropy of the system during protein folding. All these models agree on the existence of a global minimal energy state that corresponds to the native structure, the so-called Anfinsen's dogma of protein folding, also known as the thermodynamic hypothesis. Anfinsen postulated this in 1972 when he pointed out that the 3D structure of a native protein in its normal physiological environment is the one in which the Gibbs free energy of the *whole system* is lowest, determined by the totality of interatomic interactions *and hence by the amino acid sequence*.^{126,127} In that article, Anfinsen also indicated the need to predict in advance this 3D structure for major progress in the field.¹²⁸

The protein structure field tried to solve this problem and predict the native protein structure with just the sequence. In 1994, Professor John Moult and Professor Krzysztof Fidelis founded the Critical Assessment of Structure Prediction (CASP), a biennial blind protein prediction competition created to monitor the state-of-the-art (SOTA) in the field by measuring the accuracy of predictions with the Global Distance Test (GDT), which scores from 0-100. Over the CASP competitions, the scientific teams achieved median GDT values across all targets of around 60 GDT. It wasn't until the CASP14 competition that DeepMind presented the AF2 system, a machine learning (ML) model to predict protein structures that achieved a median GDT score of 92.4 and marked a paradigm shift in the field.¹⁵

DeepMind's deep neural network (DNN) model utilizes an attention-based transformer architecture that interprets the complex 'spatial graph' of proteins. AF2 is trained end-to-end (*i.e.*, the model learns from raw input data directly into 3D structures, autonomously discovering the necessary features and transformations needed to predict protein structures without segmented or manually engineered processing steps) using evolutionarily related sequences, from multiple sequence alignment (MSA) obtained from large sequence databases, and a representation of amino acid residue pairs formed by around 170,000 Protein Data Bank (PDB) structures and 355,993 self-distilled unlabeled data predicted by an undistilled model (*i.e.*, trained on just the PDB dataset).

The AF2 computation begins with a preprocessing pipeline, where the user only needs to input the protein's FASTA sequence of the protein that wants to be predicted. The first step is to generate and process an MSA with

tools like JackHMMER¹²⁹ from HMMER3¹³⁰ and HHblits¹³¹ using large sequence databases such as MGnify,¹³² UniRef90,¹³³ Uniref30 (*i.e.*, formerly known as Uniclust30),¹³⁴ and BFD.¹³⁵ These MSAs are then de-duplicated and stacked, with MSA depth varying based on the database used. Then, PDB Templates are identified using the previous UniRef90 MSA to search PDB70 with HHSearch, selecting the top 4 templates based on alignment quality.¹³⁶

Once we have the stacked MSA and chosen templates, AF2 inference can start. First, the feature embedding process is made, which creates the FASTA sequence features corresponding to the atom types and residue indexes, defined as *target_feat* and *residue_index*. Then, the MSA features, *msa_feat*, is created, which for computational and memory reasons, the MSA is reduced to 512 randomly selected sequences that are the MSA cluster centers (*i.e.*, *max_msa_clusters*), which are subsequently masked for processing. The remaining sequences are assigned to the closest cluster, improving the feature extraction of statistics from all sequences in the MSA. Additionally, a specific number of non-cluster center sequences (*i.e.*, *max_extra_msa*) are randomly sampled to create the input feature *extra_msa_feat* to promote diversity. Finally, template features are extracted to obtain the spatial information of the PDBs creating the *template_pair_feat* and *template_angle_feat*. During training for the CASP14 competition, AF2 created five models, where the *max_extra_msa* from models 1, 3, and 4 use 5120 sequences, and models 2 and 5 use 1024. In the case of template features, models 1 and 2 integrate PDB template data, while models 3 to 5 do not. After CASP14, five more models were developed, incorporating a pTM (predicted global superposition metric template modeling score)¹³⁷ prediction objective.

Once the features are created, the inference loop starts. The core computational step involves transforming the MSA and template features into "MSA representations" and "pair representations" and iteratively updating them. The MSA representation encodes sequences, highlighting similarities and differences across the alignment, while the pair representation captures relations between pairs of amino acids, essential for understanding residue interactions. The representation updating process is done in the Evoformer modules, central to AlphaFold 2's success, consisting of 48 blocks that iteratively refine these representations to enhance spatial and interactional residue insight, crucial for accurate structure prediction. Post-Evoformer, the refined representations are channeled into the Structure Module. This module applies multiple loss functions, such as Frame Aligned Point Error (FAPE) and torsion angle loss, to accurately translate the abstract representation of the protein structure (created by the Evoformer stack representations) into concrete 3D atom coordinates.

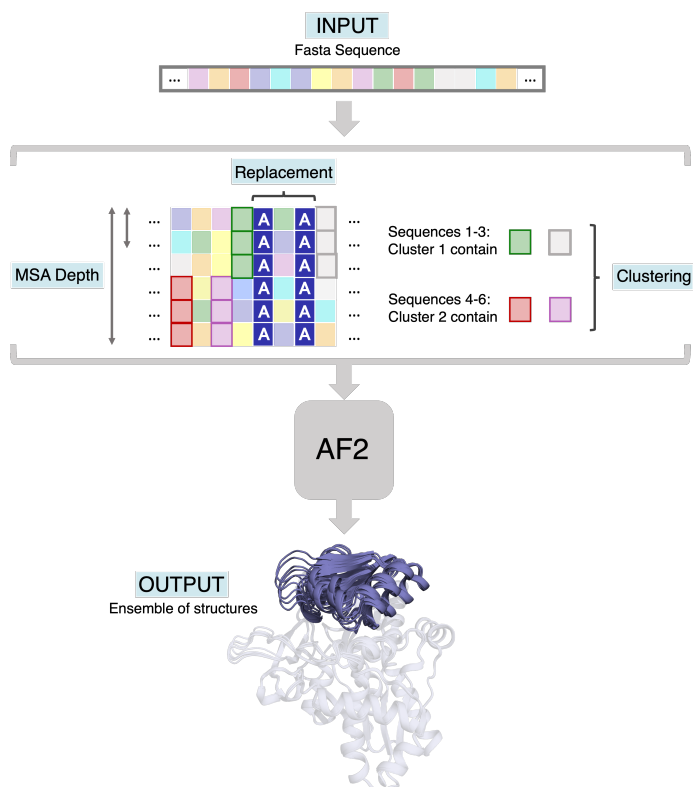


Figure 1.8: **Overview of different strategies developed for predicting different conformational states with AlphaFold2 (AF2).** Multiple Sequence Alignment (MSA) depths can be altered,¹³⁸ some of the MSA positions can be masked,¹³⁹ and the MSA can be clustered.¹⁴⁰

Recycling plays a critical role here, reprocessing the outputs from the Structure Module back as inputs in subsequent iterations in the inference loop, enabling the model to refine its predictions through multiple cycles and use different MSA inputs (*i.e.*, *msa_feat*). Optionally, multiple iterations of the Evoformer module, as done in CASP14 predicted structures with 8 iterations (*i.e.*, $N_{ensemble}=8$), can be executed with different MSA inputs and averaged before the structure module.

Finally, each protein residue's position is evaluated for confidence with per-residue local distance difference test (pLDDT) scores,¹⁴¹ reflecting the reliability of the structural prediction. This elaborate orchestration of embeddings, iterative refinement, and confidence assessment underscores AF2's advanced approach to predicting protein structures from sequence data and template PDBs.

Overall, AF2 solved the protein structure prediction problem for a wide range of cases, but not for all the proteins. AF2 struggles with highly flexible proteins, or proteins with a huge degree of native states, that is the case for proteins with a huge percentage of intrinsically disordered regions (IDR). In this regard, AF2 has been trained to find the native structure of the target sequence, which has been attributed to what Anfinsen's Dogma says. However, it is important to realize that Anfinsen defines the native structure as "the one in which the Gibbs free energy of the *whole system* is lowest", describing as "*whole system*" the protein plus solvent. In this regard, AF2 does not account for solvent effects, nor can be attributed to Anfinsen's statement that the native structure is determined *only* by the protein's amino acid sequence.¹²⁸ Although it is important to clarify that AF2 does not solve the folding problem, as during the inference iterations AF2 is not finding folded intermediates from the FEL, in the end, AF2 is focused on finding the folded native structure from the sequence. However, even though the AF2 architecture does not compute any energy features to predict the structure, a recent study has shown that AF2 has effectively learned an implicit representation of the biophysical energy landscape.¹⁴² The study further suggests that AF2 uses this internalized energy function, driven by insights from MSA and co-evolutionary data, to navigate the complex protein folding space. This enables the identification of stable conformations that likely represent the native state of the protein without explicitly calculating energy during the prediction process. The importance of this finding suggests that altering the input MSA information can successfully sample alternative confirmation states. This is further demonstrated by the work of del Alamo et al., varying the MSA depth,¹³⁸ masking specific MSA positions as demonstrated by Stein and McHaourab,¹³⁹ and clustering the MSA following methods as discussed by Kern and Ovchinnikov can influence the prediction outcomes (Fig. 1.8).¹⁴⁰ Additionally, modifying the set of provided templates, as also shown in del Alamo et al. and chapter 4 of this thesis, can further diversify the conformational states captured.¹³⁸

Chapter 2:

Objectives

This thesis aims to advance the field of enzyme design through the development of computational tools, with a focus on the insights retrieved from protein dynamics through MD simulations. The specific objectives are organized into distinct yet interconnected components, each addressing critical aspects of the enzyme design pipeline, such as mutational hot spots identification and protein conformation prediction, with a final system example:

1. **Development and Deployment of the SPM Webserver:** The first objective is to speed up the code and deploy the SPM tool, using MD simulations to discover key dynamic residues that influence enzyme functionality beyond the active site, as discussed in Chapter 3. Additionally, this objective comprises launching the SPM webserver to provide the scientific community with robust, accessible, and user-friendly computational tools for enzyme research and design.
2. **Advances in Deep Learning for Protein Design and Refining AlphaFold2 with a Template-Based Approach:** In Chapter 4, the objective is first to show the advances in DL methods and AF2 usage for protein and enzyme design, and second to enhance the conformational exploration capabilities of AlphaFold2 beyond its lowest energy state predictions by altering the MSA depth and incorporating diverse structural templates. This approach aims to enable the model to predict a broader range of functional conformations, thereby enhancing our understanding of protein dynamics. Our interest is not limited to the conformational heterogeneity obtained from AF2 predictions, as we want to prove that those structures can speed up the FEL reconstruction through MD simulations. To that end, we will use TrpS as a case example.
3. **Rational Design of Efficient Enzyme Variants from HNL to EST:** The specific goal for Chapter 5 is to use the computational tools developed to rationally design and transform a natural HNL into an efficient EST. This involves identifying and predicting the minimal set of mutations required to yield this conversion, utilizing insights from the SPM tool to guide the mutation selection. This objective is focused on demonstrating the power of computational approaches in achieving targeted enzyme functionality with precision and a rational explanation.

Chapter 3:

The shortest path method (SPM) webserver for computational enzyme design

This chapter corresponds to the following publication:

Casadevall, G.; Casadevall, J.; Duran, C.; Osuna, S. The Shortest Path Method (SPM) Webserver for Computational Enzyme Design. *Protein Eng. Des. Sel.*, **2024**, *37*, gzae005.

Reproduced with permission from: Casadevall, G.; Casadevall, J.; Duran, C.; Osuna, S. The Shortest Path Method (SPM) Webserver for Computational Enzyme Design. *Protein Eng. Des. Sel.*, **2024**, *37*, gzae005, by permission of Copyright ©2024 Oxford University Press.

The shortest path method (SPM) webserver for computational enzyme design

Guillem Casadevall¹, Jordi Casadevall², Cristina Duran¹ and Sílvia Osuna^{1,3,*}

¹Institut de Química Computacional i Catàlisi and Departament de Química, Universitat de Girona, c/Maria Aurèlia Capmany 69, Girona 17003, Spain

²Carrer Tancat 2, Vilajuïga 17493, Spain

³ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain

*To whom correspondence should be addressed. Institut de Química Computacional i Catàlisi and Departament de Química, Universitat de Girona, c/Maria Aurèlia Capmany 69, Girona 17003. E-mail: silvia.osuna@udg.edu

Edited by: Petersen, Eva

Abstract

SPMweb is the online webserver of the Shortest Path Map (SPM) tool for identifying the key conformationally-relevant positions of a given enzyme structure and dynamics. The server is built on top of the DynaComm.py code and enables the calculation and visualization of the SPM pathways. SPMweb is easy-to-use as it only requires three input files: the three-dimensional structure of the protein of interest, and the two matrices (distance and correlation) previously computed from a Molecular Dynamics simulation. We provide in this publication information on how to generate the files for SPM construction even for non-expert users and discuss the most relevant parameters that can be modified. The tool is extremely fast (it takes less than one minute per job), thus allowing the rapid identification of distal positions connected to the active site pocket of the enzyme. SPM applications expand from computational enzyme design, especially if combined with other tools to identify the preferred substitution at the identified position, but also to rationalizing allosteric regulation, and even cryptic pocket identification for drug discovery. The simple user interface and setup make the SPM tool accessible to the whole scientific community. SPMweb is freely available for academia at <http://spmosuna.com/>.

Keywords: Shortest Path Method, webserver, computational enzyme design, distal mutations

Introduction

Enzyme design aims to create novel biocatalysts with enhanced properties through the modification of their natural amino acid sequences or *via* generation of novel sequences and folds. The fascination with enzyme design and engineering is motivated by the advantageous features exhibited by these catalysts, including their capacity to function effectively under gentle biological conditions, achieving remarkable efficiency, selectivity, and specificity. Enzyme design is also an intellectual challenge, as it is a stringent examination of what we understand of enzyme stability, folding, evolution and catalysis.

Designing enzymes taking as starting point a natural or computationally reconstructed/generated scaffold involves selecting specific residues for mutagenesis, generating new variants, and employing screening protocols to assess improvements in targeted properties (Bell et al. 2021). Two main approaches exist: rational design (Damborsky and Brezovsky 2014, Romero-Rivera et al. 2017a, Maria-Solano et al. 2018) considering *de novo* and natural scaffolds, and Directed Evolution (DE) (Arnold 2015, Currin et al. 2015), which can be successfully combined to achieve higher levels of performance. Rational design focuses on predetermined hotspot positions, identified through multiple sequence alignments, structural analysis of active site pockets, potential substrate-binding tunnels, and comprehensive computational modeling (using techniques like Quantum Mechanics, Quantum Mechanics/Molecular Mechanics, Molecular Dynamics,

and MonteCarlo simulations) (Romero-Rivera, Garcia-Borràs and Osuna, 2017a, Sequeiros-Borja et al. 2020). Rational design efforts often focus on the active site pocket or in the bottleneck regions of the computed substrate binding tunnels and gates. The user-friendly tools such as CAVER (Stourac et al. 2019), AQUA-DUCT (Stourac et al. 2019), and HotSpot Wizard (Sumbalova et al. 2018), among others can be used (Sequeiros-Borja et al. 2020). In contrast, DE (Francis and Hansche 1972, Lutz and Borscheuer 2008, Borscheuer et al. 2012, Packer and Liu 2015), honored with the 2018 Nobel Prize in Chemistry, initially relied on iterative cycles of random mutagenesis. Recent advancements integrate bioinformatic tools (Jiang et al. 2008, Rothlisberger et al. 2008, Kuipers et al. 2009, Kourist et al. 2010, Kuipers et al. 2010, Siegel et al. 2010), sequence analysis (Pavelka et al. 2009, Addington et al. 2013), smarter libraries, protein engineering techniques (Kazlauskas and Borscheuer 2009, Turner 2009, Borscheuer et al. 2012), gene synthesis (Currin et al. 2014), and high-throughput screening techniques (Xiao et al. 2015). Machine-learning sequence-function models can be used to guide DE (Yang et al. 2019, Mazurenko et al. 2020). As mentioned above, the powerful DE strategy can be applied to boost the low activities of computational enzyme designs (Jaeckel et al. 2008, Romero and Arnold 2009, Renata et al. 2015) and enhance promiscuous enzymatic side-activities (Campbell et al. 2016, Leveson-Gower et al. 2019). Multiple laboratory-engineered enzymes have been reported in the literature, including enzymes for the production of drugs,

Received: December 22, 2023. Revised: February 21, 2024. Accepted: February 28, 2024

© The Author(s) 2024. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

biotherapeutics, potential bulk products, and fragrances (Buller et al. 2023).

A notable strength of DE lies in its capability to introduce mutations throughout the entire protein sequence. This contrasts with rational design approaches that are often restricted to alterations in the active site pocket or available tunnels and gates for promoting substrate binding/product release and altering the water content (Gora et al. 2013, Sequeiros-Borja et al. 2020). As observed in numerous DE studies, the remarkable fold increases in catalytic activity achieved are accomplished thanks to mutations positioned far from the active site, which are computationally very challenging to predict (Jiménez-Osés et al. 2014, Currin et al. 2015, Obexer et al. 2017, Osuna 2021). This trend extends to diverse enzymes such as cytochrome P450, Diels-Alderase, phosphotriesterase, sitagliptinase, among many additional ones (Osuna 2021). Often laboratory-evolved enzymes present mutations introduced at an average distance of around 15 Å from the active site (Currin et al. 2015). Intriguingly, there is no direct correlation between the impact of introduced mutations on enzyme turnover (k_{cat}) and their proximity to the active site, in contrast to the more deterministic role of active site mutations in specificity (Currin et al. 2015). The coupling of distal residues affecting the enzyme catalytic activity suggests a substantial influence of long-range allostery, *i.e.* regulation of catalytic activity by effector and/or protein binding, in many proteins (Gunasekaran et al. 2004). Extensive MD simulations have successfully rationalized how distal mutations influence the multiple conformations enzymes can adopt thus impacting its catalytic activity (Jiménez-Osés et al. 2014, Romero-Rivera et al., 2017b). Distal mutations often alter non-covalent interaction networks, which might favor some additional conformational states of the enzyme that are more optimal for the promiscuous activity to be enhanced and/or modify the flexibility of crucial structural elements such as loops and lids gating the active site pocket (Campbell et al. 2016, Petrović et al. 2018, Curado-Carballada et al. 2019). While computational modeling can satisfactorily explain these changes in activity induced by distal alterations, the challenge remains in predicting which distal mutations can impact and regulate enzymatic activity (Jiménez-Osés et al. 2014, Osuna 2021, Campitelli et al. 2020). Given the insights from DE that distal mutations are essential for enhancing enzyme catalytic activity, the development of computational tools capable of predicting remote mutations holds great promise, potentially advancing our underdeveloped ability to computationally design efficient Nature-like enzymes (Osuna 2021).

The effect exerted by distal mutations in enzyme design reminds the allosteric regulation effect produced by effector binding in allosteric systems or within the active sites of heterocomplexes that present synchronised transportation of substrates. Distal mutations can induce a shift in the conformational landscape, thus favouring the catalytically competent arrangement of the catalytic residues for catalysis. Given the striking similarity between these two scenarios (enzyme design and allosteric regulation), we explored the potential development and application of correlation-based tools in enzyme design (Romero-Rivera, Garcia-Borràs and Osuna, 2017b, Osuna 2021, Maria-Solano et al. 2018). We developed the Shortest Path Map (SPM, DynaComm.py) tool by constructing a first complex graph based on mean distances and correlation values between the residues that compose the enzyme

computed during MD simulations, similar to the protocol by Sethi et al. (Sethi et al. 2009) for studying allosteric systems (see Fig. 1) (Osuna 2021). In contrast to prior allosteric studies concentrating on identifying communities in the graph (Sethi et al. 2009), our SPM approach involves computing shortest path lengths using the Dijkstra algorithm implemented in the igraph module (Csárdi and Nepusz 2006). Consequently, it identifies those pairs of residues that are more correlated and have a higher impact into the enzyme conformational dynamics. Unlike community analysis that highlights important regions of the enzyme, SPM directly identifies the most crucial residues rather than regions. This feature is particularly appealing for enzyme design, enabling the direct construction of small libraries of hotspot positions.

SPM narrows down the sequence space to a subset of conformationally relevant positions, with a notable capability to pinpoint challenging distal positions that enhance activity (Osuna 2021). The successful application of SPM in identifying DE mutations in retro-aldolase, monoamine oxidase, and tryptophan synthase enzymes suggests its potential utility in the rational design of enzyme variants (Osuna 2021). The Mulholland lab utilized our SPM tool to assess changes in dynamical networks during the transition-state ensemble along DE of a computationally designed Kemp eliminase (Bunzel et al. 2021). Additionally, we used SPM to investigate allosteric communication within monomers, and in allosteric systems (Curado-Carballada et al. 2019, Calvó-Tusell et al. 2022, Castelli et al. 2024). More recently, we have also used SPM for rational enzyme design in combination with other tools to further reduce the number of identified positions and select the specific amino acid at each site, as described in the following examples. We combined SPM with ancestral sequence reconstruction for developing new stand-alone tryptophan synthase B (TrpB) variants (Maria-Solano et al. 2021). Focusing on including the ancestral amino acid in the non-conserved SPM positions, our approach increased the stand-alone activity of the new SPM6-TrpB variant by 7-fold (in terms of k_{cat}) (Maria-Solano, Kinateder, Iglesias-Fernández, Sterner and Osuna 2021). It is worth noting that, while testing only a single variant, the fold increase in k_{cat} was comparable to the 9-fold obtained through DE, which required generating and screening over 3000 variants. In a recent pre-print, we showcased the efficacy of our SPM methodology in designing efficient Nature-like enzymes. Specifically, we achieved a more than 1300-fold increase in the esterase catalytic efficiency of a hydroxynitrile lyase (HNL), surpassing the esterase activity of the reference enzyme (Casadevall et al. 2023). Altogether, these studies provide compelling evidence for the potential of our SPM methodology in computational enzyme design.

In this study, we develop and describe the webserver version of the SPM tool for its application in enzyme design for academic use. First, we discuss the user-friendly webserver generated, the input files needed and the overview of the settings that the user can alter to generate different SPM maps. Second, we show with the tryptophan synthase example how information of inter and intramolecular SPM communications networks can be withdrawn. With this tool, we hope the academic community can benefit from the application of the SPM in the study of biomolecular systems and aim to expand the current area of application of the SPM methodology.

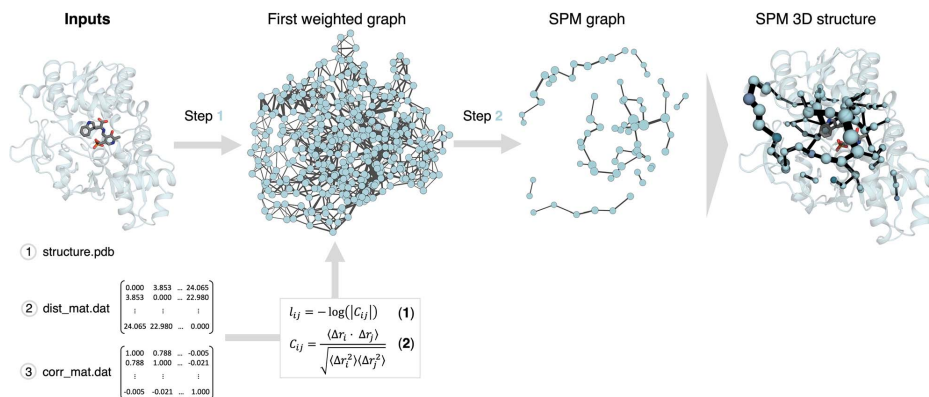


Fig. 1. Workflow and equations for Shortest Path Map (SPM) construction for computational enzyme design. The enzyme is simplified as a weighted graph as done for studying allostery (Sethi et al. 2009), however, this complex graph is simplified to identify the shortest paths (pairs of residues) that have a higher contribution to the conformational dynamics. SPM can be drawn back on the 3D structure to directly assess how the active site pocket is connected to active site and distal sites. The key equations (1 and 2) for converting a protein into a graph are also displayed. Each node in the first complex graph represents a residue. The edges linking each pair of nodes (residues) are assigned weights in accordance with equation 1 and 2, where C_{ij} is the computed correlation value, Δr_i and Δr_j are the displacement of the C_α of the i, j residue observed in the MD trajectory with respect to a reference structure.

Results

Workflow

The basic workflow for SPM construction is shown in Fig. 1. As described in the introduction, the enzyme structure and dynamics is simplified using a weighted graph (step 1), which is then further processed to identify the shortest paths to generate the final SPM graph (step 2). SPM can then be plotted back into the 3D-dimensional structure to visualize how the active site pocket is connected to more remote sites.

Generation of the first weighted graph

Initial attempts to apply graph theory to investigate allosteric regulation primarily focused on the static X-ray structure of the enzyme (Guo and Zhou 2016). In the constructed graph, two sets of nodes (residues) were linked by an edge if the distance between their representative atoms fell below a specific threshold. The significant advancement in graph construction came from Sethi et al. (Sethi et al. 2009), who employed short MD simulations (lasting a few nanoseconds) to determine the connected nodes and their respective edge weights. An edge was established between a pair of residues (nodes) if the representative atoms of each residue remained below a defined distance threshold (see Fig. 1) for a specified fraction of the MD simulation time. The edge connecting residues (i, j) was weighted based on their correlation values (C_{ij} , as outlined in equations 1 and 2, Fig. 1). Residues undergoing highly correlated conformational changes during the MD simulation (i.e. $C_{ij} \rightarrow 1$) were linked by a relatively short edge ($l_{ij} \rightarrow 0$). Conversely, a pair of residues with non-correlated movements ($C_{ij} \rightarrow 0$) were connected by relatively long edges ($l_{ij} \rightarrow \infty$).

In this protocol, the enzyme conformational dynamics is summarized through this first weighted graph (shown in Fig. 1). Further subdivision of the graph into communities, utilizing the Girvan-Newman algorithm (Girvan and Newman 2002), results in the identification of what is called the optimal community network used in the study of

allosterically-regulated enzymes (Rivalta et al. 2012, Schupfner et al. 2020). However, for computational enzyme design it is more preferred to identify a subset of positions rather than regions or communities.

Generation of the SPM

For SPM generation instead of pinpointing communities within the initial graph, we use the Dijkstra algorithm, implemented in the igrph module (Csárdi and Nepusz 2006), to calculate the shortest path lengths. The algorithm considers all nodes of the graph and determines the shortest path from the first to the last protein residue. Consequently, the method identifies the edges in the graph that are shorter, thus indicating higher correlation and more frequently used in going through all protein residues. All edges are then normalized, and only those with the most significant contribution (a visualization/significance threshold is applied, see Fig. 2) are represented in the SPM. Drawing the SPM directly onto the 3D structure of the protein, rather than its 2D graph representation (see Fig. 1), is more advantageous as one can directly see how the network expands through the 3D structure. The primary benefit of SPM over community analysis lies in directly identifying the most critical residues (as opposed to regions), making it more appealing for enzyme design, as small libraries of hotspot positions can be constructed directly. SPM enables the prediction of distal active site mutations that lead to enhanced enzymatic activity for the first time in a computational protocol (Osuna 2021).

Description of the webserver

Input files

There are three mandatory files for SPM construction: the tertiary structure of the enzyme/protein in pdb format for visualizing the results, and the distance and correlation matrices obtained often through MD simulations (but not necessarily

Fig. 2. SPM main page of the webserver. The user needs to upload in the corresponding boxes the three mandatory files that are needed for SPM construction: the enzyme 3D structure, and the two matrices: distance and correlation previously computed from the MD simulations. Two important parameters can be modified for SPM construction: the distance threshold (bottom right panel), and the significance threshold (bottom left panel). The webserver link is: <https://spmosuna.com>.

restricted to). Our recommendation is to generate the distance and correlation matrices using at least three replicates of MD simulations of 200–500 ns of simulation length in explicit solvent and considering either C_{α} or C_{β} positions. The calculation of the distance and correlation matrices can be done considering the whole MD trajectory, the last 100–200 nanoseconds of the MD simulations or using distinct sets of conformations in case of proteins undergoing large conformational changes. We, however, recommend using either the whole MD trajectory or the last 100–200 ns of the MD runs (Duran et al. 2024).

The distance and correlation matrix can be computed with different MD analysis software, but we provide as example the input file used for cpptraj included in AMBER tools:

Input files for cpptraj module for computing the correlation and proximity matrices:

We recommend taking as reference the most populated cluster from the MD trajectory. This is especially relevant for proteins undergoing large conformational changes.

```
cpptraj <parm file>
reference structure.pdb
trajin MD_trajectory.nc 1 last 1
rms reference @CA
matrix dist @CA out dist_mat.dat
matrix correl @CA out corr_mat.dat
exit
```

The three mandatory files (structure.pdb, dist_mat.dat, corr_mat.dat) can then be uploaded in the corresponding boxes included in the main page of the webserver (see Fig. 2). It should be also mentioned that the webserver also accepts the distance and correlation matrices as numpy binary files (.npz).

SPM parameters

As discussed in the previous section, two thresholds need to be defined for SPM construction. The first one is related to the mean distance value between the user defined atoms along the MD simulation (often distances between either C_{α} or C_{β}). While we recommend the use of a distance threshold of 6 Å, in some cases, it might be useful to play with the distance matrix threshold. Increasing this value to higher numbers will

of course consider a higher portion of the protein residues for each targeted site, and thus the computed SPM graph will contain a larger number of positions. In the opposite direction, rather small values for the distance matrix will only consider nearby residues thus being very local and restricted (see the distance threshold tests in the case example below).

The other important threshold is related to the number of positions represented in the final SPM graph. This visualization/significance threshold will restrict the number of edges and nodes displayed. We recommend a threshold of 0.3, as it will reduce the number of positions and will only display the ones playing a higher role in the conformational dynamics. In any case, we believe it might be also useful to play with the visualization/significance threshold as well to visualize a higher proportion of the identified edges and evaluate how the disconnected parts of the graph are actually connected. Therefore, this has been added as an extra parameter in the SPM webserver. In Fig. 2, the two boxes related to distance and significance threshold are also displayed.

Output files

SPM visualization.

After uploading the three requested input files, the SPM is built and visualized in the main screen panel (see Fig. 3). SPM can be shown in the 3D structure of the uploaded protein structure, where the important residues are marked with spheres and labelled according to its ranked ID, and edges connecting the pairs of residues are highlighted in black. Those pairs of residues that have a higher contribution to the conformational dynamics present bigger spheres and thicker edges. However, the sizes of spheres and widths of edges are mostly qualitative. By default, a distance threshold of 6 Å and a visualization/significance threshold of 0.3 is used. However, as mentioned before, these two parameters can be modified using the threshold panels. The SPM is also displayed in 2D in an additional panel below the 3D representation, in which the residue labels and connections can be more easily seen.

PyMoL script for SPM visualization.

Another interesting feature of the SPM webserver is that it generates a PyMoL script that can be executed in PyMoL

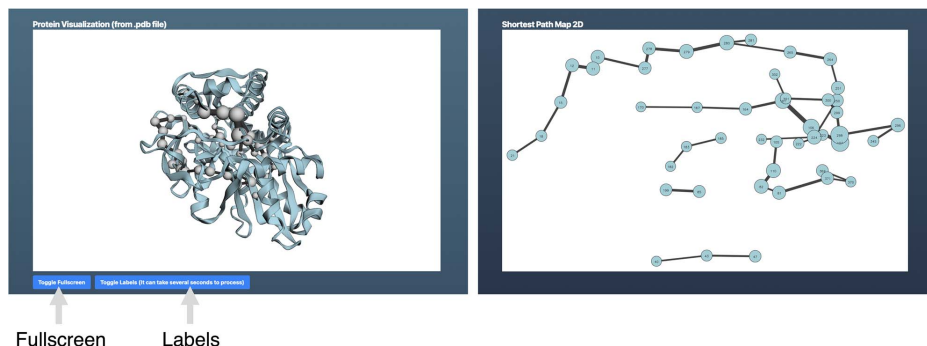


Fig. 3. SPMweb main page displaying the output after running the SPM calculation. SPM is visualized on top of the 3D structure of the protein as shown in the left panel. The 2D representation of the SPM graph is also shown (right panel). The results can be visualized as full screen by clicking the 'Toggle Fullscreen' button, and the labels of the atoms can also be added/removed by clicking the 'Toggle Labels (It can take several seconds to process)' button.

software after loading the 3D structure of the enzyme. The visualization in PyMol is rather simple, the user needs to first load the reference structure (load reference.pdb), and then in the command line execute the SPM pymol script defining the correct path where it is located (@\$PATH/pymol_shortest_path_reference). In pymol, the user can tune all parameters and also include some transparency into the cartoon of the protein structure to visualize better the SPM graph (for instance by typing the command in the command line: set cartoon_transparency, 0.6).

Case examples

SPMweb can be used to address different relevant enzymatic properties. We provide some examples of how the SPM tool can be employed: (1) to identify the conformationally relevant distal positions connected to the enzyme active site for the generation of some mutational libraries, and (2) to rationalize the existing allosteric communication between the enzyme subunits in a dimeric structure.

Case example 1. Identification of the key conformationally relevant positions either at the active site or at distal sites connected to the catalytic pocket.

SPM can be applied for identifying mutational spots not restricted to the active site and neither to the tunnel regions targeted by DE. Along the years, we have shown how SPM allows, for the first time, the prediction of which distal active site positions might lead to enhanced enzymatic activity after mutation (Romero-Rivera, García-Borràs and Osuna, 2017b). This has been tested in different unrelated enzyme families showcasing the potential of SPM for the rational design of enzyme variants (Osuna 2021). In this case example, we applied the SPM in the case of tryptophan synthase B (TrpB) subunit, as we first realized that SPM was capturing some of the DE positions (Maria-Solano et al. 2019) and subsequently applied it for designing a stand-alone TrpB (Maria-Solano, Kinateder, Iglesias-Fernández, Sterner and Osuna 2021). As shown in Fig. 4, the computed SPM in the webserver shows how the graph connects the active site pocket that holds the catalytic lysine and the PLP cofactor with remote sites that interestingly contain many of the DE mutations. For constructing this main SPM, the default parameters for the

distance and visualization/significance thresholds have been used (panel A in Fig. 4). However, as shown in panel B in Fig. 4 by changing the two threshold parameters the obtained SPM maps differ quite substantially. Despite $PfTrpB^{OB2}$ being dimeric in solution in the absence of its binding TrpA partner, we computed the distance and correlation matrices considering only one of the monomeric units. This computed SPM therefore identifies the intramolecular conformationally relevant positions with the monomeric structure connected to the active site pocket.

Case example 2. Rationalization of the allosteric pathway existing between monomers in a dimeric enzyme structure.

Another interesting feature to analyze in those enzymatic systems that are not monomeric in solution is the existing communication pathway within subunits. This is particularly relevant for allosterically regulated enzymes such as tryptophan synthase, but also in enzymes that require a higher order oligomeric structure for function like monoamine oxidase (MAO-N) (Curado-Carballada et al. 2019, Osuna 2021). We have again used the example of TrpB that adopts a dimeric structure to analyze the communication existing between the two subunits. In this case, the whole dimeric structure has been used for SPM construction: the distance and correlation matrices are computed considering the complete dimeric structure. As shown in Fig. 5, the computed SPM pathway using the default parameters now expands from one subunit to the other and does not necessarily connect the respective active site pockets of both TrpB monomers. This analysis can be used to identify residues crucial for the intersubunit (allosteric) communication and can also be relevant for explaining cooperative effects.

Conclusions

SPMweb is a new webserver for identifying a subset of conformationally relevant positions located throughout the protein structure. This unique tool can be used for rationally identifying distal sites whose conformational dynamics is connected to the enzyme active site pocket. Although the tool was initially developed for computational enzyme design as

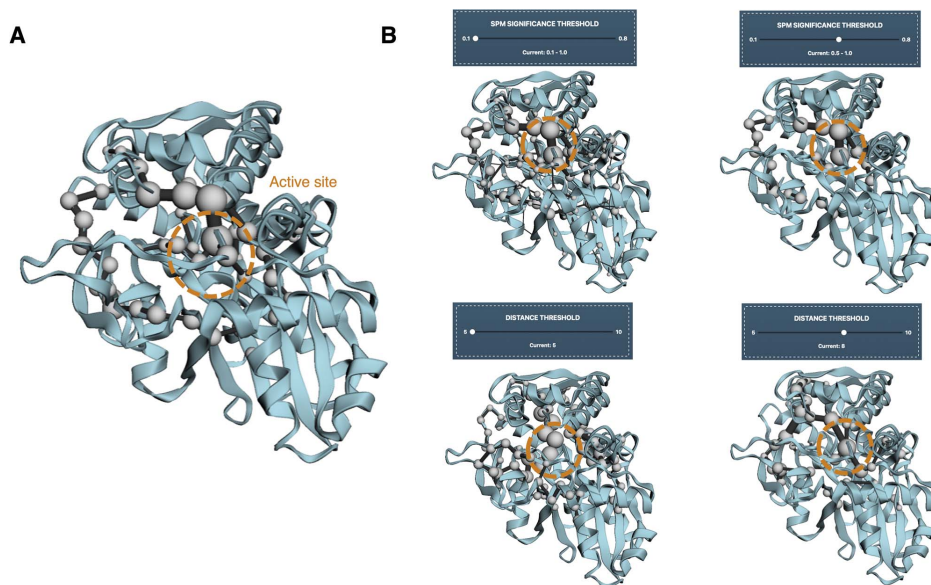


Fig. 4. Case example of computed SPM for investigating the distal sites connected to the active site pocket of the enzyme tryptophan synthase B (*PTrpB^{OB2}*) considering only the monomeric structure. (A) Visualization of the SPM using the default thresholds for significance and distance. (B) Top panel: visualization of the effect of altering the SPM significance threshold using 0.1 (left) and 0.5 (right). Bottom panel: visualization of the effect of altering the distance threshold and using a value of 5 Å (left) and 8 Å (right). The active site of the enzyme that holds the PLP-cofactor and the catalytic residues is highlighted with a discontinuous circle.

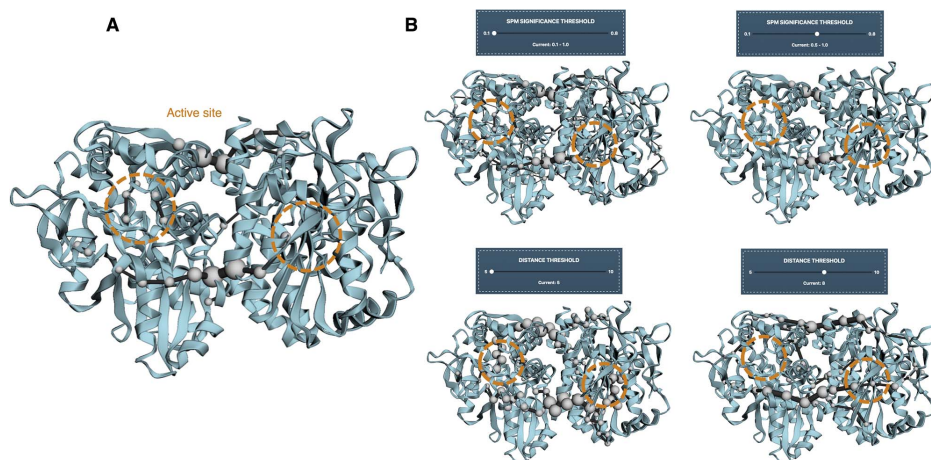


Fig. 5. Case example of computed SPM for studying the allosteric communication existing between monomers in a dimeric tryptophan synthase B (*PTrpB^{OB2}*) structure. (A) Visualization of the SPM using the default thresholds for significance and distance. (B) Top panel: visualization of the effect of altering the SPM significance threshold using 0.1 (left) and 0.5 (right). Bottom panel: visualization of the effect of altering the distance threshold and using a value of 5 Å (left) and 8 Å (right). The active site of the enzyme that holds the PLP-cofactor and the catalytic residues is highlighted with a discontinuous circle.

discussed in the whole paper, the potential applications of this novel methodology are broad. SPM can be directly used for rationalizing the allosteric communication between enzyme subunits as shown in the case example discussed above. However, it could also be potentially applied for instance for identifying cryptic pockets for designing allosteric inhibitors in drug discovery. We hope that by releasing this webserver to the scientific community, the number of applications and successful cases in which SPM can be applied is expanded.

Author contributions

Guillem Casadevall (Conceptualization [equal], Methodology [equal], Software [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Jordi Casadevall (Conceptualization [equal], Methodology [equal], Software [equal]), Cristina Duran (Conceptualization [equal], Methodology [equal], Software [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Silvia Osuna (Conceptualization [equal], Methodology [equal], Software [equal], Writing—original draft [equal], Writing—review & editing [equal]).

Supplementary data

Supplementary data is available at *PROENG Journal* online.

Funding

We thank the Generalitat de Catalunya for the consolidated group TCBioSys (SGR 2021 00487), Spanish MICIN for grant projects PID2021-129034NB-I00 and PDC2022-133950-I00. 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, ERC-2022-POC-101112805, and ERC-2022-CoG-101088032), and the Human Frontier Science Program (HFSP) for project grant RGP0054/2020. C.D. was supported by the Spanish MINECO for a PhD fellowship (PRE2019-089147), and G. C. by a research grant from ERC-StG (ERC-2015-StG-679001) and ERC-POC (ERC-2022-POC-101112805).

References

- Addington, T.A., Mertz, R.W., Siegel, J.B. *et al.* (2013) *J Mol Biol*, **425**, 1378–1389. <https://doi.org/10.1016/j.jmb.2013.01.034>.
- Arnold, F.H. (2015) *Q Rev Biophys*, **48**, 404–410. <https://doi.org/10.1017/S003358351500013X>.
- Bell, E.L., Finnigan, W., France, S.P. *et al.* (2021) *Nat Rev Methods Primers*, **1**, 46. <https://doi.org/10.1038/s43586-021-00044-z>.
- Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J. *et al.* (2012) *Nature*, **485**, 185–194. <https://doi.org/10.1038/nature11117>.
- Buller, R., Lutz, S., Kazlauskas, R.J. *et al.* (2023) *Science*, **382**, eadh8615. <https://doi.org/10.1126/science.adh8615>.
- Bunzel, H.A., Anderson, J.L.R., Hilvert, D. *et al.* (2021) *Nat Chem*, **13**, 1017–1022. <https://doi.org/10.1038/s41557-021-00763-6>.
- Calvó-Tusell, C., Maria-Solano, M.A., Osuna, S. *et al.* (2022) *J Am Chem Soc*, **144**, 7146–7159. <https://doi.org/10.1021/jacs.1c12629>.
- Campbell, E., Kaltenbach, M., Correy, G.J. *et al.* (2016) *Nat Chem Biol*, **12**, 944–950. <https://doi.org/10.1038/nchembio.2175>.
- Campitelli, P., Modi, T., Kumar, S. *et al.* (2020) *Annu Rev Biochem*, **49**, 267–288. <https://doi.org/10.1146/annurev-biochem-052118-115517>.
- Casadevall, G., Pierce, C., Guan, B. *et al.* (2023) *bioRxiv* 2023.2008.2023.554512. <https://doi.org/554510.551101/552023.554508.554523.554512>.
- Castelli, M., Marchetti, F., Osuna, S. *et al.* (2024) *J Am Chem Soc*, **146**, 901–919. <https://doi.org/10.1021/jacs.3c11396>.
- Csárdi, G. and Nepusz, T. (2006) *InterJournal*, **Complex Systems**, 1695–1704.
- Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J. *et al.* (2019) *Angew Chem Int Ed*, **58**, 3097–3101. <https://doi.org/10.1002/anie.201812532>.
- Curran, A., Swainston, N., Day, P.J. *et al.* (2014) *Protein Eng Des Sel*, **27**, 273–280. <https://doi.org/10.1093/protein/gzu029>.
- Curran, A., Swainston, N., Day, P.J. *et al.* (2015) *Chem Soc Rev*, **44**, 1172–1239. <https://doi.org/10.1039/C4CS00351A>.
- Damborsky, J. and Brezovsky, J. (2014) *Curr Opin Chem Biol*, **19**, 8–16. <https://doi.org/10.1016/j.cbpa.2013.12.003>.
- Duran C, Casadevall G, Osuna S. 2024; submitted for publication.
- Francis, J.C. and Hansche, P.E. (1972) *Genet*, **70**, 59–73. <https://doi.org/10.1093/genetics/70.1.59>.
- Girvan, M. and Newman, M.E.J. (2002) *Proc Natl Acad Sci U S A*, **99**, 7821–7826. <https://doi.org/10.1073/pnas.122653799>.
- Gora, A., Brezovsky, J. and Damborsky, J. (2013) *Chem Rev*, **113**, 5871–5923. <https://doi.org/10.1021/cr300384w>.
- Gunasekaran, K., Ma, B. and Nussinov, R. (2004) *Proteins*, **57**, 433–443. <https://doi.org/10.1002/prot.20232>.
- Guo, J. and Zhou, H.-X. (2016) *Chem Rev*, **116**, 6503–6515. <https://doi.org/10.1021/acs.chemrev.5b00590>.
- Jaeckel, C., Kast, P. and Hilvert, D. (2008) *Annu Rev Biophys*, **37**, 153–173. <https://doi.org/10.1146/annurev-biochem-032807.125832>.
- Jiang, L., Althoff, E.A., Clemente, F.R. *et al.* (2008) *Science*, **319**, 1387–1391. <https://doi.org/10.1126/science.1152692>.
- Jiménez-Osés, G., Osuna, S., Gao, X. *et al.* (2014) *Nat Chem Biol*, **10**, 431–436. <https://doi.org/10.1038/nchembio.1503>.
- Kazlauskas, R.J. and Bornscheuer, U.T. (2009) *Nat Chem Biol*, **5**, 526–529. <https://doi.org/10.1038/nchembio0809-526>.
- Kourist, R., Jochens, H., Bartsch, S. *et al.* (2010) *Chembiochem*, **11**, 1635–1643. <https://doi.org/10.1002/cbic.2010000213>.
- Kuipers, R.K.P., Joosten, H.-J., Verwiél, E. *et al.* (2009) *Proteins*, **76**, 608–616. <https://doi.org/10.1002/prot.22374>.
- Kuipers, R.K., Joosten, H.-J., van Berkel, W.J.H. *et al.* (2010) *Proteins*, **78**, 2101–2113. <https://doi.org/10.1002/prot.22725>.
- Leveson-Gower, R.B., Mayer, C. and Roelfes, G. (2019) *Nat Rev Chem*, **3**, 687–705. <https://doi.org/10.1038/s41570-019-0143-x>.
- Lutz S, Bornscheuer UT. *Protein Engineering Handbook*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2008. <https://doi.org/10.1002/26>.
- Maria-Solano, M.A., Serrano-Hervás, E., Romero-Rivera, A. *et al.* (2018) *Chem Commun*, **54**, 6622–6634. <https://doi.org/10.1039/C8CC02426j>.
- Maria-Solano, M.A., Iglesias-Fernández, J. and Osuna, S. (2019) *J Am Chem Soc*, **141**, 13049–13056. <https://doi.org/10.1021/jacs.9b03646>.
- Maria-Solano, M.A., Kinatered, T., Iglesias-Fernández, J. *et al.* (2021) *ACS Catal*, **11**, 13733–13743. <https://doi.org/10.1021/acscatal.1c03950>.
- Mazurenko, S., Prokop, Z. and Damborsky, J. (2020) *ACS Catal*, **10**, 1210–1223. <https://doi.org/10.1021/acscatal.9b04321>.
- Obexer, R., Godina, A., Garrabou, X. *et al.* (2017) *Nat Chem*, **9**, 50–56. <https://doi.org/10.1038/nchem.2596>.
- Osuna, S. (2021) *Wiley Interdiscip Rev Comput Mol Sci*, **11**, e1502. <https://doi.org/10.1002/wcms.1502>.
- Packer, M.S. and Liu, D.R. (2015) *Nat Rev Genet*, **16**, 379–394. <https://doi.org/10.1038/nrg3927>.
- Pavelka, A., Chovanova, E. and Damborsky, J. (2009) *Nucleic Acids Res*, **37**, W376–W383. <https://doi.org/10.1093/nar/gkp410>.
- Petrović, D., Rizzo, V.A., Kamerlin, S.C.L. *et al.* (2018) *J R Soc Interface*, **15**, 20180330. <https://doi.org/10.1098/rsif.2018.0330>.
- Renata, H., Wang, Z.J. and Arnold, F.H. (2015) *Angew Chem Int Ed*, **54**, 3351–3367. <https://doi.org/10.1002/anie.201409470>.

- Rivalta, I., Sultan, M.M., Lee, N.-S. et al. (2012) *Proc Natl Acad Sci U S A*, **109**, E1428–E1436. <https://doi.org/10.1073/pnas.1120536109>.
- Romero, P.A. and Arnold, F.H. (2009) *Nat Rev Mol Cell Biol*, **10**, 866–876. <https://doi.org/10.1038/nrm2805>.
- Romero-Rivera, A., Garcia-Borràs, M. and Osuna, S. (2017a) <https://doi.org/10.1039/C6CC06055B>.
- Romero-Rivera, A., Garcia-Borràs, M. and Osuna, S. (2017b) *ACS Catal*, **7**, 8524–8532. <https://doi.org/10.1021/acscatal.7b02954>.
- Rothlisberger, D., Khersonsky, O., Wollacott, A.M. et al. (2008) *Nature*, **453**, 190–195. <https://doi.org/10.1038/nature06879>.
- Schupfner, M., Straub, K., Busch, F. et al. (2020) *Proc Natl Acad Sci U S A*, **117**, 346–354. <https://doi.org/10.1073/pnas.1912132117>.
- Sequeiros-Borja, C.E., Surpeta, B. and Brezovsky, J. (2020) *Brief Bioinform*, **22**, 1–15. <https://doi.org/10.1093/bib/bbaa150>.
- Sethi, A., Eargle, J., Black, A.A. et al. (2009) *Proc Natl Acad Sci U S A*, **106**, 6620–6625. <https://doi.org/10.1073/pnas.0810961106>.
- Siegel, J.B., Zanghellini, A., Lovick, H.M. et al. (2010) *Science*, **329**, 309–313. <https://doi.org/10.1126/science.1190239>.
- Stourac, J., Vavra, O., Kokkonen, P. et al. (2019) *Nucleic Acids Res*, **47**, W414–W422. <https://doi.org/10.1093/nar/gkz378>.
- Sumbalova, L., Stourac, J., Martinek, T. et al. (2018) *Nucleic Acids Res*, **46**, W356–W362. <https://doi.org/10.1093/nar/gky417>.
- Turner, N.J. (2009) *Nat Chem Biol*, **5**, 567–573. <https://doi.org/10.1038/nchembio.203>.
- Xiao, H., Bao, Z. and Zhao, H. (2015) *Ind Eng Chem Res*, **54**, 4011–4020. <https://doi.org/10.1021/ie503060a>.
- Yang, K.K., Wu, Z. and Arnold, F.H. (2019) *Nat Methods*, **16**, 687–694. <https://doi.org/10.1038/s41592-019-0496-6>.

Chapter 4:

AlphaFold2 and Deep Learning for Estimating Conformational Heterogeneity and Designing Proteins: The Case of Tryptophan Synthase

This chapter corresponds to the following publications:

Casadevall, G.; Duran, C.; Osuna, S. AlphaFold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design. *JACS Au*, **2023**, 3 (6), 1554–1562.

Casadevall, G.; Duran, C.; Estévez-Gay, M.; Osuna, S. Estimating Conformational Heterogeneity of Tryptophan Synthase with a Template-based AlphaFold2 Approach. *Prot. Sci.*, **2022**, 31 (10), e4426.

AlphaFold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design

Guillem Casadevall, Cristina Duran, and Sílvia Osuna*

Cite This: *JACS Au* 2023, 3, 1554–1562

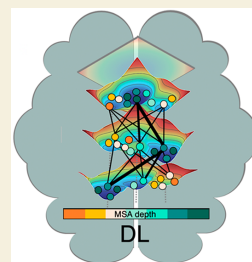
Read Online

ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: The recent success of AlphaFold2 (AF2) and other deep learning (DL) tools in accurately predicting the folded three-dimensional (3D) structure of proteins and enzymes has revolutionized the structural biology and protein design fields. The 3D structure indeed reveals key information on the arrangement of the catalytic machinery of enzymes and which structural elements gate the active site pocket. However, comprehending enzymatic activity requires a detailed knowledge of the chemical steps involved along the catalytic cycle and the exploration of the multiple thermally accessible conformations that enzymes adopt when in solution. In this Perspective, some of the recent studies showing the potential of AF2 in elucidating the conformational landscape of enzymes are provided. Selected examples of the key developments of AF2-based and DL methods for protein design are discussed, as well as a few enzyme design cases. These studies show the potential of AF2 and DL for allowing the routine computational design of efficient enzymes.



KEYWORDS: AlphaFold2, conformational heterogeneity, free energy landscape, enzyme design, deep learning

1. INTRODUCTION

The 60-year problem of knowing the folded structure from the primary sequence of proteins (and enzymes) was thought to be solved by the recent success of AlphaFold2 (AF2).^{1–3} AF2 is a deep-learning (DL) algorithm that incorporates novel neural network architectures based on the evolutionary, physical, and geometric constraints of protein structures and is able to predict with high levels of accuracy the three-dimensional structure of proteins. AF2 is recognized as one of the milestones in protein structure prediction and has boosted the application of DL methods for many other applications.⁴ Despite the impressive performance of AF2 algorithms in predicting the native lowest in energy structure of enzymes, knowing the single static folded structure is not sufficient for understanding and engineering function, as recently highlighted.^{5,6} As discussed below, another limitation of these methods is that nonprotein parts (i.e., cofactors, substrates, metal ions) are not predicted.

The three-dimensional structure of the enzymes indeed provides very relevant information on the arrangement of the catalytic machinery and structural elements gating the active-site pocket, but understanding enzymatic function requires the exploration of the ensemble of thermally accessible conformations that enzymes adopt in solution. This ensemble of conformations can be represented in the so-called Free Energy Landscape (FEL, see Figure 1 for FELs at different reaction stages),⁷ which displays the relative stabilities of the thermally accessible conformations, as well as the kinetic barriers separating them. Conformational changes that can directly

impact catalytic function include side-chain conformational changes in the fast time scale, loop motions often playing a key role in substrate binding/product release in slower time scales, and in some cases allosteric transitions that usually correspond to the slowest processes. The evaluation of the conformational landscapes of natural and evolved enzymes has provided relevant new insights. Experimental X-ray structures and associated B-factors,⁸ room-temperature X-ray experiments,^{9,10} and NMR experiments¹¹ have been used to explore the changes in the conformational landscape induced by mutations along several enzyme variants generated with the experimental Directed Evolution technique. From a computational perspective, the reconstruction of the FEL and how this is shifted after mutation provides crucial information for understanding enzyme function (and also for design).⁷

It has been recently shown by different groups that AF2 can be actually tuned to provide multiple conformations of the same protein, which suggests the potential of AF2 for elucidating the conformational landscape of enzymes and proteins.^{12,13} Given the rather low computational cost of AF2, especially if compared to the computationally demanding

Received: April 14, 2023

Revised: May 22, 2023

Accepted: May 22, 2023

Published: June 6, 2023



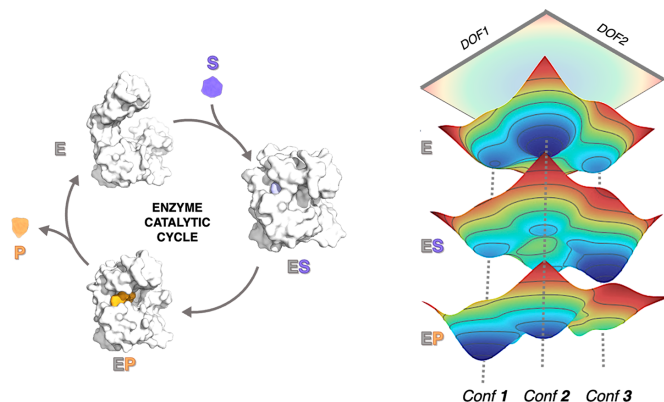


Figure 1. Schematic representation of a catalytic cycle of a model enzyme and associated conformational changes represented in the Free Energy Landscape (FEL) at the different steps: free enzyme (E), enzyme–substrate (ES), and enzyme–product (EP). For FEL reconstruction some key degrees of freedom (DOF) need to be defined, as explained in section 3.

Molecular Dynamics (MD) simulations, its application for assessing the effect of mutations on the conformational landscape is highly appealing. This could impact the development of AF2-based conformationally focused enzyme designs protocols.^{7,14}

Multiple reviews are available in the literature covering the available tools for rationalizing the changes in activity induced by mutations in several enzymes^{15–17} and for rationally designing novel enzymes by means of computational protocols based on Quantum Mechanics (QM), hybrid QM and Molecular Mechanics (QM/MM), Empirical Valence Bond (EVB), MD, and Monte Carlo simulations, or combinations of them.¹⁴ Instead, in this Perspective we will cover a few of the most recent applications of DL strategies for elucidating enzyme conformational flexibility and for its application for enzyme design.

2. THE ROLE OF CONFORMATIONAL DYNAMICS IN ENZYME FUNCTION AND EVOLUTION

Enzymes present highly preorganized active-site pockets with the catalytic residues perfectly arranged for efficiently stabilizing the transition state(s) of a specific reaction.^{18,19} This preorganization is complemented by the enzyme ability to adopt multiple conformations of importance for substrate binding and/or product release. The importance of conformational flexibility was clearly shown with the design of catalytic antibodies presenting an ideal complementary structure to the transition state, which showed a clearly inferior catalytic activity with respect to enzymes.²⁰ This shows that efficient catalysis requires not only transition-state stabilization but also the optimization of the conformational ensemble.^{14,21} In fact, the ability of enzymes to adapt and evolve toward novel functions either by natural or laboratory evolution has been connected to their inherent conformationally rich dynamic nature.^{7,22–25} Enzymes display a high degree of flexibility and versatility as shown by their promiscuous side activities²⁶ and also their tolerance to evolve toward novel functions.^{7,22–25}

Along the enzymatic cycle the following steps take place: (1) first, the substrate(s) bind to the catalytic pocket, which often

require and/or induce a change in the conformation of loops and flexible domains regulating the access to the active site;^{27,28} (2) the substrate(s) are activated to facilitate productive formation of the Enzyme–Substrate (ES) complex; (3) this is followed by the stabilization of the transition state(s) for the formation of multiple reaction intermediates and product(s); (4) finally, once the Enzyme–Product (EP) complex is formed the product(s) are released from the pocket, which is often accompanied by conformational changes that initiate the next round of the catalytic cycle. All of these steps are essential for maximizing catalytic activity by optimized throughput of the overall pathway. The binding of the substrate for ES formation can also modulate the conformational landscape as shown for the multienzyme complex pyruvate dehydrogenase complex.²⁹ In Figure 1, the conformational changes that take place along the catalytic itinerary of the enzyme adenylate kinase (AdK) is shown as a model. The catalytic cycle involves the conformational change from open to closed structures of a lid that covers the active site. The computational evaluation of the chemical steps along the catalytic itinerary (steps 2 and 3) require the use of QM, hybrid QM/MM, and EVB, which are too expensive to be applied for analyzing the conformational changes taking place through the cycle and the processes of substrate binding and product release (steps 1 and 4).^{7,14,15} This explains the large available number of computational approaches developed along the years. Current computational strategies put mostly the focus to only some of the above-mentioned features, in part explaining the often low success in achieving high levels of enzymatic activity.¹⁴

3. COMPUTATIONAL RECONSTRUCTION OF THE FREE ENERGY LANDSCAPE

The ensemble of conformations that enzymes adopt in solution can be represented in the free energy landscape (FEL). The free energy (G) is proportional to the negative logarithm of the population distribution in $k_B T$ units; thus, a maximum in this distribution is a minimum in the FEL. The FEL therefore provides crucial information on the thermodynamics (i.e.,

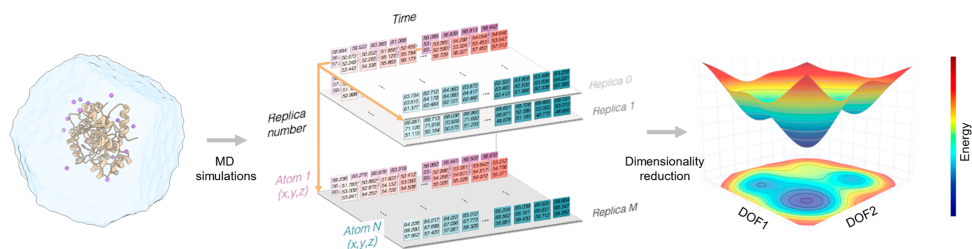


Figure 2. Schematic representation of the Free Energy Landscape (FEL) reconstruction process. The high dimensional data of the MD simulations needs to be reduced and projected into a set of key collective degrees of freedom (DOF) for probability distribution calculation to reconstruct the FFEL.

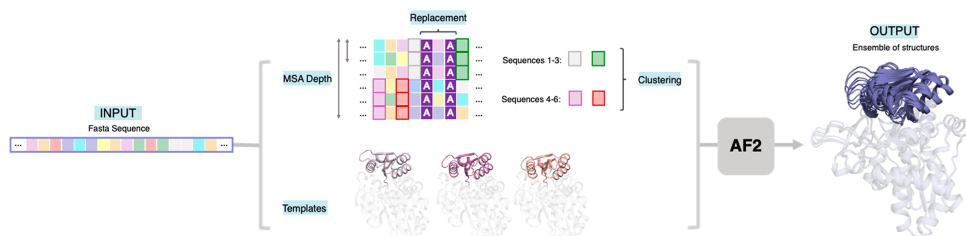


Figure 3. Overview of strategies developed for predicting alternate states with AlphaFold2 (AF2). As done in del Alamo et al.¹² the Multiple Sequence Alignment (MSA) depth can be altered, some of the MSA positions can be masked as shown by Stein and McHaourab,¹³ and the MSA can be clustered as in the Kern and Ovchinnikov preprint paper.³⁷ The provided set of templates can also be changed as done in some cases in del Alamo et al.¹² and also in our recent publication.⁴⁴

which are the lowest in energy conformations at a given set of conditions) and the kinetics for the conformational transitions. These energy barriers separating the different minima will determine the time scale of the conformational exchange: fast conformational changes occur in the picosecond to microsecond time scales (this is the case of loop motions crucial for enzyme catalysis), whereas slow motions will take place in millisecond to seconds.

Enzymes can be captured in different conformational states by means of X-ray, room-temperature, and time-resolved X-ray, cryo-EM, NMR, and biophysical techniques can be applied for providing complementary kinetic information.³⁰ These multiple conformations of the same enzyme deposited in the protein data bank (PDB) played an important role in AF2 training but also for the AF2 application for assessing the conformational heterogeneity of biological systems (as discussed below). Computational methods are particularly appropriate for reconstructing the FEL: MD simulations sample the population distribution by integrating Newton's laws of motion. By defining a reduced set of collective degrees of freedom (DOFs) the high dimensional data obtained in the MD runs can be projected for probability distribution calculation and thus FEL reconstruction (see Figure 2). The selection of the reduced set of DOFs can be made manually or automatically by means of different dimensionality reduction schemes.^{7,31} The accurate exploration of the conformational changes for FEL reconstruction requires extensive MD simulations, and depending on the time scale of the conformational transitions enhanced sampling techniques need to be applied.^{7,14,15} These techniques have a high

computational cost associated with them (from weeks to many months of simulations), which limits the applicability of these strategies for computationally designing and ranking enzyme designs.

4. APPLICATION OF AF2 FOR CAPTURING CONFORMATIONAL HETEROGENEITY

The standard AF2 protocol requires the primary sequence of the enzyme, a multiple sequence alignment (MSA) generated with information on evolutionary related proteins, and the 3D coordinates of a small number of homologous structures named templates (see Figure 3). Although AF2 was designed to predict single *static* structures, some recent papers have shown that by reducing the depth of the input MSAs used in the AF2 algorithm (in addition to decrease the number of recycles) accurate models in multiple conformations can be generated.^{12,13,32} In particular, del Alamo and coworkers showed that multiple conformations of transporters and G-protein-coupled receptors can be obtained by altering the AF2 pipeline and providing a reduced number of MSA sequences (as low as 16 sequences only).¹² They generated up to 50 different models of each protein receptor for each MSA size, as opposed to the standard AF2 protocol that provides conformationally homogeneous and nearly identical models. Interestingly, they observed limited conformational sampling for proteins that were contained in the AF2 training set. In another study, Stein and McHaourab reported a universal method for biasing the models generated by AF2 based on the replacement of specific residues within the MSA to alanine or another residue.¹³ AF2 was used to generate initial models, and

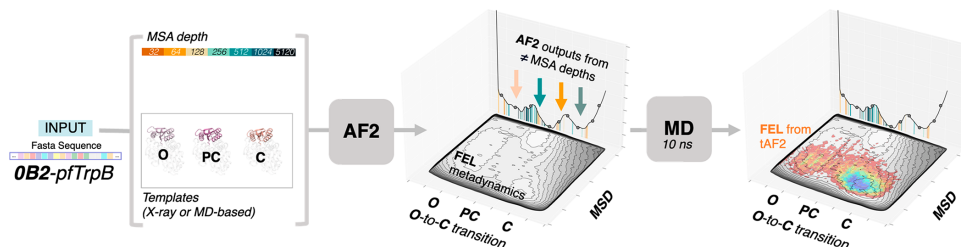


Figure 4. Our template-based AF2 (tAF2) approach for estimating the conformational heterogeneity. Different Multiple Sequence Alignment (MSA) depths and set of templates taken from either selected X-ray structures or Molecular Dynamics (MD) snapshots are provided to AF2.³⁴ The multiple output models generated by AF2 at the different MSA depths shown as vertical lines in the central plot are then subjected to short MD simulations for FEL reconstruction. The new FEL generated from the 10 ns MD simulations starting from the ca. 1000 AF2 outputs at different MSA is shown in a blue to red colormap on top of the computationally reconstructed FEL obtained via well-tempered multiple-walker metadynamics simulations (in gray).^{33,34} The x and y axis of the reconstructed FELs indicate the Open-to-Closed (O-to-C) transition of the COMM domain of TrpB that covers the active site, and the Mean Square Deviation (MSD) from the path of the generated O-to-C structures, respectively.^{33–35} The input sequence is the 0B2-*pf*TrpB variant.³⁸ Reproduced with permission from ref 34. Copyright 2022 John Wiley & Sons, Inc.

the MSA was modified based on possible contact points in the initial structures, prior structural information, or regions of uncertainty within the main structure. They found that the replacement of certain amino acid columns to alanine or other residues turns the attention of the network to other parts of the MSA allowing for AF2 to find alternative conformations based on other coevolved residues. One of the provided examples is AdK that undergoes a large-scale conformational change of a lid and a flap that gate the crystal site, as revealed by the unbound and inhibitor-bound active structures (see Figure 1). By masking some residues and replacing them to alanine, closed and open structures of AdK were obtained. Although none of the AF2 open structures reached the level of opening of the crystal structure (PDB: 4AKE), the set of generated AF2 models displaying a different level of closure showed the potential of the methodology for predicting alternate conformations describing the conformational heterogeneity of the systems.

Inspired by these previous publications showing AF2's ability to sample additional conformational states, we developed a template-based AF2 approach to assess the conformational heterogeneity and how this is altered by mutations on the β subunit of several tryptophan synthase enzymes (TrpB, see Figure 4).^{33–35} As done in the work of del Alamo et al.,¹² we tested the effect of reducing the provided number of sequences in the MSA, but we additionally assessed how AF2 predictions are altered when different templates displaying multiple conformational states are provided.³⁴ We tested the template-based AF2 pipeline by providing either X-ray based or conformations extracted from MD simulations as templates. With these settings AF2 revealed major differences in the conformational landscapes among the analyzed systems. Interestingly, this was further demonstrated by running multiple short MD simulations from the set of AF2 structures and reconstructing the associated FELs (Figure 4). The comparison of the generated FEL from the template-based AF2 predictions were in line with the computationally expensive FELs generated with well-tempered multiple-walker metadynamics simulations. This is exciting as it shows the potential of AF2 for rapidly and accurately assessing the FELs of different systems, which could be applied for conforma-

tionally driven enzyme design approaches.^{7,14} The multiple outputs obtained via AF2 at different MSA depths were also recently combined with Reweighted Autoencoded Variational Bayes for Enhanced (RAVE) sampling.³⁶

In a recent preprint paper, Kern and Ovchinnikov showed that clustering the input MSA by sequence similarity allows AF2 to visit multiple conformational states of some metamorphic proteins known to display large conformational changes.³⁷ They also identified two mutations that according to AF2 predictions could switch the circadian rhythm protein KaiB between the two major conformational states. Their developed methodology was also applied for searching for alternative conformational states in other protein families and found a putative alternate state for the oxidoreductase DsBE. These computational predictions were, however, not tested experimentally in the published preprint paper.

5. APPLICATION OF AF2 AND OTHER DEEP LEARNING TECHNIQUES FOR PROTEIN AND ENZYME DESIGN

Inspired by the AF2 approach, other DL techniques have been recently developed for elucidating the folded structure of enzymes and providing some metrics to be potentially used for protein design. The field is advancing fast, and the number of DL strategies developed especially for protein design is constantly increasing. In this section, we aim to provide a brief overview of the most representative techniques developed, and we put special emphasis to those strategies that are particularly relevant for enzyme design. Some recent reviews focused on structure-based protein design with DL strategies are available,^{39,40} as well as a review related to the design of more stable enzymes.⁴¹

These available strategies for structure prediction can be classified depending on the number of input parameters used: those that require the input query sequence, MSA, and set of templates for accurate predictions and those that predict the folded structure based on the input sequence only. Similarly to AF2, the RoseTTAFold (RF) algorithm developed almost at the same time as AF2 requires an MSA and a set of initial templates to make accurate predictions of the folded structure. RF showed improved accuracy toward protein–protein

complex prediction as compared to AF2 and AF2 multimer.^{42,43} OpenFold2 was also developed to replicate the AF2 algorithm and make it accessible to the structural biology community.⁴⁴ AlphaLink was also introduced to incorporate experimental distance restraint information, thus generating a modified version of the AF2 network architecture.⁴⁵

Sequences contain implicit information about the enzyme structure and function, as the position of each amino acid in the sequence is determined by the spatial arrangement and the possible interactions established between them. The main advantage is the comparison of sequences is computationally cheap (at least as compared to physics-based approaches) and provides crucial information about the most frequent residues at each position, conservation score, and correlated mutation pairs that have emerged during evolution. Covarying mutations have been associated with function, tertiary contacts, and binding. It has also been shown that the use of language models previously used for Natural Language Processing (NLP) could be applied in the context of the biology language to generate “content-aware” data representations from large-scale sequence data sets. This is the case of ESM-2, which corresponds to the largest language model of protein sequences developed to date.⁴⁶ ESMFold was then developed, which was found to perform end-to-end folded structure predictions with similar accuracies to AF2, *albeit* at an order of magnitude faster.⁴⁶ OmegaFold (OF) is another end-to-end structure prediction algorithm developed that combines a pretrained language model and a geometrical transformer model for reconstructing the structure.⁴⁷ Similarly to ESMfold, OF only needs the input sequence and is 10-fold faster than AF2 and RF. More importantly, OF was found to do a better job in predicting the folded structure of orphan proteins, i.e., those proteins that do not have any assigned functional family.

Apart from the different methodologies developed to predict the folded structure of proteins, different NLP and deep-learning architectures have been developed to generate new non-natural sequences. These different strategies have targeted different objectives that range from generating new sequences for maintaining some natural activities^{48,49} to imagining new folds and sophisticated symmetric assemblies,⁵⁰ among others.

The generative language models ProGen and ProGen2 trained on millions of raw protein sequences were developed to generate de novo artificial proteins that express well and maintain enzymatic function.^{51,52} ProtGPT2 is an unsupervised language model that can generate new sequences based on the principles of natural ones.⁵³ Similarly, variational autoencoders trained on a data set of luciferase-like oxidoreductases were also used to generate new sequences maintaining the luciferase activity.⁴⁹ ProteinGAN, which is based on a self-attention-based variant of the generative adversarial network, learns natural protein sequences for generating new functional variants.⁵⁴ The conditional language model ZymCTRL trained on the BRENDA database of enzymes has also been recently developed, which is able to provide new artificial enzymes within a user-defined Enzyme Classification (EC)-based enzymatic class.⁵⁵ Language models have also been used to obtain a set of sequences that are likely to fold into a given desired structure. This is, for instance, the case for recently developed LM-Design⁵⁶ and ProteinDT.⁵⁷ Yu and co-workers have recently developed CLEAN based on contrastive learning that is able to assign EC numbers to a given sequence.⁵⁸

The transform-restrained Rosetta (trRosetta) was developed by the Baker lab in 2020 to design a variety of proteins by randomly modifying the starting sequences to find sharply predicted residue–residue interdistances.⁵⁹ The combination of trRosetta and the physics-based Rosetta was shown to provide more funneled energy landscapes: trRosetta was used to disfavor alternative states, and high-resolution Rosetta was used for creating a deep energy minimum at the designed target structure.⁶⁰ Small β barrel proteins and proteins with discontinuous functional sites were developed with trRosetta.^{61,62} Recently, Dauparas and co-workers developed a method called ProteinMPNN, which is a graph neural network that was found to rescue previously failed designs targeted with Rosetta or AF2.⁶³ ProteinMPNN was recently applied to generate de novo luciferases.⁶⁴ MutComput is a convolutional neural network (CNN) that was successfully applied for designing new hydrolases for poly(ethylene terephthalate) depolymerization.⁶⁵ Another more recent CNN for protein sequence design was provided by Anand et al. to generate a de novo TIM-barrel protein backbone.⁶⁶ Holographic CNNs have also been developed to learn the shape of protein micro-environments to predict the impact of mutations on stability and binding of protein complexes.⁶⁷

Different protocols based on the use of AF2 for predicting the structure of the generated sequences and use the output AF2 metrics for the design of new proteins have also been developed. The AlphaDesign computational framework was constructed to enable the rapid prediction of completely novel protein monomers starting from random sequences.⁶⁸ The potential application of AlphaDesign for designing proteins that bind to prespecific target proteins was also shown. AF2 was also used for the rapid and accurate fixed backbone design of sequences that are strongly predicted to fold to a specific backbone.⁶⁹ The Baker lab combined ProteinMPNN with AF2 to design closed repeat proteins with central pockets⁵⁰ and generate symmetric protein assemblies.⁵⁰ Similarly, RF instead of AF2 was used for designing high-affinity protein binders⁷⁰ or proteins with prespecified functional motifs.⁷¹ RF has also the potential to predict the effect of mutations on protein function.⁷²

The RF-based diffusion model (named *RFdiffusion*) has been recently developed by the Baker lab.⁷³ *RFdiffusion* can very rapidly and accurately design topology-constrained protein monomers, protein binders, symmetric oligomers, metal-binding proteins, and even enzyme scaffolds containing specific active-site residues.⁷³ The performance of *RFdiffusion* outperforms hallucination in terms of success rate, accuracy, and speed. Even though *RFdiffusion* does not explicitly consider the substrate molecule, it can be implicitly modeled using an external potential to guide the generation of the active-site pocket.

As mentioned in the Introduction, catalytic function requires substrate binding and product release, and in many cases enzymatic activity is dependent on cofactor and metal ion binding. In this direction, different strategies based on DL have also been generated to dock ligands, substrates, and missing cofactors into potential pockets. AlphaFill uses sequence and structural similarity to include the missing organic molecules and metal ions into the AF2 models.⁷⁴ The diffusion generative model DiffDock was designed to dock small molecules into potential protein pockets. This strategy was shown to outperform previous traditional and DL docking protocols.⁷⁵ Meller and co-workers also developed an AF2-based strategy to

find cryptic pockets.⁷⁶ DL has also been applied for finding potential location sites of transition metals in proteins (Metal1D and Metal3D).⁷⁷ The coevolution based MetalNet pipeline has also been recently created to predict potential metal-binding sites.⁷⁸

6. OUTLOOK AND FUTURE PERSPECTIVES

Enzyme catalysis is a complex multidimensional process that requires the optimal sequence and structure for allowing substrate(s) binding, catalyzing the chemical steps and product(s) release, and optimizing the multiple conformations needed for developing its function. This high complexity makes the task of enzyme design, especially toward non-natural reactions or substrates in high efficiencies very challenging. The selected examples highlighted in this review show the potential of DL techniques to generate new functional variants mostly within the allowed biological constraints of the sequence space. The application of DL strategies for computational enzyme design for any target reaction and non-natural substrate is only at its beginning. For many years, the lack of precision in incorporating the desired active-site residues into protein scaffolds in computational enzyme design has been considered one of the many limitations of the overall process. This point, however, seems to be solved with the recent RosettaFold-based diffusion model developed by the Baker lab. The incorporation of the QM-based models of the enzyme active site into new non-natural scaffolds specifically designed to hold the functional motifs in place might no longer be the limiting factor, but instead predicting which scaffolds might be more appropriate for the optimization of the conformational ensemble for efficient catalysis will most likely be essential. Considering the huge advances especially in the field of structure prediction and protein design seen these recent years, the combination of DL methods with physics-based approaches will play a key role the coming years for finding optimal solutions for the rational and routine design of highly efficient and stable enzymes for non-natural reactions and substrates.

AUTHOR INFORMATION

Corresponding Author

Silvia Osuna – Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, 17003 Girona, Spain; ICREA, 08010 Barcelona, Spain; orcid.org/0000-0003-3657-6469; Email: silvia.osuna@udg.edu

Authors

Guillem Casadevall – Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, 17003 Girona, Spain

Cristina Duran – Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, 17003 Girona, Spain

Complete contact information is available at: <https://pubs.acs.org/10.1021/jacsau.3c00188>

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. CRediT: Guillem Casadevall writing-review &

editing; Cristina Duran writing-review & editing; Silvia Osuna writing-review & editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the Generalitat de Catalunya for the consolidated group TCBioSys (SGR 2021 00487) and grant projects PID2021-129034NB-I00 and PDC2022-133950-I00 funded by Spanish MICIN. 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, ERC-2022-POC-10112805, and ERC-2022-CoG-101088032) and the Human Frontier Science Program (HFSP) for project grant RGP0054/2020. G.C. was supported by a research grant from ERC-StG (ERC-2015-StG-679001) and HFSP RGP0054/2020. C.D. was supported by the Spanish MINECO for a PhD fellowship (PRE2019-089147).

ABBREVIATIONS

AF2, AlphaFold2; DL, Deep Learning; FEL, Free Energy Landscape; MD, Molecular Dynamics

REFERENCES

- (1) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Židek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577* (7792), 706–710.
- (2) Ourmazd, A.; Moffat, K.; Lattman, E. E. Structural biology is solved — now what? *Nat. Methods* **2022**, *19* (1), 24–26.
- (3) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.
- (4) Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **2020**, *588* (7837), 203–204.
- (5) Clementi, C. Fast track to structural biology. *Nat. Chem.* **2021**, *13* (11), 1032–1034.
- (6) Jones, D. T.; Thornton, J. M. The impact of AlphaFold2 one year on. *Nat. Methods* **2022**, *19* (1), 15–20.
- (7) 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* (50), 6622–6634.
- (8) Campbell, E.; Kaltenbach, M.; Correy, G. J.; Carr, P. D.; Porebski, B. T.; Livingstone, E. K.; Afriat-Jumou, L.; Buckle, A. M.; Weik, M.; Hoffelder, F.; Tokuriki, N.; Jackson, C. J. The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* **2016**, *12* (11), 944–950.
- (9) Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T. Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (39), 16247–16252.
- (10) Broom, A.; Rakotoharisoa, R. V.; Thompson, M. C.; Zarifi, N.; Nguyen, E.; Mukhametzhano, N.; Liu, L.; Fraser, J. S.; Chica, R. A. Ensemble-based enzyme design can recapitulate the effects of

- laboratory directed evolution in silico. *Nat. Commun.* **2020**, *11* (1), 4808.
- (11) Otten, R.; Pádua, R. A. P.; Bunzel, H. A.; Nguyen, V.; Pitsawong, W.; Patterson, M.; Sui, S.; Perry, S. L.; Cohen, A. E.; Hilvert, D.; Kern, D. How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. *Science* **2020**, *370* (6523), 1442–1446.
- (12) del Alamo, D.; Sala, D.; McHaourab, H. S.; Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **2022**, *11*, e75751.
- (13) Stein, R. A.; McHaourab, H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLoS Comput. Biol.* **2022**, *18* (8), e1010483.
- (14) Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, e1502.
- (15) Romero-Rivera, A.; Garcia-Borràs, M.; Osuna, S. Computational tools for the evaluation of laboratory-engineered biocatalysts. *Chem. Commun.* **2017**, *53* (2), 284–297.
- (16) Himo, F.; de Visser, S. P. Status report on the quantum chemical cluster approach for modeling enzyme reactions. *Commun. Chem.* **2022**, *5* (1), 29.
- (17) Świderek, K.; Tuñón, I.; Moliner, V. Predicting enzymatic reactivity: from theory to design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (5), 407–421.
- (18) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.* **2006**, *106* (8), 3210–3235.
- (19) Marti, S.; Roca, M.; Andres, J.; Moliner, V.; Silla, E.; Tunon, I.; Bertran, J. Theoretical insights in enzyme catalysis. *Chem. Soc. Rev.* **2004**, *33* (2), 98–107.
- (20) Winkler, C. K.; Schrittwieser, J. H.; Kroutil, W. Power of Biocatalysis for Organic Synthesis. *ACS Cent. Sci.* **2021**, *7* (1), 55–71.
- (21) Lovelock, S. L.; Crawshaw, R.; Basler, S.; Levy, C.; Baker, D.; Hilvert, D.; Green, A. P. The road to fully programmable protein catalysis. *Nature* **2022**, *606* (7912), 49–58.
- (22) Tokuriki, N.; Tawfik, D. S. Protein Dynamism and Evolvability. *Science* **2009**, *324* (5924), 203–207.
- (23) Petrović, D.; Risso, V. A.; Kamerlin, S. C. L.; Sanchez-Ruiz, J. M. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* **2018**, *15* (144), 20180330.
- (24) Campbell, E. C.; Correy, G. J.; Mabbitt, P. D.; Buckle, A. M.; Tokuriki, N.; Jackson, C. J. Laboratory evolution of protein conformational dynamics. *Curr. Opin. Struct. Biol.* **2018**, *50*, 49–57.
- (25) Crean, R. M.; Gardner, J. M.; Kamerlin, S. C. L. Harnessing Conformational Plasticity to Generate Designer Enzymes. *J. Am. Chem. Soc.* **2020**, *142* (26), 11324–11342.
- (26) Kherosky, O.; Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **2010**, *79*, 471–505.
- (27) Boehr, D. D.; Nussinov, R.; Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5* (11), 789–796.
- (28) Hammes, G. G.; Benkovic, S. J.; Hammes-Schiffer, S. Flexibility, Diversity, and Cooperativity: Pillars of Enzyme Catalysis. *Biochem.* **2011**, *50* (48), 10422–10430.
- (29) Prajapati, S.; Haselbach, D.; Wittig, S.; Patel, M. S.; Chari, A.; Schmidt, C.; Stark, H.; Tittmann, K. Structural and Functional Analyses of the Human PDH Complex Suggest a “Division-of-Labor” Mechanism by Local E1 and E3 Clusters. *Structure* **2019**, *27* (7), 1124–1136.
- (30) Baldwin, A. J.; Kay, L. E. NMR spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.* **2009**, *5* (11), 808–814.
- (31) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121* (16), 9722–9758.
- (32) Roney, J. P.; Ovchinnikov, S. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Phys. Rev. Lett.* **2022**, *129* (23), 238101.
- (33) 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* (33), 13049–13056.
- (34) Casadevall, G.; Duran, C.; Estévez-Gay, M.; Osuna, S. Estimating conformational heterogeneity of tryptophan synthase with a template-based AlphaFold2 approach. *Protein Sci.* **2022**, *31* (10), e4426.
- (35) Maria-Solano, M. A.; Kinatader, T.; Iglesias-Fernández, J.; Sterner, R.; Osuna, S. In Silico Identification and Experimental Validation of Distal Activity-Enhancing Mutations in Tryptophan Synthase. *ACS Catal.* **2021**, *11* (21), 13733–13743.
- (36) Vani, B. P.; Aranganathan, A.; Wang, D.; Tiwary, P. AlphaFold2-RAVE: From Sequence to Boltzmann Ranking. *J. Chem. Theory Comput.* **2023**, DOI: 10.1021/acs.jctc.3c00290.
- (37) Wayment-Steele, H. K.; Ovchinnikov, S.; Colwell, L.; Kern, D. Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. *bioRxiv* **2022**, DOI: 10.1101/2022.10.17.512570.
- (38) Buller, A. R.; Brinkmann-Chen, S.; Romney, D. K.; Herger, M.; Murciano-Calles, J.; Arnold, F. H. Directed evolution of the tryptophan synthase β -subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112* (47), 14599–14604.
- (39) Ovchinnikov, S.; Huang, P.-S. Structure-based protein design with deep learning. *Curr. Opin. Chem. Biol.* **2021**, *65*, 136–144.
- (40) Gao, W.; Mahajan, S. P.; Sulam, J.; Gray, J. J. Deep Learning in Protein Structural Modeling and Design. *Patterns* **2020**, *1* (9), 100142.
- (41) Ming, Y.; Wang, W.; Yin, R.; Zeng, M.; Tang, L.; Tang, S.; Li, M. A review of enzyme design in catalytic stability by artificial intelligence. *Brief. Bioinform.* **2023**, 1–19.
- (42) Baek, M.; Baker, D. Deep learning and protein structure modeling. *Nat. Methods* **2022**, *19* (1), 13–14.
- (43) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Dijk, A. A. v.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373* (6557), 871–876.
- (44) Ahdritz, G.; Bouatta, N.; Kadyan, S.; Xia, Q.; Gerecke, W.; O'Donnell, T. J.; Berenberg, D.; Fisk, I.; Zanichelli, N.; Zhang, B.; Nowaczynski, A.; Wang, B.; Stepniowska-Dziubinska, M. M.; Zhang, S.; Ojewole, A.; Guney, M. E.; Biderman, S.; Watkins, A. M.; Ra, S.; Lorenzo, P. R.; Nivon, L.; Weitzner, B.; Ban, Y.-E. A.; Sorger, P. K.; Mostaque, E.; Zhang, Z.; Bonneau, R.; AlQurashi, M. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv* **2022**, DOI: 10.1101/2022.11.20.517210.
- (45) Stahl, K.; Graziadei, A.; Dau, T.; Brock, O.; Rappsilber, J. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nat. Biotechnol.* **2023**, DOI: 10.1038/s41587-023-01704-z.
- (46) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Costa, A. d. S.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379* (6637), 1123–1130.
- (47) Wu, R.; Ding, F.; Wang, R.; Shen, R.; Zhang, X.; Luo, S.; Su, C.; Wu, Z.; Xie, Q.; Berger, B.; Ma, J.; Peng, J. High-resolution de novo structure prediction from primary sequence. *bioRxiv* **2022**, DOI: 10.1101/2022.07.21.500999.
- (48) Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; Fraser, J. S.; Naik, N. Deep neural language modeling enables

functional protein generation across families. *bioRxiv* **2021**, DOI: [10.1101/2021.07.18.452833](https://doi.org/10.1101/2021.07.18.452833).

(49) Hawkins-Hooker, A.; Depardieu, F.; Baur, S.; Couairon, G.; Chen, A.; Bikard, D. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.* **2021**, *17* (2), e1008736.

(50) Wicky, B. I. M.; Milles, L. F.; Courbet, A.; Ragotte, R. J.; Dauparas, J.; Kinfu, E.; Tipps, S.; Kibler, R. D.; Baek, M.; DiMaio, F.; Li, X.; Carter, L.; Kang, A.; Nguyen, H.; Bera, A. K.; Baker, D. Hallucinating symmetric protein assemblies. *Science* **2022**, *378* (6615), 56–61.

(51) Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; Fraser, J. S.; Naik, N. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **2023**, DOI: [10.1038/s41587-022-01618-2](https://doi.org/10.1038/s41587-022-01618-2).

(52) Nijkamp, E.; Ruffolo, J.; Weinstein, E. N.; Naik, N.; Madani, A. ProGen2: Exploring the Boundaries of Protein Language Models. *arXiv* **2022**, DOI: [10.48550/arXiv.2206.13517](https://doi.org/10.48550/arXiv.2206.13517).

(53) Ferruz, N.; Schmidt, S.; Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **2022**, *13* (1), 4348.

(54) Repecka, D.; Jauniskis, V.; Karpus, L.; Rembeza, E.; Rokaitis, I.; Zrimec, J.; Poviloniene, S.; Laurynenas, A.; Viknander, S.; Abuajwa, W.; Savolainen, O.; Meskys, R.; Engqvist, M. K. M.; Zelezniak, A. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **2021**, *3* (4), 324–333.

(55) Munsamy, G.; Lindner, S.; Lorenz, P.; Ferruz, N. Zymctrl: a conditional language model for the controllable generation of artificial enzymes. In *NeurIPS Machine Learning in Structural Biology Workshop*, 2022.

(56) Zheng, Z.; Deng, Y.; Xue, D.; Zhou, Y.; YE, F.; Gu, Q. Structure-informed Language Models Are Protein Designers. *arXiv* **2023**, DOI: [10.48550/arXiv.2302.01649](https://doi.org/10.48550/arXiv.2302.01649).

(57) Liu, S.; Zhu, Y.; Lu, J.; Xu, Z.; Nie, W.; Gitter, A.; Xiao, C.; Tang, J.; Guo, H.; Anandkumar, A. A Text-guided Protein Design Framework. *arXiv* **2023**, DOI: [10.48550/arXiv.2302.04611](https://doi.org/10.48550/arXiv.2302.04611).

(58) Yu, T.; Cui, H.; Li, J. C.; Luo, Y.; Jiang, G.; Zhao, H. Enzyme function prediction using contrastive learning. *Science* **2023**, *379* (6639), 1358–1363.

(59) Anishchenko, I.; Pellock, S. J.; Chidyausiku, T. M.; Ramelot, T. A.; Ovchinnikov, S.; Hao, J.; Bafna, K.; Norn, C.; Kang, A.; Bera, A. K.; DiMaio, F.; Carter, L.; Chow, C. M.; Montelione, G. T.; Baker, D. De novo protein design by deep network hallucination. *Nature* **2021**, *600* (7889), 547–552.

(60) Norn, C.; Wicky, B. I. M.; Juergens, D.; Liu, S.; Kim, D.; Tischer, D.; Koepnick, B.; Anishchenko, I.; Baker, D.; Ovchinnikov, S.; Coral, A.; Bubar, A. J.; Boykov, A.; Pérez, A. U. V.; MacMillan, A.; Lubow, A.; Mussini, A.; Cai, A.; Ardill, J.; Seal, A.; Kalantarian, A.; Failer, B.; Lackertsen, B.; Chagot, B.; Haight, B. R.; Taştan, B.; Uitham, B.; Roy, B. G.; Cruz, B. R. d. M.; Echols, B.; Lorenz, B. E.; Blair, B.; Kestemont, B.; Eastlake, C. D.; Bragdon, C. J.; Vardeman, C.; Salerno, C.; Comisky, C.; Hayman, C. L.; Landers, C. R.; Zimov, C.; Coleman, C. D.; Painter, C. R.; Ince, C.; Lynagh, C.; Malania, D.; Wheeler, D. C.; Robertson, D.; Simon, V.; Chisari, E.; Kai, E. L. J.; Rezae, F.; Lengyel, F.; Tabotta, F.; Padelletti, F.; Boström, F.; Gross, G. O.; McIlvaine, G.; Beecher, G.; Hansen, G. T.; Jong, G. d.; Feldmann, H.; Borman, J. L.; Quinn, J.; Norrgard, J.; Truong, J.; Diderich, J. A.; Canfield, J. M.; Photakis, J.; Slone, J. D.; Madzio, J.; Mitchell, J.; Stomierowski, J. C.; Mitch, J. H.; Altenbeck, J. R.; Schinkler, J.; Weinberg, J. B.; Burbach, J. D.; Costa, J. C. S. d.; Juarez, J. F. B.; Gunnarsson, J. P.; Harper, K. D.; Joo, K.; Clayton, K. T.; DeFord, K. E.; Scully, K. F.; Gildea, K. M.; Abbey, K. J.; Kohli, K. L.; Stenner, K.; Takács, K.; Poussaint, L. L.; Manalo, L. C.; Withers, L. C.; Carlson, L.; Wei, L.; Fisher, L. R.; Carpenter, L.; Ji-hwan, M.; Ricci, M.; Belcastro, M. A.; Leniec, M.; Hohmann, M.; Thompson, M.; Thayer, M. A.; Gaebel, M.; Cassidy, M. D.; Fagiola, M.; Lewis, M.; Pfützenreuter, M.; Simon, M.; Elmassy, M. M.; Benevides, N.; Kerr, N. K.; Verma, N.; Shannon, O.; Yin, O.; Wolfteich, P.; Gummersall, P.; Thušcik, P.; Gajar, P.; Triggiani, P. J.; Guha, R.; Innes,

R. B. M.; Buchanan, R.; Gamble, R.; Leduc, R.; Spearing, R.; Gomes, R. L. C. d. S.; Estep, R. D.; DeWitt, R.; Moore, R.; Shneider, S. G.; Zaccanelli, S. J.; Kuznetsov, S.; Burillo-Sanz, S.; Mooney, S.; Vasily, S.; Butkovich, S. S.; Hudson, S. B.; Pote, S. L.; Denne, S. P.; Schwegmann, S. A.; Ratna, S.; Kleinfelder, S. C.; Bausewein, T.; George, T. J.; Almeida, T. S. d.; Yeginer, U.; Barnettler, W.; Pulley, W. R.; Wright, W. S.; Willlyanto; Lansford, W.; Hochart, X.; Gaiji, Y. A. S.; Lagodich, Y.; Christian, V. Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (11), e2017228118.

(61) Kim, D. E.; Jensen, D. R.; Feldman, D.; Tischer, D.; Saleem, A.; Chow, C. M.; Li, X.; Carter, L.; Milles, L.; Nguyen, H.; Kang, A.; Bera, A. K.; Peterson, F. C.; Volkman, B. F.; Ovchinnikov, S.; Baker, D. De novo design of small beta barrel proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120* (11), e2207974120.

(62) Tischer, D.; Lisanza, S.; Wang, J.; Dong, R.; Anishchenko, I.; Milles, L. F.; Ovchinnikov, S.; Baker, D. Design of proteins presenting discontinuous functional sites using deep learning. *bioRxiv* **2020**, DOI: [10.1101/2020.11.29.402743](https://doi.org/10.1101/2020.11.29.402743).

(63) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Courbet, A.; Haas, R. J. d.; Bethel, N.; Leung, P. J. Y.; Huddy, T. F.; Pellock, S.; Tischer, D.; Chan, F.; Koepnick, B.; Nguyen, H.; Kang, A.; Sankaran, B.; Bera, A. K.; King, N. P.; Baker, D. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **2022**, *378* (6615), 49–56.

(64) Yeh, A. H.-W.; Norn, C.; Kipnis, Y.; Tischer, D.; Pellock, S. J.; Evans, D.; Ma, P.; Lee, G. R.; Zhang, J. Z.; Anishchenko, I.; Coventry, B.; Cao, L.; Dauparas, J.; Halabiya, S.; DeWitt, M.; Carter, L.; Houk, K. N.; Baker, D. De novo design of luciferases using deep learning. *Nature* **2023**, *614* (7949), 774–780.

(65) Lu, H.; Diaz, D. J.; Czarniecki, N. J.; Zhu, C.; Kim, W.; Shroff, R.; Acosta, D. J.; Alexander, B. R.; Cole, H. O.; Zhang, Y.; Lynd, N. A.; Ellington, A. D.; Alper, H. S. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **2022**, *604* (7907), 662–667.

(66) Anand, N.; Eguchi, R.; Mathews, I. I.; Perez, C. P.; Derry, A.; Altman, R. B.; Huang, P.-S. Protein sequence design with a learned potential. *Nat. Commun.* **2022**, *13* (1), 746.

(67) Pun, M. N.; Ivanov, A.; Bellamy, Q.; Montague, Z.; LaMont, C.; Bradley, P.; Otwinowski, J.; Nourmohammad, A. Learning the shape of protein micro-environments with a holographic convolutional neural network. *arXiv* **2022**, DOI: [10.1101/2022.10.31.514614](https://doi.org/10.1101/2022.10.31.514614).

(68) Jendrusch, M.; Korbel, J. O.; Sadiq, S. K. AlphaDesign: A de novo protein design framework based on AlphaFold. *bioRxiv* **2021**, DOI: [10.1101/2021.10.11.463937](https://doi.org/10.1101/2021.10.11.463937).

(69) Moffat, L.; Greener, J. G.; Jones, D. T. Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design. *bioRxiv* **2021**, DOI: [10.1101/2021.08.24.457549](https://doi.org/10.1101/2021.08.24.457549).

(70) Torres, S. V.; Leung, P. J. Y.; Lutz, I. D.; Venkatesh, P.; Watson, J. L.; Hink, F.; Huynh, H.-H.; Yeh, A. H.-W.; Juergens, D.; Bennett, N. R.; Hoofnagle, A. N.; Huang, E.; MacCoss, M. J.; Expòsit, M.; Lee, G. R.; Korkmaz, E. N.; Nivala, J.; Stewart, L.; Rogers, J. M.; Baker, D. De novo design of high-affinity protein binders to bioactive helical peptides. *bioRxiv* **2022**, DOI: [10.1101/2022.12.05.19862](https://doi.org/10.1101/2022.12.05.19862).

(71) Wang, J.; Lisanza, S.; Juergens, D.; Tischer, D.; Watson, J. L.; Castro, K. M.; Ragotte, R.; Saragovi, A.; Milles, L. F.; Baek, M.; Anishchenko, I.; Yang, W.; Hicks, D. R.; Expòsit, M.; Schlichthaerle, T.; Chun, J.-H.; Dauparas, J.; Bennett, N.; Wicky, B. I. M.; Muenks, A.; DiMaio, F.; Correia, B.; Ovchinnikov, S.; Baker, D. Scaffolding protein functional sites using deep learning. *Science* **2022**, *377* (6604), 387–394.

(72) Mansoor, S.; Baek, M.; Juergens, D.; Watson, J. L.; Baker, D. Accurate Mutation Effect Prediction using RoseTTAFold. *bioRxiv* **2022**, DOI: [10.1101/2022.11.04.515218](https://doi.org/10.1101/2022.11.04.515218).

(73) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffer, V.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek,

M.; Baker, D. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv* **2022**, DOI: [10.1101/2022.12.09.519842](https://doi.org/10.1101/2022.12.09.519842).

(74) Hekkelman, M. L.; de Vries, I.; Joosten, R. P.; Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **2023**, *20* (2), 205–213.

(75) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *arXiv* **2023**, DOI: [10.48550/arXiv.2210.01776](https://doi.org/10.48550/arXiv.2210.01776).

(76) Meller, A.; Bhakat, S.; Solieva, S.; Bowman, G. R. Accelerating Cryptic Pocket Discovery Using AlphaFold. *J. Chem. Theory Comput.* **2023**, DOI: [10.1021/acs.jctc.2c01189](https://doi.org/10.1021/acs.jctc.2c01189).

(77) Dürr, S. L.; Levy, A.; Rothlisberger, U. Accurate prediction of transition metal ion location via deep learning. *bioRxiv* **2022**, DOI: [10.1101/2022.08.22.504853](https://doi.org/10.1101/2022.08.22.504853).

(78) Cheng, Y.; Wang, H.; Xu, H.; Liu, Y.; Ma, B.; Chen, X.; Zeng, X.; Wang, X.; Wang, B.; Shiau, C.; Ovchinnikov, S.; Su, X.-D.; Wang, C. Co-evolution-based prediction of metal-binding sites in proteomes by machine learning. *Nat. Chem. Biol.* **2023**, *19*, 548–555.

Estimating conformational heterogeneity of tryptophan synthase with a template-based Alphafold2 approach

Guillem Casadevall¹  | Cristina Duran¹  | Miquel Estévez-Gay¹  |
Sílvia Osuna^{1,2} 

¹CompBioLab Group, Institut de Química Computacional i Catàlisi and Departament de Química, Universitat de Girona, Girona, Spain

²ICREA, Barcelona, Spain

Correspondence

Sílvia Osuna, CompBioLab Group, Institut de Química Computacional i Catàlisi and Departament de Química, Universitat de Girona, Carrer Maria Aurèlia Capmany 69, 17003 Girona, Spain.
Email: silvia.osuna@udg.edu

Present address

Sílvia Osuna, ICREA, Pg. Lluís Companys 23, 08010, Barcelona, Spain.

Funding information

H2020 European Research Council, Grant/Award Number: ERC-2015-StG-679001; Human Frontier Science Program, Grant/Award Number: RGP0054/2020; Ministerio de Ciencia e Innovación, Grant/Award Number: PGC2018-102192-B-I00

Review Editor: John Kuriyan

Abstract

The three-dimensional structure of the enzymes provides very relevant information on the arrangement of the catalytic machinery and structural elements gating the active site pocket. The recent success of the neural network Alphafold2 in predicting the folded structure of proteins from the primary sequence with high levels of accuracy has revolutionized the protein design field. However, the application of Alphafold2 for understanding and engineering function directly from the obtained single *static* picture is not straightforward. Indeed, understanding enzymatic function requires the exploration of the ensemble of thermally accessible conformations that enzymes adopt in solution. In the present study, we evaluate the potential of Alphafold2 in assessing the effect of the mutations on the conformational landscape of the beta subunit of tryptophan synthase (TrpB). Specifically, we develop a template-based Alphafold2 approach for estimating the conformational heterogeneity of several TrpB enzymes, which is needed for enhanced stand-alone activity. Our results show the potential of Alphafold2, especially if combined with molecular dynamics simulations, for elucidating the changes induced by mutation in the conformational landscapes at a rather reduced computational cost, thus revealing its plausible application in computational enzyme design.

KEYWORDS

Alphafold2, computational enzyme design, conformational heterogeneity, tryptophan synthase

1 | INTRODUCTION

“What are the features that make proteins evolvable?” questioned Tokuriki and Tawfik in their seminal review paper.¹ As opposite to the traditional view of one well-defined structure of proteins, a new “avant-garde view”

in which proteins display conformational variability key for their evolvability was proposed. They described that evolution operates by enriching pre-existing diversities, which provide the protein the ability to acquire new functions. These ensembles of pre-existing conformations in thermal equilibrium with the so-called *native state* are

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

the basis for proteins' evolutionary adaptability^{1–5} and provide an explanation for the observed divergency originated from a few common ancestors,⁶ as well as their versatility as shown by the enzymatic promiscuous side activities.⁶ However, the ability of proteins and enzymes to adopt multiple conformations might a priori seem to counter with their characteristics of being proficient, accurate, and specific. Indeed, the high catalytic activity of enzymes is mostly attributed to their highly pre-organized active site pockets presenting the catalytic machinery well-positioned for efficiently stabilizing the transition state(s) of the reactions.^{7,8} However, the importance of conformational flexibility was demonstrated with the design of catalytic antibodies.⁹ Their modest efficiencies as compared to enzymes were attributed to the imperfect steric and electrostatic environment, but also to their restricted conformational heterogeneity.¹⁰ This shows that efficient catalysis requires a delicate balance between active site pre-organization for transition state stabilization, and also the optimization of the conformational ensemble along the catalytic itinerary. Following an enzymatic cycle in detail, the major steps that take place along a general catalytic itinerary are the following: (1) binding of the substrate(s) in the catalytic pocket, which often involves the exploration of additional conformational states presenting properly positioned loops and flexible domains gating the active site access,^{11,12} (2) activation of the substrate(s) for productive enzyme-substrate (ES) formation; (3) stabilization of the transition state(s) leading to the formation of the multiple reaction intermediates and product(s); (4) release of the product(s), which again is often accompanied by conformational changes to restart the catalytic cycle. All these steps are key players for enhanced catalytic activity.

Since the proposal of Tokuriki and Tawfik of the ensemble-based conformational diversity key for evolvability, many experimental and computational studies have been reported in the literature supporting this idea.^{1,3–5,13} The study of the conformational landscape of natural and laboratory-evolved enzymes showed that by introducing mutations at the active site and also at remote positions changes in the stabilities of the pre-existing conformations can be induced. This was experimentally demonstrated in the laboratory evolution of a phosphotriesterase into an arylesterase (AE) enzyme.^{2,13} The AE activity was gradually increased by changing the fluctuation of some key active site gating loops, as shown by the B-factors of the multiple x-ray structures obtained along the laboratory-evolution path. NMR and room-temperature x-ray crystallography of several HG3 Kemp eliminases also showed a change in the conformational ensembles along laboratory evolution.^{14,15} From a computational perspective, this ensemble view of enzymes

can be represented in the so-called free energy landscape (FEL, Figure 1 for FELs at different reaction stages),⁴ which can be reconstructed by means of molecular dynamics (MD) simulations and enhanced sampling techniques.^{16–18} In the reconstructed FEL, the relative stabilities of the thermally accessible conformations, as well as the kinetic barriers separating them are represented. Depending on the barrier height that separates a given pair of conformational states, the timescale associated to the transition is faster or slower. Conformational changes that can directly impact catalytic function include side-chain conformational changes in the fast timescale, loop motions often playing a key role in substrate binding/product release in slower timescales, and in some cases allosteric transitions that usually correspond to the slowest processes. The reconstruction of the FEL and how this is shifted after mutation provides crucial information for understanding and designing enzyme function.⁴ The introduced mutations located at the active site and many times at remote sites induce a long-range effect affecting enzymatic catalysis. Induced by the mutations introduced, catalytically productive conformational states are stabilized, whereas the non-productive ones for the novel functionality are disfavored, thus converting computational enzyme design into a population shift problem.¹⁹ These observations promoted the exploration of enzyme conformational dynamics for enzyme design.^{3,4,13} The reconstruction of ancestral enzymes displaying a higher degree of flexibility with respect to the modern counterparts and their use as initial scaffolds for enzyme design yielded interesting new insights.²⁰ The higher flexibility of many ancestral variants was found to be key for achieving higher levels of catalytic activity with only a few mutations located at the active site. In this direction, several ancestrally reconstructed enzymes have been used as starting points for enzyme design, for instance for enhancing some residual catalytic promiscuity contained in an enzyme family, for altering the allosteric regulation of some heterodimeric enzymes, among others.^{20–22}

The recent success of the neural network Alphafold2 (AF2) in predicting the folded structure from the primary sequence with high levels of accuracy has revolutionized the field.^{23–26} The novel AF2 neuronal network incorporates information on the evolutionary, physical and geometric constraints of existing protein structures. AF2 is recognized as one of the milestones in protein structure prediction, and has boosted the application of deep-learning methods for many other applications.²⁶ Despite the impressive performance of AF2 algorithms in predicting the native lowest in energy structure of proteins, application of AF2 for understanding and engineering function directly from the obtained single *static* picture is

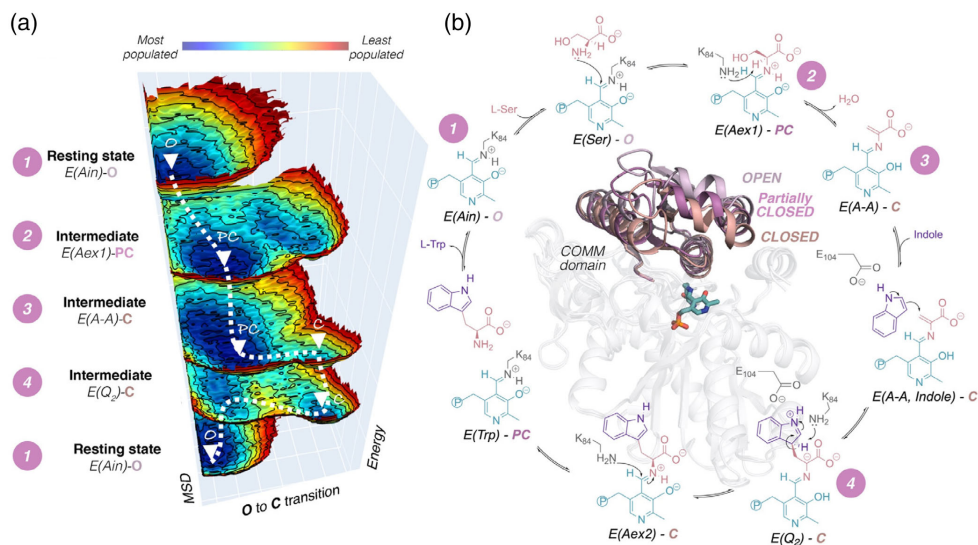


FIGURE 1 (a) Representation of the reconstructed conformational landscape (Source: Data from Ref. 17) of tryptophan synthase B (TrpB) at several reaction intermediates along the catalytic itinerary. The enzyme displays a different conformation of the catalytically relevant COMM domain that covers the active site (shown in pink in the structure displayed at the center of panel b): open (O) states are adopted in the resting state E(Ain), partially closed (PC) at the reaction intermediates E(Aex1) and E(A-A), and closed (C) at E(Q₂) states. Most stable conformations are represented in blue, whereas least stable ones in red. (b) Reaction mechanism of TrpB subunit.²⁹ The conformational states of the COMM domain according to available x-ray data at each reaction intermediate along the catalytic cycle are displayed. Overlay of the different COMM domain conformational states: O highlighted in lilac, PC in pink, and C in brown. Pyridoxal phosphate cofactor is shown in teal, L-Ser in pink and L-Trp in lilac. TrpB, tryptophan synthase B

not straightforward. However, some recent studies have suggested that AF2 can additionally predict multiple conformations of the same protein, and thus it can be potentially used to elucidate the conformational plasticity of biological systems.^{27,28} This is exciting as it suggests that AF2 could be applied for assessing the effect of the introduced mutations on the conformational landscape at a rather reduced computational cost, which would boost the development of conformationally driven enzyme designs protocols.^{4,19}

In this study, we evaluate the potential of AF2 in assessing the effect of the mutations on the conformational landscape of tryptophan synthase (TrpS). TrpS is a heterodimeric enzyme complex (based on TrpA and TrpB subunits) that performs a multistep reaction mechanism together with a sophisticated allosteric signal communication. TrpA catalyzes the retro-aldol cleavage of indole glycerol phosphate producing glyceraldehyde 3-phosphate and indole, the latter being able to diffuse through an internal TrpA-TrpB tunnel to reach the TrpB subunit (Figure 1). For this enzyme, the allosteric

communication between TrpA and TrpB keeps the proper conformations along the cycle optimizing the catalytic steps, thus the absence of the protein partner leads to a deficient conformational ensemble.¹⁷ In the case of TrpB, this involves the change of conformation of a COMM domain that covers the active site, which is known to adapt closed, partially closed, and open conformations (Figure 1).^{17,30,31} This fine-tuning of the conformational ensemble induced by the binding partner makes both TrpB and TrpA substantially less efficient when isolated.^{32–37} By applying laboratory-evolution, the Arnold lab enhanced the stand-alone activity of *Pyrococcus furiosus* TrpB and generated a new variant named 0B2-*p*TrpB exhibiting a 2.9-fold increase in k_{cat} with respect to the original complex.^{32,33} The reconstruction of the last bacterial common ancestor (LBCA) TrpB by means of ancestral sequence reconstruction showed a high level of stand-alone activity, which was found to be lost along evolution.^{22,38} We previously explored the FEL of the ancestrally reconstructed LBCA TrpS in complex and as stand-alone catalyst (LBCA-TrpB), as well as the

wild-type *pfTrpS* complex, isolated *pfTrpB*, and laboratory-evolved stand-alone *OB2-pfTrpB* enzyme.¹⁷ Our results showed that the low stand-alone activity of isolated *pfTrpB* is due to the restricted conformational heterogeneity of the COMM domain, and the inability to adopt catalytically productive closed conformations. The distal mutations introduced in *OB2-pfTrpB* recovered the conformational flexibility of the COMM domain to similar levels to those observed for the allosterically regulated *pfTrpS* complex. However, the closed conformation of the COMM domain was substantially more stable in the case of *OB2-pfTrpB*, which explains its superior catalytic activity with respect to *pfTrpS*. Similar observations were found for *LBCA-TrpB* whose stand-alone activity was mainly attributed to its conformational heterogeneity and ability to adopt catalytically productive closed conformations of the COMM domain.³⁹ These two works elucidated the conformational ensemble that a stand-alone catalyst has to display for being efficient, and revealed dramatic changes in the COMM domain conformation, which are important for the multi-step catalytic pathway of *TrpB* (as shown in Figure 1).^{17,39} This information is pivotal for designing new stand-alone *TrpB* variants, which requires the fine-tuning of the conformational ensemble. In a recent paper,³⁹ we applied the Shortest Path Map (SPM) methodology^{18,19} together with ancestral sequence reconstruction to predict distal activity-enhancing mutations and design a new *TrpB* variant, that we named *SPM6-TrpB*. The experimental validation of *SPM6-TrpB* design demonstrated its superior stand-alone activity in the absence of *TrpA* to similar levels to those achieved with laboratory evolution (seven-fold increase in k_{cat} with respect to the starting ancestral *ANC3-TrpB* enzyme).³⁹ Still the stand-alone activity of the designed *SPM6-TrpB* was far from that of the reference *LBCA-TrpB* (k_{cat} of 0.5 and 0.2 s^{-1} for *LBCA* and *SPM6*, respectively). This was mostly due again to a restricted conformational heterogeneity and the lack of catalytically productive closed conformations of the COMM domain, as revealed by the FEL reconstruction.

In this work, we evaluate the potential of AF2 for quickly estimating the conformational heterogeneity of different *TrpB* displaying different levels of stand-alone activity. We first evaluate the effect of using different multiple sequence alignment (MSA) depths in the AF2 predictions for all *TrpB* systems. We then develop a template-based AF2 approach consisting on providing a set of either x-ray based templates or conformations extracted from MD simulations to estimate the conformational heterogeneity of the different systems. Finally, we run short nanosecond timescale MD simulations from each AF2 prediction to quickly estimate the FEL. Our results show the potential of AF2, especially if combined

with MD simulations, for elucidating the changes induced by mutation in the conformational landscapes.

2 | RESULTS

2.1 | Exploring the conformational heterogeneity by altering the multiple sequence alignment depths in AF2

Inspired by recent pre-print papers in which the conformational heterogeneity of some proteins was estimated by reducing the depth of the input MSAs used in AF2 algorithm (as well as the number of recycles),^{27,28} we decided to test this methodology in *pfTrpB*, *OB2-pfTrpB*, *LBCA-TrpB*, and *SPM6-TrpB* (see more details in Section 4). For these systems, we have previously reconstructed the FEL and have the experimental characterization of the stand-alone activity.^{17,39} In Figure 2, the previously reconstructed FEL of the *OB2-pfTrpB* variant¹⁷ is shown together with the predictions of AF2 for the different analyzed systems considering different MSA depths (represented with vertical lines colored from orange to dark blue depending on the MSA depth). The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1–5 (open, **O**), 6–10 (partially-closed, **PC**), and 11–15 (closed, **C**). The predictions obtained by AF2 for *pfTrpB* and *OB2-pfTrpB* are very similar: in both cases **PC** conformations of the COMM domain are predicted when the MSA depth is higher than 512 (teal lines). Indeed, by increasing the MSA depth more structures closer to the native **PC** state as predicted by the original AF2 are obtained (Table S1). **O** structures are also predicted when a reduced number of MSA (32–64) is used instead (the standard deviation of the **O**-to-**C** path of the predicted structures is ca. 2 for both systems at low MSA depths of 32–64 indicating that several levels of closure of the COMM domain are predicted, Table S1). For these variants no **C** conformations are obtained. This is completely changed in the case of *LBCA-TrpB* and *SPM6-TrpB*. In both cases, **C** conformations of the COMM domain are predicted when high MSA depths are used (256–5120), and by reducing the MSA depth to 32–64 only **PC** structures (no **O** structures) are instead obtained (Table S1). Although most of the predicted structures for *LBCA* and *SPM6-TrpB* (with high MSAs) fall in the range of **C** conformations (mean value of 11 in the **O**-to-**C** pathway in Figure 2 and Table S1), a higher flexibility of the COMM domain is predicted for the ancestral enzyme: *LBCA-TrpB* predictions have **O**-to-**C** values in the 11–15 range, whereas *SPM6-TrpB* in the 13–15 (see x axis in Figure 2, and larger deviation in Table S1). It should be emphasized that the native state

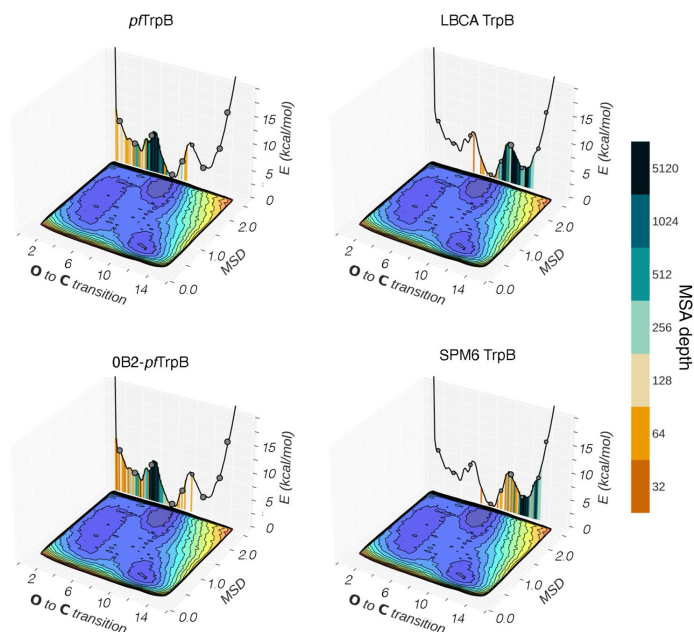


FIGURE 2 Representation of the previously reconstructed FEL of the OB2-*pf*TrpB variant.¹⁷ The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1–5 (open, **O**), 6–10 (partially closed, **PC**), to 11–15 (closed, **C**), the y axis is the MSD deviation from the path of **O**-to-**C** structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.¹⁷ The predictions of AF2 for the different analyzed systems are represented on the 2D-FEL representation using vertical lines colored from orange to dark blue depending on the MSA depth: AF2 predictions obtained with a 32 MSA depth are shown with a vertical orange line, 64 in light orange, 128 in light brown, 256 in light cyan, 512 in cyan, 1024 in teal, and 5120 in dark blue. Black dots indicate some representative available x-ray structures, the size of the spheres is proportional to the sequence identity of the x-ray with respect to the studied TrpB system. FEL, free energy landscape; MSD, mean square deviation

predicted by the default AF2 for both LBCA and SPM6 TrpB has a **C** conformation of the COMM domain. Altogether these results obtained by reducing the MSA depths suggest that OB2-*pf*TrpB and *pf*TrpB have a higher ability to visit **O** and **PC** conformations, whereas LBCA and SPM6 TrpB **PC** and **C** structures. Interestingly, a higher conformational flexibility is predicted for LBCA-TrpB, especially if compared with SPM6-TrpB, which is in line with our previous reconstructed FELs that show a much-limited conformational heterogeneity of the designed SPM6 variant with respect to LBCA TrpB.³⁹ Finally, to quantitatively assess the AF2 predicted conformational heterogeneity in the context of structural variance as observed in x-ray data, we applied Principal Component Analysis (PCA, Figure S3). We focused on the carbon alpha distances of the conserved amino acids of the set of x-ray structures used. The first two components describe 74% and 12.5% of the total variance. The projection of the

predicted AF2 structures with different depths of MSA shows no major deviations from the space generated with experimentally determined structures, thus providing evidence for the validity of the predictions even with a low MSA depth.

2.2 | Exploring the conformational heterogeneity by altering the multiple sequence alignment depths and using x-rays as templates

The inputs for AF2 calculation are the primary sequence of the enzyme, a MSA generated with information of evolutionary related proteins, and the three-dimensional (3D) coordinates of a small number of homologous structures named templates. In the previous section, we reduced the depth of the input MSAs used in AF2

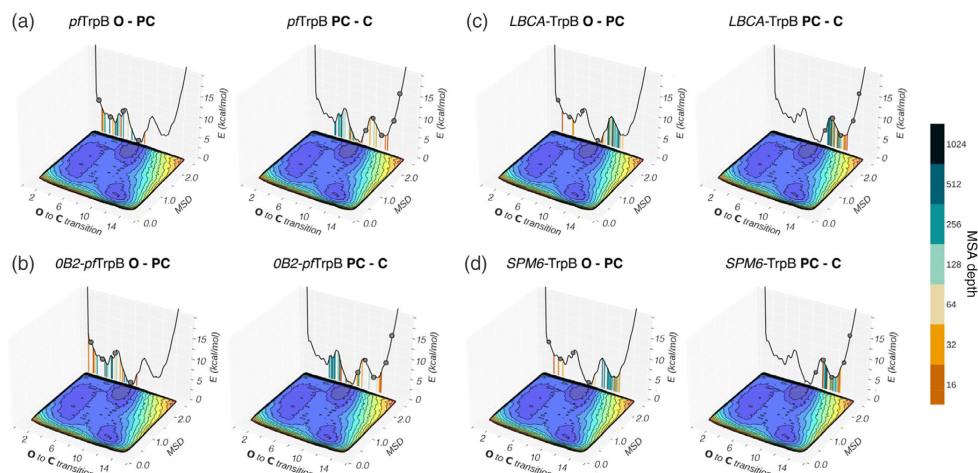


FIGURE 3 Representation of the previously reconstructed FEL of the 0B2-*pfTrpB* variant,¹⁷ and the predictions of the x-ray template-based AF2 approach for the different analyzed systems: *pfTrpB* (a), (b) 0B2-*pfTrpB*, (c) LBCA-TrpB, (d) SPM6-TrpB. The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1–5 (open, O), 6–10 (partially closed, PC), to 11–15 (closed, C), the y axis is the MSD deviation from the path of O-to-C structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.¹⁷ The predictions of the x-ray template-based AF2 for the different analyzed systems are represented on the 2D-FEL representation using vertical lines colored from orange to dark blue depending on the MSA depth: AF2 predictions obtained with a 16 MSA depth are shown with a vertical orange line, 32 in light orange, 64 in light brown, 128 in light cyan, 256 in cyan, 512 in teal, and 1024 in dark blue. For each studied case, the predictions obtained by AF2 using x-ray templates with O-PC conformations of the COMM domain are shown in the left, whereas the results with x-ray templates presenting PC-C conformations in the right. Black dots indicate the used x-ray structures as input templates, and the size of the spheres is proportional to the sequence identity of the x-ray with respect to the studied TrpB system. FEL, free energy landscape; MSD, mean square deviation

algorithm to evaluate the conformational heterogeneity of four TrpB enzymes displaying different levels of stand-alone activity, and found mostly C conformations for the ancestral and designed TrpB, whereas PC for *pfTrpB* based variants. Our hypothesis is that by additionally fine-tuning the other input parameter of AF2, that is, the set of templates used to extract the 3D information, more information regarding the enzyme ability to adopt O, PC, and C conformations can be derived. In fact, in the original formulation of AF2 five models are provided that use different number of MSA depths and template structures to encourage diversity in the predictions.²⁵ It should be also noted that in a recent paper the effect of including different x-ray templates on AF2 predictions was tested only for a specific protein target that was exclusively modeled in only one of the conformations (even with low MSA depths).²⁷ Different decoy structures were used as templates in a recent pre-print paper based on assessing the coevolution dependency of the AF2 learned potential function for scoring protein structures.⁴⁰ In this study, we

have tested the hypothesis that by altering the set of AF2 templates the conformational heterogeneity of the target enzymes can be estimated. We have used a reduced number of template structures based on the available x-ray structures presenting a sequence identity larger than 70% with respect to all systems (Table S2 and Figure S1). The side-chain conformation was kept in the template, however, as done in a previous study⁴⁰ we also run the simulations by hiding side-chain information from the template and providing only the coordinates of the carbon beta (or carbon alpha in the case of glycine, Figures S4 and S5 for the results without side-chain). In Figures 3 and S6, the results from this template-based AF2 strategy are displayed. For the same primary sequence, the prediction from this template-based AF2 approach suggests a different level of closure of the COMM domain depending on the x-ray template structures used (either in a C, PC, or O conformation, Figures 3 and S6). When C and PC x-ray templates are used, AF2 prediction for *pfTrpB* and 0B2-*pfTrpB* mostly suggests PC conformations of the

COMM domain, especially at high MSA depths (ca. 72% of **PC** and **C** structures are predicted, Table S3). The differences between the predicted structures for both systems are small (they only differ in six mutations), although a slightly higher number of **C** conformations are suggested for OB2-*pf*TrpB (74 vs. 71% for OB2-*pf*TrpB and *pf*TrpB, respectively, Table S3). When **O** and **PC** templates are used instead, **PC** conformations are predicted similarly for both cases (79 and 21% of **O-PC** and **PC-C**, respectively, for both cases). The same strategy when applied in the case of LBCA and SPM6 TrpB shows that **C** conformations of the COMM domain are most frequently predicted irrespectively of the x-ray template structure used and the MSA depth applied, as shown in the previous section (Table S3). As discussed in the previous section, the conformational variance of the obtained predictions with this x-ray template-based approach is in line with the structural variance observed in x-ray data (Figure S3).

2.3 | Exploring the conformational heterogeneity by altering the multiple sequence alignment depths and using molecular dynamics conformations as templates

In the specific case of TrpB multiple x-ray structures displaying different conformations of the COMM domain are actually available (Figure S1). However, this is not the case for most of the systems. In this section, we assessed and compared the outcome of the template-based AF2 using conformations extracted from MD simulations instead of x-ray structures. As far as we know, this was not tested in any of the previously mentioned studies based on using AF2 to extract information of the conformational landscape.^{27,28} In particular, we used as input a reduced number of conformations displaying either **C**, **PC**, or **O** conformations of the COMM domain extracted from our recently published FELs at the Q₂-bound state

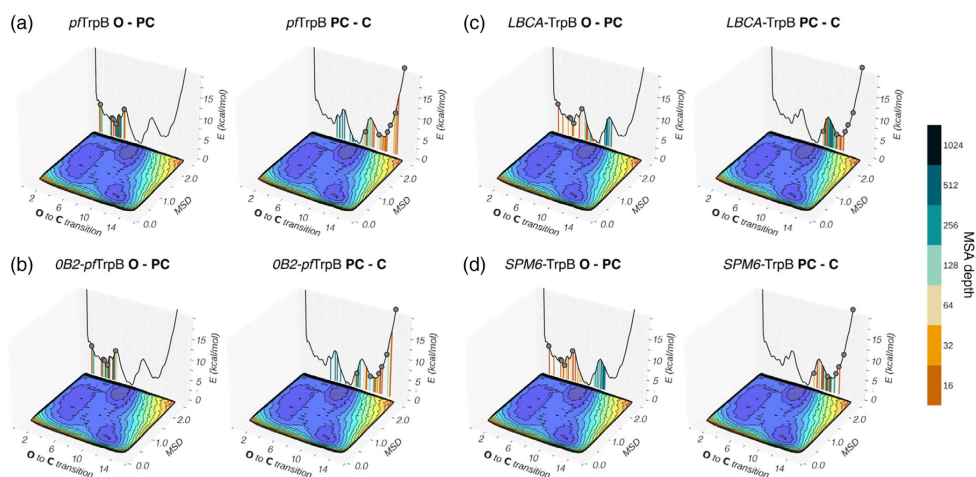


FIGURE 4 Representation of the previously reconstructed FEL of the OB2-*pf*TrpB variant,¹⁷ and the predictions of the MD extracted template-based AF2 approach for the different analyzed systems: *pf*TrpB (a), (b) OB2-*pf*TrpB, (c) LBCA-TrpB, and (d) SPM6-TrpB. The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1–5 (open, **O**), 6–10 (partially closed, **PC**), to 11–15 (closed, **C**), the y axis is the MSD deviation from the path of **O**-to-**C** structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.¹⁷ The predictions of the MD template-based AF2 for the different analyzed systems are represented on the 2D-FEL representation using vertical lines colored from orange to dark blue depending on the MSA depth: AF2 predictions obtained with a 16 MSA depth are shown with a vertical orange line, 32 in light orange, 64 in light brown, 128 in light cyan, 256 in cyan, 512 in teal, and 1024 in dark blue. For each studied case, the predictions obtained by AF2 using as templates conformations extracted from MD simulations with **O-PC** conformations of the COMM domain are shown in the left, whereas the results with MD templates presenting **PC-C** conformations in the right. (a–d) With vertical lines colored from yellow to teal depending on the MSA depth. The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1–5 (open, **O**), 6–10 (partially closed, **PC**), to 11–15 (closed, **C**). Black dots indicate the used representative MD conformations as input templates. FEL, free energy landscape; MD, molecular dynamics; MSD, mean square deviation; MSA, multiple sequence alignment

of the most evolved OB2-*pf*TrpB variant (gray dots in Figure 4).¹⁷ Similarly to what was done with x-ray-based templates, the side-chain conformation was included in the 3D template (Figures S9 and S10 for the results without side-chain information). When MD-extracted C conformations are used as input templates, the predicted structures for *pf*TrpB and OB2-*pf*TrpB present PC conformations of the COMM, especially if high MSA depths are used (green lines in Figures 4 and S11). This is in line with the results obtained in the previous sections where either the MSA depths were only altered or different x-ray templates combined with different levels of MSA were used. Interestingly, by comparing the range of predicted structures at high MSA depths for both *pf*TrpB and OB2-*pf*TrpB systems, which differ only in six distal active site mutations, a slightly higher ability to adapt C conformations of the COMM domain is predicted for the stand-alone OB2-*pf*TrpB variant (74% of the predicted structures adopt PC-C conformations, whereas 67% in the case of *pf*TrpB, Table S3). Indeed, some structures present O-to-C values in the 12–15 range for OB2 instead of the 10–13 for *pf*TrpB with MSA depths higher than 256 (teal vertical lines in Figures 4 and S11). As expected, the use of O and PC conformations of the COMM domain as input templates generate mostly PC conformations for both *pf*TrpB systems irrespective of the MSA depth (ca. 85% of the predicted structures present O-PC conformations for both systems, Table S3). Altogether these results suggest that AF2 predicts as the lowest in energy conformation PC structures for *pf*TrpB and OB2-*pf*TrpB. However, by altering the MSA depths and providing as input templates different conformations of the COMM domain taken from MD simulations some hints about the conformational heterogeneity can be extracted: a higher ability to adopt the catalytically productive C conformation is predicted for the stand-alone OB2-*pf*TrpB, if compared to *pf*TrpB.

The same analysis was performed on the ancestral LBCA-TrpB and SPM6-TrpB design. When O conformations are used as input templates, PC and C structures are predicted for LBCA at high MSA levels, whereas for SPM6 more C conformations are obtained (Table S3). At low MSA (brown colored lines in Figure 4), O, PC, and C structures are similarly predicted for both cases. These results are again suggesting a higher conformational flexibility for the ancestral LBCA-TrpB as compared to the SPM6 design, which is accordance with our previously computed FELs. By using C conformations as templates, the predictions for both systems at either high and low levels of MSA depths yield C structures of the COMM domain (ca. 91% of the predicted structures present PC-C conformations of the COMM domain in both systems,

Table S3). As found for the x-ray template-based AF2 predictions, the conformational variance of the structures generated with this MD template-based approach is in line with the structural variance observed in x-ray data (Figure S3).

2.4 | Exploring the conformational heterogeneity by short nanosecond timescale molecular dynamics simulations from the x-ray template-based AF2 predicted structures

The previous sections have shown that by altering the MSA depth and providing different sets of templates (either based on x-ray structures or conformations extracted from MD simulations) the conformational landscape of different TrpB systems can be estimated. To further validate AF2 predictions and assess its potential application for rapidly estimating the conformational heterogeneity, we decided to run multiple replica short nanosecond timescale MD simulations starting from the set of AF2 structures predicted in Section 2 (2 replicas of 10 ns MD simulations starting from the ca. 60 different AF2 outputs obtained in the x-ray template based AF2 approach, that is, ca. 1200 ns of accumulated MD simulation time for each TrpB variant). In Figure 5, the reconstructed FEL from the set of MD simulations performed starting from AF2 output structures is shown on top of the previously reconstructed FELs of *pf*TrpB, OB2-*pf*TrpB, LBCA-TrpB, and SPM6-TrpB.^{17,39} As discussed in the previous sections, a higher conformational heterogeneity is predicted for OB2-*pf*TrpB and *pf*TrpB with respect to LBCA and SPM6-TrpB that mostly adopt C conformations of the COMM domain. Interestingly, larger differences are observed when comparing OB2-*pf*TrpB and *pf*TrpB that only differ in six mutations (98.4% of sequence identity): although there is only one minimum at C conformations of the COMM domain in both cases, the estimation of the FEL for OB2-*pf*TrpB suggest the existence of an additional minima at O conformations. The O-to-C value of the COMM domain at the C minima is ca. 9 in *pf*TrpB, whereas ca. 10.5 in OB2-*pf*TrpB, thus suggesting a higher ability for adopting the catalytically productive C conformation in OB2-*pf*TrpB, as found in our previous study.¹⁷ In the case of LBCA and SPM6-TrpB a much more restricted conformational heterogeneity is found, in line with the previously reconstructed FELs.³⁹ In fact, in our previous study we found that at the Q₂ intermediate LBCA-TrpB has a wide energy minima at C conformations (mostly presenting a larger deviation along the y axis),

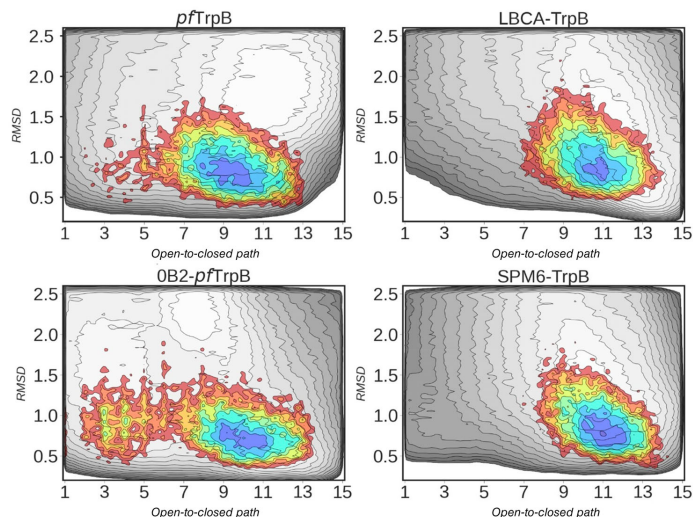


FIGURE 5 Representation of the previously reconstructed FELs of the *pf*TrpB, 0B2-*pf*TrpB, LBCA-TrpB, and SPM6-TrpB (shown in gray scale).^{17,39} The estimated FEL from multiple replica short nanosecond timescale MD simulations performed starting at the x-ray template-based AF2 predictions for the different analyzed systems is shown in color on top of the previously reconstructed FELs. The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1–5 (open, O), 6–10 (partially closed, PC), to 11–15 (closed, C), the y axis is the MSD deviation from the path of O-to-C structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.¹⁷ FEL, free energy landscape; MD, molecular dynamics; MSD, mean square deviation

which confers the enzyme the ability to visit both catalytically productive and unproductive conformations of the COMM domain. The estimated FEL obtained here from the ensemble of short MD simulations starting at the different AF2 structures also suggest a higher deviation along the y axis, and the ability to visit C conformation of the COMM domain (O-to-C values of ca. 11) in line with its high stand-alone activity. The estimated FEL for LBCA and SPM6-TrpB are similar (Figure 5), although the C minima for SPM6-TrpB is wider along the O-to-C axis as it ranges between 9.5 and 12. Although the estimated FELs from this rather short MD simulations present some deviations from the previously reconstructed FELs based on well-tempered multiple-walker metadynamics simulations, the conformational heterogeneity of the different systems can be estimated, which suggests its potential application for rapidly evaluating the effect of mutations into the conformational landscape of enzymes. Finally, the projection of the accumulated MD dataset into the principal component space generated based on the ensemble of available x-ray structures shows no major deviations with experimentally determined structures (Figure S12).

3 | DISCUSSION AND CONCLUSIONS

Tryptophan synthase is a heterodimeric enzyme that features a mechanically complex reaction mechanism (as shown in Figure 1), together with a fine-tuned conformational ensemble that needs to be optimized for enhanced function.^{17,32–37,39} Our previously reconstructed conformational landscapes of the heterodimeric complex as well as several isolated TrpB enzymes showed that by altering the relative stabilities of the open, partially closed, and closed conformations of the catalytically relevant COMM domain, the reaction steps along the catalytic itinerary are optimized.¹⁷ Such conformational changes play an important role in pre-organizing the active site for efficient catalysis, for the binding of the two substrates and for product release. By computationally analyzing multiple TrpBs displaying different stand-alone activities, we found that enhanced stand-alone activity requires the ability to adopt closed conformations of the COMM domain in the absence of the binding partner, as well as a high conformational flexibility to allow substrate binding and product release.³⁹ The high computational cost associated to FEL reconstruction limits the

exploration of such conformational changes to only a few selected enzyme systems, which is a clear limitation for the routine computational design of new stand-alone variants.¹⁹ In this study, we aimed to test the ability of AF2 to quickly estimate the conformational heterogeneity and changes in the conformational landscapes induced by mutations. We focused on the following systems: the allosterically regulated *pf*TrpB enzyme that in the absence of its binding TrpA partner has restricted conformational heterogeneity and thus low catalytic activity; the laboratory-evolved OB2-*pf*TrpB that presents stand-alone activity thanks to the six distal mutations introduced that recover the allosterically regulated conformational ensemble; the ancestrally reconstructed LBCA TrpB that does not require TrpA to operate efficiently as it presents the ability to adopt different levels of closed conformations of the COMM; and our recently designed SPM6 TrpB variant displaying some stand-alone activity.^{17,39} By changing the depth of the MSA used and altering the input template structures (either from x-ray data or conformations taken from MD simulations), the conformational heterogeneity of the systems can be estimated. This is particularly evidenced by running multiple short nanosecond timescale MD simulations from the provided set of structures by this tuned AF2 approach and reconstructing the associated conformational landscapes.

Interestingly, by altering the MSA depth and including either x-ray or MD-based structures as templates, AF2 predicts mostly partially closed (PC) conformations of the COMM domain for OB2 and *pf*TrpB, whereas closed (C) structures for the ancestral LBCA and SPM6 TrpB design. By further analyzing the output structures provided by using either C or O templates in the different systems, one can estimate a higher conformational heterogeneity for *pf*TrpB-based systems, as O, PC, and C conformations can be generated at both higher and lower MSA depths. In contrast, LBCA and SPM6 TrpB predictions are much more restricted to the PC and C ensemble. The lack of O structures for LBCA-TrpB and SPM6-TrpB is also in accordance with our previous calculations at the Q₂ intermediate that suggested an infrequent transition towards O states as the reaction progresses, thus suggesting that product release might be rate-limiting.³⁹ A high similitude in the predictions is observed when comparing OB2 and *pf*TrpB systems, which is attributed to the high sequence identity between both systems (only six mutations are introduced in OB2, that is, 98.4% of sequence identity). However, AF2 predictions suggest a higher number of C conformations of the COMM domain for OB2-*pf*TrpB variant (74% vs. 67% of PC-C structures when using C MD templates). This increased number of C structures for the evolved variant is in accordance with the reconstructed FELs that show

that at the Q₂ intermediate the C conformation of the COMM is much more accessible for OB2 than for *pf*TrpB in the presence of TrpA (i.e., for *pf*TrpS complex).¹⁷ The estimation of the conformational landscapes from multiple replica short nanosecond timescale MD simulations starting at the different x-ray template-based AF2 predictions are in line with the previously reconstructed computationally expensive FELs obtained from well-tempered multiple-walker metadynamics simulations.³⁹ This suggests that the developed tuned AF2 approach combined with short MD simulations could be potentially applied for rapidly estimating changes on the conformational landscape at a rather reduced computational cost.

Altogether, the distribution of the AF2 predictions in the reconstructed FEL highlights how AF2 learned to locate the global minimum for the input sequence. In this regard, the increase of co-evolutionary information from the MSA forces the network to predict structures close to the global minimum, even if a deviated template is provided. As it can be rationalized from Figures S4 and S6, AF2 predicted as the most probable FEL regions those containing the templates with the highest sequence identities (see largest spheres in Figures S4 and S6). Thus, including co-evolutionary information limits the conformational exploration, which highlights the importance of developing tuned template-based AF2 approaches for assessing the conformational heterogeneity of protein structures. The results provided in this study indicate that by altering the MSA depth and using either x-ray structures or conformations taken from MD simulations, the conformational heterogeneity of related TrpB variants can be quickly estimated. This is specially the case if AF2 predictions are then further evaluated by means of multiple short nanosecond timescale MD simulations. Although much drastic differences are observed when comparing systems presenting lower sequence identities, subtle conformational changes induced by a small number of mutations can also be potentially captured. This is exciting as it suggests that AF2 could be applied for assessing the effect of the introduced mutations on the conformational landscape at a rather reduced computational cost, and opens the door to new AF2-based computational enzyme design approaches.

4 | MATERIALS AND METHODS

AF2 structure prediction starts with a FASTA sequence as an input that is used to generate the MSA and find structural templates, with which AF2 was trained. Five models are obtained as a result, which come from different combinations of random seeds, and considering a

different number of structural templates and extra sequences. This blend of strategies leads to a larger diversity in the predictions.⁴¹ Even so, the obtained structures are often very similar as they are exploring the lowest in energy conformations of the protein. In recent papers, different strategies have been proposed to overcome this static picture provided by AF2: (a) del Alamo et al.²⁷ the number of sequences of the MSA provided to AF2 was modified to contain as few as 16 sequences, they also reduced the number of recycles to 1, and avoid the final MD simulation to reduce the computational cost of the pipeline; (b) Stein and coworkers proposed to replace some specific residues within the MSA to alanine or another residue to potentially manipulate the distance matrices leading to alternate conformations.²⁸ To investigate whether AF has learned a coevolution-independent potential function for scoring protein structures Roney and Ovchinnikov evaluated the effect of using decoy structures as templates with missing amino acids.⁴⁰ In this work, we fine-tune several parameters that differ from the default AF2: (1) MSA depths, as described in the previously mentioned papers,^{27,28} giving less coevolution information and thus leading to an increase of the conformation diversity and (2) the set of templates used, that come either from a subset of x-ray structures (as done in del Alamo et al.²⁷), or from conformations taken from previous MD simulations¹⁷ (this was not tested in any of the previously mentioned papers).

In particular, we used the following protocol: Starting from the MSA depth alteration, we reduced the number of recycles to one (“*num_recycle* = 1”) because of performance reasons. Similarly, Amber minimization was also deactivated, so no structure relaxation was requested (“*amber_relaxed*” = none). Each of the five models were run with 10 different MSA depths. This value is controlled by “*max_extra_msa*” and “*max_msa_clusters*” parameters. The first parameter is described as the number of extra sequences used; and the latter determines the number of the sequence clusters used for the AF2 neural network. Here, the first parameter was comprised between 5120 and 32. Note that, as described in a previous paper,²⁷ we set the latter parameter as the half of the former parameter except when 5120 sequences are used. In that particular case, “*max_msa_clusters*” parameter was set to 512.²⁷ In order to include templates, two strategies were performed: (1) considering x-ray structures and (2) conformations extracted from our previously reconstructed free energy landscape of OB2-*pf*TrpB.¹⁷ Nine x-ray structures (Figure S2) and 11 MD structures (Figure S8) were used as templates presenting different levels of closure of the COMM domain. The parameters: “*num_recycle*” and “*amber_relaxed*” were also maintained as described before. We focused on

“*model_ptm_2*” as it presented the most confident structure results in terms of the predicted LDDT-C α score (pLDDT) and TM-score (pTM) values. In the case of template-based calculations, the MSA depth was altered in the 1024–16 range (“*max_extra_msa*” comprised between 1024 and 16, and “*max_msa_clusters*” was set at its half). Finally, the parameter “*reduce_msa_clusters_by_max_templates*” was deactivated. Also, AF2 calculations considering only the targeted sequence in the MSA was done to ignore co-evolution information (Figures S5 and S10).

4.1 | Molecular dynamics simulations

The starting structures for the four enzymes (*pf*TrpB, OB2-*pf*TrpB, LBCA-TrpB, and SPM6-TrpB) were generated with the predictions of the x-ray template-based AF2 approach. We performed two replicas of 10 ns MD simulations at the Q₂ intermediate starting from a total of 60, 59, 62, and 59 AF2 structures for *pf*TrpB, OB2-*pf*TrpB, LBCA-TrpB, and SPM6-TrpB systems, respectively. All calculations were performed using a modification of the amber99 force field (ff14SB) using AMBER 20 (see Appendix S1 for a complete description of the methods).⁴²

AUTHOR CONTRIBUTIONS

Silvia Osuna: Conceptualization (equal); funding acquisition (lead); supervision (lead); writing – original draft (lead); writing – review and editing (lead). **Guillem Casadevall:** Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Cristina Duran:** Data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Miquel Estévez-Gay:** Data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGMENTS


We thank the Generalitat de Catalunya for the emerging group CompBioLab (2017 SGR-1707) and Spanish MINECO for project PGC2018-102192-B-I00. Cristina Duran was supported by the Spanish MINECO for a PhD fellowship (PRE2019-089147), and Guillem Casadevall and Miquel Estévez-Gay by a research grant from ERC-StG (ERC-2015-StG-679001). Silvia Osuna 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 Dr. Miguel Á. Maria-Solano for sharing the jupyter-notebook of the free energy landscape of the different TrpB systems analyzed.

ORCID

Guillem Casadevall  <https://orcid.org/0000-0003-4442-1600>

Cristina Duran  <https://orcid.org/0000-0003-3094-8823>

Miquel Estévez-Gay  <https://orcid.org/0000-0002-8576-8777>

Sílvia Osuna  <https://orcid.org/0000-0003-3657-6469>

REFERENCES

1. Tokuriki N, Tawfik DS. Protein dynamism and evolvability. *Science*. 2009;324:203–207.
2. Campbell E, Kaltenbach M, Correy GJ, et al. The role of protein dynamics in the evolution of new enzyme function. *Nat Chem Biol*. 2016;12:944–950.
3. Crean RM, Gardner JM, Kamerlin SCL. Harnessing conformational plasticity to generate designer enzymes. *J Am Chem Soc*. 2020;142:11324–11342.
4. Maria-Solano MA, 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.
5. Petrović D, Rizzo VA, Kamerlin SCL, Sanchez-Ruiz JM. Conformational dynamics and enzyme evolution. *J R Soc Interface*. 2018;15:20180330.
6. Khersonsky O, Tawfik DS. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu Rev Biochem*. 2010;79:471–505.
7. Warshel A, Sharma PK, Kato M, Xiang Y, Liu H, Olsson MHM. Electrostatic basis for enzyme catalysis. *Chem Rev*. 2006;106:3210–3235.
8. Marti S, Roca M, Andres J, et al. Theoretical insights in enzyme catalysis. *Chem Soc Rev*. 2004;33:98–107.
9. Winkler CK, Schrittwieser JH, Kroutil W. Power of biocatalysis for organic synthesis. *ACS Cent Sci*. 2021;7:55–71.
10. Hilvert D. Critical analysis of antibody catalysis. *Annu Rev Biochem*. 2000;69:751–793.
11. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol*. 2009;5:789–796.
12. Hammes GG, Benkovic SJ, Hammes-Schiffer S. Flexibility, diversity, and cooperativity: Pillars of enzyme catalysis. *Biochemistry*. 2011;50:10422–10430.
13. Campbell EC, Correy GJ, Mabbitt PD, Buckle AM, Tokuriki N, Jackson CJ. Laboratory evolution of protein conformational dynamics. *Curr Opin Struct Biol*. 2018;50:49–57.
14. Broom A, Rakotoharisoa RV, Thompson MC, et al. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat Commun*. 2020;11:4808.
15. Otten R, Pádua RAP, Bunzel HA, et al. How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. *Science*. 2020;370:1442–1446.
16. Curado-Carballada C, Feixas F, Iglesias-Fernández J, Osuna S. Hidden conformations in aspergillus Niger monoamine oxidase are key for catalytic efficiency. *Angew Chem Int Ed*. 2019;58:3097–3101.
17. Maria-Solano MA, Iglesias-Fernández J, Osuna S. Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution. *J Am Chem Soc*. 2019;141:13049–13056.
18. Romero-Rivera A, Garcia-Borrás M, Osuna S. Role of conformational dynamics in the evolution of retro-aldolase activity. *ACS Catal*. 2017;7:8524–8532.
19. Osuna S. The challenge of predicting distal active site mutations in computational enzyme design. *Wiley Interdiscip Rev Comput Mol Sci*. 2021;11:e1502.
20. Gardner JM, Biler M, Rizzo VA, Sanchez-Ruiz JM, Kamerlin SCL. Manipulating conformational dynamics to repurpose ancient proteins for modern catalytic functions. *ACS Catal*. 2020;10:4863–4870.
21. Devamani T, Rauwerdink AM, Lunzer M, et al. Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J Am Chem Soc*. 2016;138:1046–1056.
22. 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*. 2020;117:346–354.
23. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577:706–710.
24. Ourmazd A, Moffat K, Lattman EE. Structural biology is solved—Now what? *Nat Methods*. 2022;19:24–26.
25. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–589.
26. Callaway E. ‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures. *Nature*. 2020;588:203–204.
27. del Alamo D, Sala D, McHaourab HS, Meiler J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife*. 2022;11:e75751.
28. Stein RA, McHaourab HS. Modeling alternate conformations with Alphafold2 via modification of the multiple sequence alignment. *bioRxiv*. 2021;470469.
29. Dunn MF. Allosteric regulation of substrate channeling and catalysis in the tryptophan synthase bienzyme complex. *Arch Biochem Biophys*. 2012;519:154–166.
30. Hioki Y, Ogasahara K, Lee SJ, et al. The crystal structure of the tryptophan synthase beta subunit from the hyperthermophile *Pyrococcus furiosus*. Investigation of stabilization factors. *Eur J Biochem*. 2004;271:2624–2635.
31. Lee SJ, Ogasahara K, Ma JC, et al. Conformational changes in the tryptophan synthase from a hyperthermophile upon alpha (2)beta(2) complex formation: Crystal structure of the complex. *Biochemistry*. 2005;44:11417–11427.
32. Buller AR, Brinkmann-Chen S, Romney DK, Herger M, Murciano-Cales J, Arnold FH. Directed evolution of the tryptophan synthase beta-subunit for stand-alone function recapitulates allosteric activation. *Proc Natl Acad Sci U S A*. 2015;112:14599–14604.

33. Buller AR, van Roye P, Cahn JKB, Scheele RA, Herger M, Arnold FH. Directed evolution mimics allosteric activation by stepwise tuning of the conformational ensemble. *J Am Chem Soc.* 2018;140:7256–7266.
34. Romney DK, Murciano-Calles J, Wehrmuller JE, Arnold FH. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J Am Chem Soc.* 2017;139:10769–10776.
35. Buller AR, van Roye P, Murciano-Calles J, Arnold FH. Tryptophan synthase uses an atypical mechanism to achieve substrate specificity. *Biochemistry.* 2016;55:7043–7046.
36. Herger M, van Roye P, Romney DK, Brinkmann-Chen S, Buller AR, Arnold FH. Synthesis of beta-branched tryptophan analogues using an engineered subunit of tryptophan synthase. *J Am Chem Soc.* 2016;138:8388–8391.
37. Murciano-Calles J, Romney DK, Brinkmann-Chen S, Buller AR, Arnold FH. A panel of TrpB biocatalysts derived from tryptophan synthase through the transfer of mutations that mimic allosteric activation. *Angew Chem Int Ed Engl.* 2016;55:11577–11581.
38. Busch F, Rajendran C, Heyn K, Schlee S, Merkl R, Sterner R. Ancestral tryptophan synthase reveals functional sophistication of primordial enzyme complexes. *Cell Chem Biol.* 2016;23:709–715.
39. Maria-Solano MA, Kinateder T, Iglesias-Fernández J, Sterner R, Osuna S. In silico identification and experimental validation of distal activity-enhancing mutations in tryptophan synthase. *ACS Catal.* 2021;11:13733–13743.
40. Roney JP, Ovchinnikov S. State-of-the-art estimation of protein model accuracy using AlphaFold. *bioRxiv.* 2022;484043. <https://doi.org/10.1101/2022.03.11.484043>.
41. Hegedűs T, Geisler M, Lukács GL, Farkas B. Ins and outs of AlphaFold2 transmembrane protein structure predictions. *Cell Mol Life Sci.* 2022;79:73.
42. Case DA, Belfon K, Ben-Shalom IY, et al. AMBER 2020. San Francisco, CA: University of California, San Francisco, 2020.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Casadevall G, Duran C, Estévez-Gay M, Osuna S. Estimating conformational heterogeneity of tryptophan synthase with a template-based Alphafold2 approach. *Protein Science.* 2022;31(10):e4426. <https://doi.org/10.1002/pro.4426>



Chapter 5:

Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

This chapter corresponds to the following publication:

Casadevall, G.; Pierce, C.; Guan, B.; Iglesias-Fernandez, J.; Lim, H.-Y.; Greenberg, L. R.; Walsh, M. E.; Shi, K.; Gordon, W.; Aihara, H.; Evans, R. L.; Kazlauskas, R.; Osuna, S. Designing efficient enzymes: Eight predicted mutations convert a hydroxynitrile lyase into an efficient esterase. *BioRxiv*, **2023**. DOI:10.1101/2023.08.23.554512.

Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

Guillem Casadevall^{a‡}, Colin Pierce^{b‡}, Bo Guan^b, Javier Iglesias-Fernandez^a, Huey-Yee Lim^b, Lauren R. Greenberg^b, Meghan E. Walsh^b, Ke Shi^b, Wendy Gordon^b, Hideki Aihara^b, Robert L. Evans III^b, Romas Kazlauskas^b, Sílvia Osuna^{a,c}

^aInstitut de Química Computacional i Catàlisi and Departament de Química, Universitat de Girona, Carrer Maria Aurèlia Capmany 69, 17003 Girona, Spain

^bBiotechnology Institute and Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, 1479 Gortner Avenue, Saint Paul, MN 55108 USA

^cICREA, Barcelona, Spain

[‡]co-first authors

evans858@umn.edu

rjk@umn.edu

silvia.osuna@udg.edu

Abstract: Hydroxynitrile lyase from rubber tree (*HbHNL*) shares 45% identical amino acid residues with the homologous esterase from tobacco, SABP2, but the two enzymes catalyze different reactions. The x-ray structures reveal a serine-histidine-aspartate catalytic triad in both enzymes along with several differing amino acid residues within the active site. Previous exchange of three amino acid residues in the active site of *HbHNL* with the corresponding amino acid residue in SABP2 (T11G-E79H-K236M) created variant HNL3, which showed low esterase activity toward p-nitrophenyl acetate. Further structure comparison reveals additional differences surrounding the active site. *HbHNL* contains an improperly positioned oxyanion hole residue and differing solvation of the catalytic aspartate. We hypothesized that correcting these structural differences would impart good esterase activity on the corresponding *HbHNL* variant. To predict the amino acid substitutions needed to correct the structure, we calculated shortest path maps for both *HbHNL* and SABP2, which reveal correlated movements of amino acids in the two enzymes. Replacing four amino acid residues (C81L-N104T-V106F-G176S) whose movements are connected to the movements of the catalytic residues yielded variant HNL7TV (stabilizing substitution H103V was also added), which showed an esterase catalytic efficiency comparable to that of SABP2. The x-ray structure of an intermediate variant, HNL6V, showed an altered solvation of the catalytic aspartate and a partially corrected oxyanion hole. This dramatic increase in catalytic efficiency demonstrates the ability of shortest path maps to predict which residues outside the active site contribute to catalytic activity.

Introduction

One of the most important qualities of an enzyme is its ability to efficiently catalyze reactions. Nature's enzymes are efficient with natural substrates, but many potential applications require enzymes to work with unnatural substrates or even catalyze new chemical steps. Engineering enzymes to Nature-like efficiencies remains an unsolved

problem.^[1] Protein design often yields enzymes millions of times slower than natural enzymes.

Improving enzyme efficiency is challenging for several reasons. First, an efficient enzyme must simultaneously optimize substrate binding, transition state stabilization, and product release. Second, proteins move continuously,^[2-6] but catalysis requires precisely positioning the substrate and catalytic groups for reaction. It is difficult to predict how to shift the conformational landscape to favor the catalytically competent conformations. Third, residues outside the active site, not in direct contact with substrate, contribute to catalysis as shown by many directed evolution experiments, but it is difficult to predict how these distant residues impact catalysis.

The objective of this paper is to test whether correlated protein motions (shortest path maps, SPM^[1, 4]) can predict residues outside the active site that contribute to catalysis. These SPM's identify residues that move together during molecular dynamics simulations. We hypothesize that residues whose motion is correlated with the motions of the catalytic residues are those that contribute most strongly to efficient catalysis. Mutating these correlated positions should shift global minimum energy conformation towards catalytically competent conformations. Previously, SPM's identified locations of beneficial substitutions previously identified by directed evolution,^[1, 4] and substitutions that allosterically activated tryptophan synthase B a modest 7-fold in k_{cat} and 4-fold in k_{cat}/K_M .^[7]

The test case is to increase the efficiency of an inefficient esterase (a modified hydroxynitrile lyase, HNL3V) by computational design. The engineering uses a homologous esterase, SABP2, for comparison and focuses on transferring substitutions from the homologous esterase to the modified hydroxynitrile lyase. This approach limits the scope of the problem, but it remains challenging since the esterase and modified hydroxynitrile lyase differ by 146 residues. There are 2^{146} or approximately 10^{44} possible variants one could create by exchanging residues between the two enzymes.

Previous design of esterase activity achieved only modest catalytic efficiencies;^[8-12] the best k_{cat}/K_M was $6,600 \text{ M}^{-1} \text{ min}^{-1}$.^[11] In contrast, the work below reports an eight-substitution variant with excellent catalytic efficiency (k_{cat}/K_M of $120,000 \text{ M}^{-1} \text{ min}^{-1}$), which is two-fold higher than that for the target esterase. In addition, computational and x-ray structure analysis of the variants reveal the molecular basis of the improved esterase activity: improved positioning of an oxyanion-hole-stabilizing residue and altered solvation of the catalytic aspartate.

Results

Homologs SABP2 and *HbHNL* catalyze different reactions with similar active sites

HbHNL and SABP2 are homologous enzymes with similar catalytic triads but different catalytic activity. *HbHNL* shares 44% sequence identity and 62% similarity over 260 positions with the modern esterase SABP2 (salicylic acid binding protein 2 from tobacco), Supplementary Fig. 1. *HbHNL* and SABP2 contain the same Ser-His-Asp catalytic triad, but *HbHNL* catalyzes cyanohydrin cleavage,^[13] while SABP2 catalyzes ester hydrolysis.^[14] Both

enzymes belong to the α/β -hydrolase fold superfamily.^[15, 16] *HbHNL* and other hydroxynitrile lyases diverged from the esterases approximately 100 million years ago.^[17, 18]

SABP2 and other serine esterases contain a catalytic triad (Ser-His-Asp) and an oxyanion hole, which consists of two main chain N-H's that form hydrogen bonds to the substrate carbonyl oxygen, Fig. 1. In the first step of the reaction, His238 abstracts a proton from the O_γ of Ser81, which initiates a nucleophilic attack on the carbonyl carbon of the pNPAc substrate (Fig. 1a). The resultant oxyanion is stabilized through H-bonds with backbone N-H's at positions 13 and 82.

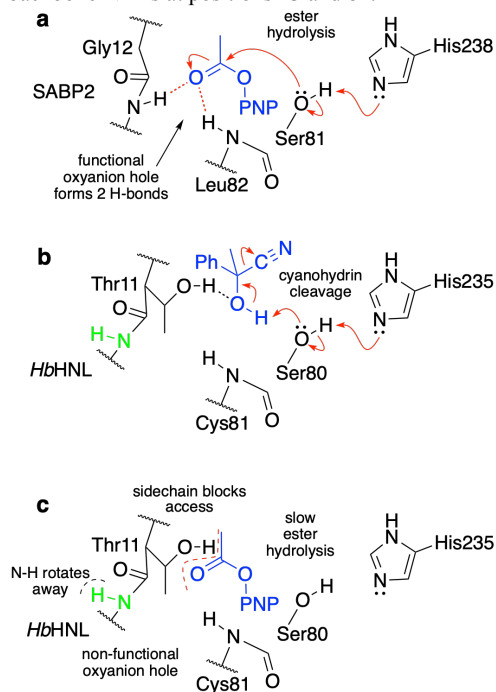


Fig. 1 | Although esterase SABP2 and hydroxynitrile lyase *HbHNL* both contain a Ser-His-Asp catalytic triad, *HbHNL* lacks a functional oxyanion hole preventing it from catalyzing ester hydrolysis. a, The carbonyl oxygen of the *p*-nitrophenyl acetate substrate (blue) accepts hydrogen bonds from two main chain N-H's (Gly12, Leu82). These simultaneous hydrogen bonds are known as the oxyanion hole. Hydrolysis of the ester starts with a nucleophilic attack of Ser81 on the carbonyl carbon to form a tetrahedral intermediate (not shown). The oxyanion hole stabilizes the negative charge that forms on the oxygen in the tetrahedral intermediate. **b**, *HbHNL* catalyzes cleavage of a cyanohydrin (blue) in a single step. The reaction does not involve an oxyanion intermediate, so catalysis does not require an oxyanion hole. **c**, *HbHNL* catalyzes slow *p*-nitrophenyl acetate hydrolysis (blue). The oxyanion hole in *HbHNL* is disrupted by the side chain of Thr11, which hinders access of the

substrate to this region and by a twist in the main chain that points the N-H (green) away from the substrate. For clarity none of the diagrams show the aspartate of the catalytic triad.

HbHNL catalyzes the enantioselective cleavage of mandelonitrile, an aromatic cyanohydrin (Fig. 1b).^[19, 20] In contrast to the multi-step ester hydrolysis catalyzed by SABP2, this lyase reaction occurs in a single step. Catalysis of mandelonitrile cleavage uses the catalytic triad for simple acid-base chemistry. The catalytic Ser deprotonates the substrate hydroxyl; subsequent elimination of cyanide yields benzaldehyde. The lyase reaction does not involve an oxyanion intermediate.

HbHNL also catalyzes promiscuous hydrolysis of pNPAc, but 500-fold slower than SABP2 (k_{cat} of 0.25 min⁻¹ vs. 130 min⁻¹ for SABP2), Fig. 1c. One reason for the low esterase activity of *HbHNL* is that access to the oxyanion hole is blocked by the side chain of Thr11. During lyase catalysis, the side chain hydroxyl of Thr11 can donate a hydrogen bond to the cyanohydrin hydroxyl, Fig. 1b. The corresponding residue in SABP2 is Gly12, which allows full access to the oxyanion hole. Replacement of Gly12 with threonine in SABP2 decreased the esterase activity 2000-fold^[21] confirming that the threonine hinders ester hydrolysis. However, replacement of Thr11 in *HbHNL* with glycine increases the promiscuous hydrolysis of pNPAc only slightly^[19] suggesting that additional structural differences between the oxyanion hole in *HbHNL* and SABP2 contribute to the low esterase activity of *HbHNL*.

The positioning of one oxyanion hole residue differs significantly between SABP2 and *HbHNL*.

The x-ray crystal structures of SABP2 and *HbHNL* show similar placement for the catalytic atoms. The x-ray structure of apo SABP2 and three x-ray structures of apo *HbHNL* were overlaid to best fit the positions of all corresponding C α , Fig. 2. The comparison was restricted to structures without a ligand in the active site to avoid structure changes caused by the ligand. The comparison of multiple structures may also account for some of the motion of the atoms in solution. Four of the catalytic atoms (serine O γ , histidine N ϵ 2, aspartate O δ 2, and oxyanion hole residue OX2 N) align closely between SABP2 and the three *HbHNL* structures with RMSD of 0.4-0.9 Å.

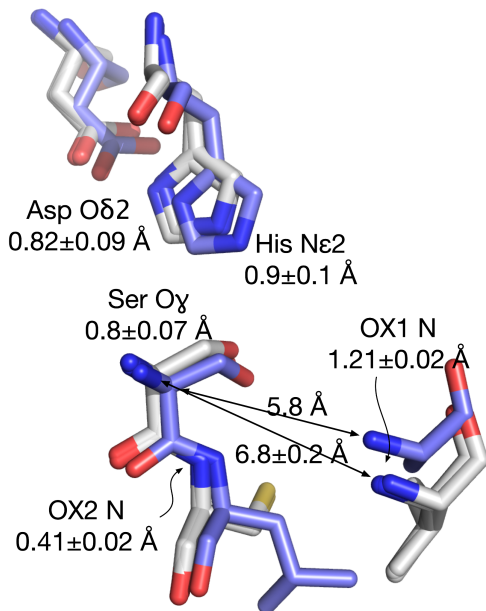


Fig. 2 | Overlay of the catalytic residues of three apo *HbHNL* structures (pdb id = 6yas, 3c6x, 2g4l; white carbons) onto the structure of apo SABP2 (1y7h; blue carbons). The alignment minimized the RMSD between the corresponding Ca atoms in the entire protein. The average deviation was 0.68 Å over 213-218 aligned Ca atoms out of 256. The five catalytic triad and oxyanion hole atoms deviated by slightly more than the average Ca deviation, an average of 0.8 ± 0.3 Å. The largest deviation was the catalytic nitrogen atom of oxyanion residue OX1 (Ala13 in SABP2, Ile12 in *HbHNL*), which deviated by 1.21 ± 0.02 Å. This deviation likely contributes to the poor esterase activity of *HbHNL*. Another way to measure this difference is the serine Ca to OX1 N distance within each structure. For SABP2, this distance is 5.8 Å, while for the three *HbHNL* structures, this distance is longer, 6.8 ± 0.2 Å.

The remaining catalytic atom (oxyanion hole residue OX1 N) differs significantly between SABP2 and *HbHNL*: 1.2 Å. We hypothesized that this difference contributes to the poor promiscuous esterase activity of *HbHNL*. The k_{cat}/K_M value for hydrolysis of *p*-nitrophenyl acetate by *HbHNL* is only 0.1% of the value for SABP2, Table 1. Since OX1 N is a main chain atom, fixing this deviation requires shifting the protein backbone. A disrupted oxyanion hole in *HbHNL* is not expected to hinder lyase catalysis because it does not involve an oxyanion intermediate.

Table 1. Steady-state kinetics parameters for *p*-nitrophenyl acetate hydrolysis catalyzed by SABP2, *HbHNL* and *HbHNL* enzyme variants. See Supplementary Table 1 for a more extensive list of enzyme variants.

Variant	k_{cat} (min ⁻¹)	K_M (mM)	k_{cat}/K_M (M ⁻¹ ·min ⁻¹)	k_{cat} (% relative to SABP2)	k_{cat}/K_M (% relative to SABP2)
SABP2	134 ± 4	2.2 ± 0.2	61,000	100%	100%
<i>HbHNL</i>	0.25 ± 0.02	3.0 ± 0.4	83	0.2%	0.1%
HNL3	0.33 ± 0.02	0.71 ± 0.1	460	0.2%	0.8%
HNL3V	0.32 ± 0.02	0.65 ± 0.1	490	0.2%	0.9%
H N L 3 V N I 0 4 A - G176S	2.1 ± 0.17	0.04 ± 0.01	52,000	1.6%	87%
HNL6V ^a	0.59 ± 0.14	0.03 ± 0.003	20,000	0.4%	33%
HNL7V	3.9 ± 0.5	0.16 ± 0.07	25,000	2.9%	40%
HNL7TV	9.3 ± 0.3	0.08 ± 0.04	120,000	7.0%	190%
HNL8V	8.1 ± 0.4	0.35 ± 0.06	23,000	6.1%	38%

^a Preliminary data for HNL6V measured at the typical temperature of 21°C. k_{cat} at 29°C was 2.3 ± 0.02 min⁻¹, K_M 0.13 ± 0.01 mM

The internal distance between serine Ca and OX1 N is significantly longer in the *HbHNL* structures (6.8 ± 0.2 Å) than in SABP2 (5.8 Å). This difference indicates that the ester carbonyl group, which interacts with the serine O_y and OX1 N in the transition state, cannot make the same interactions in *HbHNL* and SABP2.

Previous engineering within active site yielded only inefficient esterase activity

Three amino acid residues in the active site of *HbHNL* are thought to have a mechanistic role in the catalysis of the lyase reaction, but hinder ester hydrolysis.^[21] The side chain of Thr11 helps orient the hydroxynitrile substrate, but blocks access of the ester substrate to the oxyanion hole (Fig. 1c). Lys236, oriented by Glu79, stabilizes the leaving cyanide from hydroxynitriles but hinders the loss of a hydrophobic group from an ester. Replacement of these three residues in *HbHNL* with the corresponding residues in SABP2 (*HbHNL*-T11G-E79H-K236M) to create HNL3 increased the esterase catalytic efficiency (k_{cat}/K_M) of

HbHNL 5.6-fold from 84 to 470 $M^{-1}\cdot\text{min}^{-1}$, Table 1. For experimental convenience, we created HNL3V, which contains an additional H103V substitution that stabilizes the protein. This H103V substitution did not affect the esterase activity of HNL3 and is analogous to the stabilizing H103L substitution in the homologous HNL from *Manihot esculenta*.^[22] We hypothesized that additional substitutions outside the active site are required to reposition the catalytic machinery of *HbHNL*, including the oxyanion hole, to enable efficient esterase activity.

Shortest path map identifies residues outside the active site that alter the conformational landscape

Since proteins move and flex continuously, repositioning of the catalytic atoms requires altering the conformational landscape of the enzyme. The shortest path map (SPM) methodology identified the positions that contribute most strongly to conformational dynamics in SABP2 and HNL3V and replaced those amino acids that differed between them within the SPM closest to the active site. SPM starts with a molecular dynamics simulation and identifies which residues move together and have a higher contribution to the conformational dynamics.^[1,4] Importantly, SPM predicts which amino acids outside the active site can alter the positioning of the catalytic groups within the active site.

The SPMs constructed for HNL3V and SABP2 revealed differences in the motions involving the catalytic residues, Fig. 3. Based on those SPM's, two regions were identified for mutagenesis. In the first region two substitutions are expected to add a correlated motion with OX1 and in the second region three substitutions are expected to remove a correlated motion to the catalytic Asp207. The SPM of SABP2 indicates that movements of OX1 (Ala13) are directly connected to motions of residues Cys14, Gly12, Ser179, and Leu 82, Supplementary Fig. 2. The first two residues are conserved between SABP2 and HNL3V, but the second two residues differ. Positions Ser179 and Leu82 in SABP2 correspond to Gly176 and Cys81 in HNL3V. Residue Cys81 does not appear in the SPM of HNL3V indicating that its motions are not strongly correlated with the movements of any other residues including OX1 (Ile12). While the motion of Ser179 is directly correlated to OX1 (Ala13) in SABP2, the motion between OX1 (Ile12) and Gly176 in HNL3V differs because it correlates indirectly via Cys13. Therefore, SPM analysis predicted that the C81L and G176S substitutions in HNL3V would create motions directly correlated with the oxyanion hole residue Ile12 and may fix its orientation. OX1 (Ala13), which corresponds to Ile12 in HNL, is also part of the correlated motion in SABP2. We did not include an Ile12Ala substitution in the variants because the Ala13Leu substitution had no effect on esterase activity.^[23]

The second region for mutagenesis involves removing correlated motions from HNL3V. The residues 104-106 in HNL3V are located at the loop connecting the β -sheet β_6 and the beginning of the lid domain. In HNL3V, Asn104 is directly connected to Trp203, whose movement is correlated to Thr204, Gln206, and the catalytic Asp207. The SPM of SABP2 shows no such correlations. Residues Ala105 and Ala106 (correspond to Asn104 and

Ser105 in HNL3V) do not appear in the SPM indicating that their motions are not strongly correlated to any other residues. Residue Phe107 (corresponds to Val106 in HNL3V) appears in the SPM, but in contrast to HNL3V, its motion is not correlated with the catalytic aspartate. Therefore, the SPM predicted that substitutions Asn104Ala, Ser105Ala, Val106Phe in HNL3V would remove the connection of motions between this loop and the catalytic aspartate.

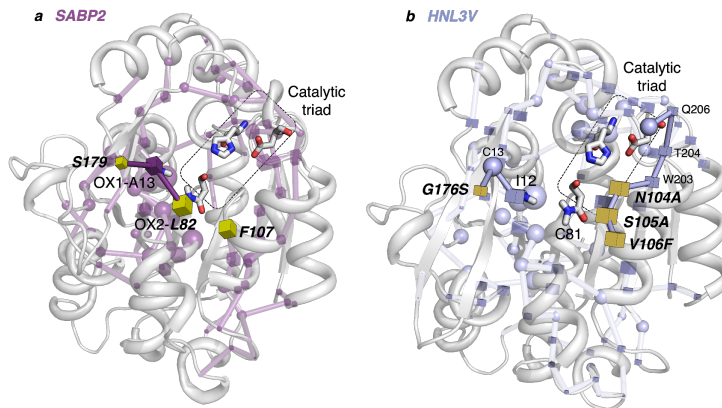


Fig. 3 | Shortest path maps of (a) SABP2 and (b) HNL3V showing five substitutions predicted to make the conformational dynamics of catalytic residues in HNL3V more like that of SABP2. The substitutions Gly176Ser and Cys81Leu add a connection to OX1 (Ile11) similar to that present in SABP2 where OX1 corresponds to Ala13. The substitutions at 104-106 in HNL3V remove a connection to the catalytic Asp207 that is present in HNL3V, but absent in SABP2. Adding these five substitutions to HNL3V created HNL8V. The spheres in the SPM indicate residues conserved between the two proteins, while cubes indicate residues that differ between the two proteins. Catalytic residues and the amides of the oxyanion hole are shown in sticks.

Four of the five SPM mutations predicted to fix the oxyanion hole orientation and the correlated motions of the catalytic aspartate are in the second or third shell outside the active site (Fig. 4). Residues 104-106 are on a loop adjacent to the catalytic serine and histidine. The G176S substitution is on a loop adjacent to one of the oxyanion hole residues. The fifth of the five substitutions is in the active site since it replaces the oxyanion hole residue Cys81 with leucine. Variant HNL8V contains the HNL3V substitutions and all five substitutions predicted by the SPM. Variant HNL7V contains the HNL3V substitutions and only four of the substitutions predicted by the SPM; the Ser105Ala substitution is omitted because the similarity of serine and alanine suggested that this substitution may be less important.

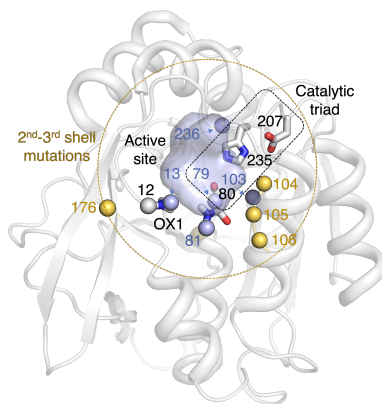


Fig. 4 | Location of the five substitutions (C81L N104A, S105A V106F G176S) added to HNL3V to create HNL8V. The catalytic triad residues (Ser80, Asp207, His235) and the oxyanion hole residues (Ile12, Cys81Leu) are blue spheres at the Ca with the side chains shown as sticks. The initial set of three substitutions (blue spheres at the Ca) were next to catalytic residues: Thr11Gly next to Ile12, Glu79His next to Ser80, Lys236Met next to His235. Four of the substitutions predicted by the SPM (yellow) are 2nd and 3rd shell changes outside the loops holding the catalytic residues: Gly176Ser outside Ile12, Asn104Ala, Ser105Ala, Val106Phe outside Ser80 and His235. One of the substitutions predicted by the SPM was the replacement of a catalytic residue (Cys81Leu, OX2) next to the catalytic serine (Ser80). Stabilizing substitution His103Val is shown as a blue sphere at the Ca. Part of the main chain trace is not shown for clarity; in reality the active site is buried.

Designed substitutions yield an esterase more efficient than SABP2

The predicted variants HNL8V (HNL3V plus the five predicted SPM substitutions) and HNL7V (omit the Ser105Ala substitution from HNL8V) both proved to be good esterases, Fig. 5, Table 1. HNL7V was 50-fold more catalytically efficient ($k_{\text{cat}}/K_{\text{M}}$ of 25,000 $\text{M}^{-1} \text{min}^{-1}$) than HNL3V and 290-fold more efficient than *HbHNL*. HNL8V showed a slightly lower catalytic efficiency ($k_{\text{cat}}/K_{\text{M}}$ of 23,000 $\text{M}^{-1} \text{min}^{-1}$), but the k_{cat} was 2-fold higher (k_{cat} of $8.1 \pm 0.4 \text{ min}^{-1}$). HNL6V, containing substitutions C81L, N104A, and G176S, showed a 2.4-fold improvement in turnover (k_{cat} of $0.59 \pm 0.14 \text{ min}^{-1}$) and a 100-fold improvement in binding (K_{M} of $0.03 \pm 0.003 \text{ mM}$) relative to *HbHNL*, resulting in a 240-fold improvement in catalytic efficiency ($k_{\text{cat}}/K_{\text{M}}$ of 20,000 $\text{M}^{-1} \text{min}^{-1}$).

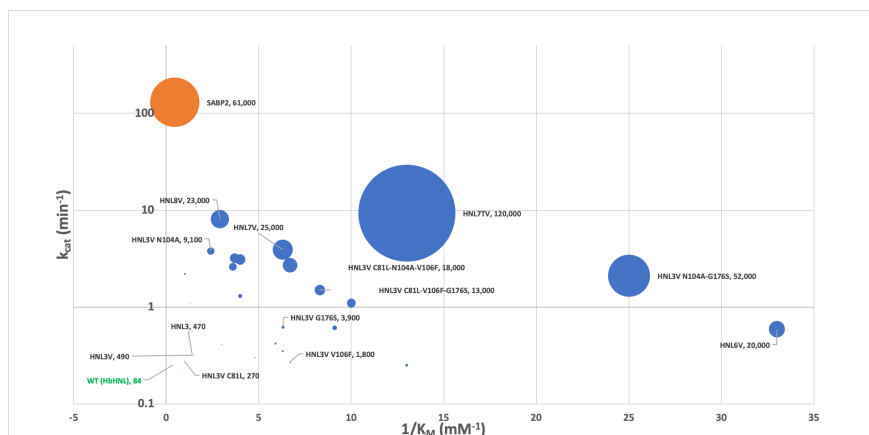


Fig. 5 | Protein engineering of *HbHNL* (green text) for improved esterase activity (k_{cat}/K_M represented by ball size) yielded a variant (largest blue ball) with two-fold better catalytic efficiency than SABP2 (orange ball). Most variants showed better binding than SABP2, i.e. are further right on the x-axis, but the catalytic step was slower, i.e. lower on the y-axis. All *HbHNL* variants are shown with blue balls. The k_{cat}/K_M values in units of $M^{-1} \text{min}^{-1}$ are shown for selected variants.

Of the four substitutions added to HNL3V to create HNL7V, the Asn104Ala substitution had the largest effect on catalytic activity. Excluding HNL7TV (discussed below), 9 of the 10 fastest variants and the 5 most efficient variants all contained N104A. HNL3V N104A's turnover rate was as fast as HNL7V (k_{cat} of 3.8 ± 0.2 vs. $3.9 \pm 0.4 \text{ min}^{-1}$), though its catalytic efficiency (k_{cat}/K_M of 9,100) was nearly 3-fold lower because of poorer binding (K_M of 0.42 ± 0.08 vs. 0.16 ± 0.07 for HNL7V). All of the intermediate variants between HNL3V N104A and HNL7V - those containing some combination of mutations C81L, N104A, V106F, and G176S - showed lower activity and poorer efficiency than HNL3V N104A and HNL7V. Mutations C81L, N104A, V106F, and G176S therefore interact epistatically and result in non-linear fitness effects. To confirm the importance of N104A for esterase catalysis, we made the reverse substitution in SABP2 to create SABP2 A105N and found a 4.7-fold decrease in enzyme turnover (k_{cat} dropped from 134 to 29 min^{-1}). A decrease in binding (K_M increased from 2.2 to 3.0 mM) resulted in a 6.3-fold decrease in catalytic efficiency (k_{cat}/K_M dropped from 61,000 to $9,600 M^{-1} \text{min}^{-1}$) as compared to SABP2. Relative to the best *HbHNL*-based variant that does not contain N104A (HNL3V V106F-G176S), SABP2 A105N showed a 9-fold higher turnover rate. We expected a larger decrease in catalysis in SABP2 A105N given the importance of N104A for improved esterase activity in *HbHNL*-based variants. This smaller than expected change suggests that epistasis plays a role in esterase activity for both SABP2- and *HbHNL*-based variants.

The best variant, HNL7TV, contained eight substitutions and was 1390-fold more catalytically efficient ($k_{\text{cat}}/K_{\text{M}}$ of 120,000 $\text{M}^{-1} \text{min}^{-1}$) than *HbHNL* and two-fold more catalytically efficient than the benchmark esterase SABP2 (Fig. 5). We made the N104T substitution because although SABP2 contains Ala at position 105, Thr is the most highly conserved amino acid at that position among homologous esterases (Supplementary Fig. 3). Thus the N104T mutation was expected to yield soluble, active protein. HNL7TV showed the highest activity (k_{cat} of 9.3 ± 0.3) of all variants, a 37-fold improvement, and a 28-fold improvement in binding compared to *HbHNL* (K_{M} of $0.08 \pm 0.04 \text{ mM}$).

All of the *HbHNL* variants containing SPM-predicted substitutions demonstrated an improvement in substrate binding and catalytic efficiency over *HbHNL*. Notably, 65% and 59% of the variants showed at least 5-fold enhancements in binding over *HbHNL* and SABP2, respectively. Despite the increase in catalytic activity, the fastest variants did not necessarily translate to superior binders; only three out of the ten fastest variants ranked in the top ten for binding and we found no correlation ($R^2 = 0.01$) between k_{cat} and K_{M} (Supplementary Fig 4, Supplementary Table 2).

Molecular dynamics simulations reveal restored oxyanion hole

Molecular dynamics (MD) simulations confirm a distorted OX1 N positioning in HNL3V as compared to SABP2 (Fig. 6a) similar to that seen in the x-ray structure comparisons of *HbHNL* and SABP2 above (Fig. 2). Multiple replica nanosecond timescale MD simulations reveal a distribution in the distance between OX1 N and the serine Ca of ca. 5.5 Å in SABP2 as compared to 5.8 Å in the x-ray, but a much longer distance of ca. 6.2 Å for HNL3V as compared to 6.8 Å in the x-ray of *HbHNL*. This shorter distance in SABP2 allows OX1 N to orient properly to stabilize the developing negative charge during catalysis.

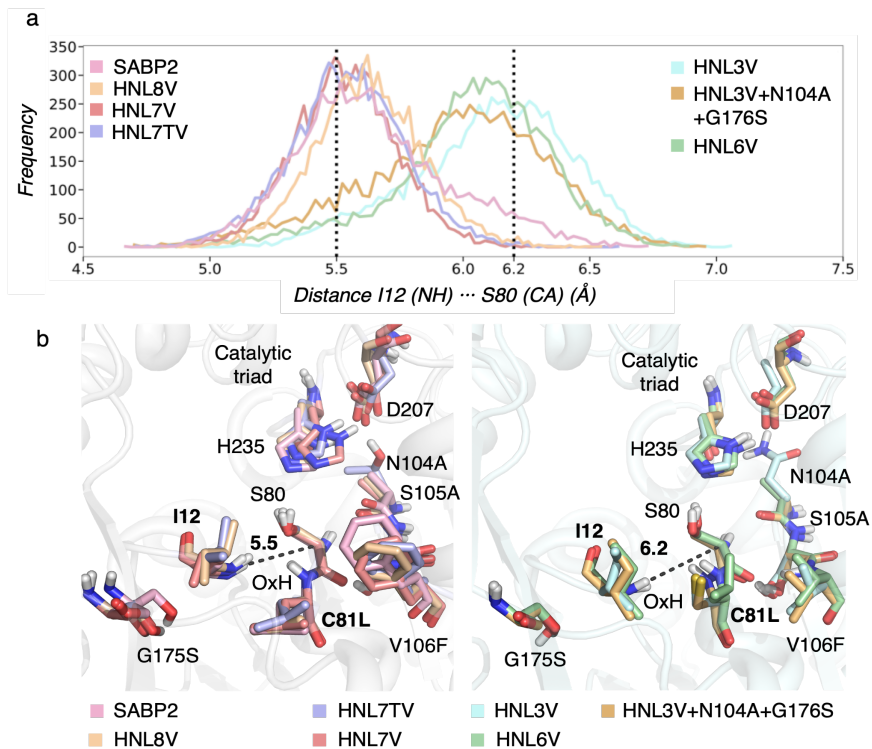


Fig. 6 | The OX1 N position of the most active variants mimic that of SABP2. **a**, Histogram of the distances (in Å) between the amide backbone of OX1 N (Ile12 or Ala13) and Ca of the catalytic serine (Ser80 or 81) for the following proteins: SABP2 (pink), HNL3V (blue), HNL3V N104A G176S (orange), HNL6V (green), HNL7V (dark pink), HNL7TV (purple), *HbHNL8V* (brown). **b**, Overlay of most populated conformations of the variants presenting properly preorganized oxyanion hole residues (i.e., distances of ca. 5.5 Å, left panel): SABP2 (pink), HNL7V (dark pink), HNL7TV (purple), HNL8V (brown), and those presenting a non-optimal oxyanion hole positioning (i.e., distances of 6.2 Å): HNL3V (blue), HNL3V N104A G176S (orange), HNL6V (green). The labels refer to *HbHNL* residue numbering.

MD simulations of many variants displaying higher levels of esterase activity matched the OX1 N positioning in SABP2, while those with lower esterase activity matched the OX1 N positioning in HNL3V (Fig. 6). An overlay of the most populated conformations in the multiple replica nanosecond timescale MD simulations confirm a restored oxyanion hole similar to that found in SABP2 for HNL7V, HNL7TV, and HNL8V. HNL3V and

HNL6V do not present a catalytically productive positioning of the oxyanion hole residues, in line with their inferior esterase catalytic efficiency.

However, the OX1 N positioning in the MD simulation does not match the esterase activity in several cases. The second most efficient variant, i.e., HNL3V N104A-G176S, adopts longer distances between the catalytic serine and the oxyanion hole residue similar to those found in the less efficient HNL3V and HNL6V. Similarly, HNL7V and HNL7TV show similar OX1 N positioning, but their esterase activity differs. Thus, the properly restored oxyanion hole explains only part of the enhancements in esterase activity.

MD simulations identify changes in the pK_a of catalytic aspartate

One of the substitutions that enhanced esterase catalytic efficiency is N104A/T, which lies close to the catalytic histidine and aspartate. The electrostatic environment of the catalytic triad has a profound effect on the catalytic activity of cysteine and serine proteases.^[24, 25] The catalytic Asp in the serine peptidase trypsin and a cysteine peptidase of the papain superfamily lie in different electrostatic environments creating different pK_a values for the Asp-His of the triad. MD simulations of variants with and without the N104A substitution show similar positions of OX1 N indicating that this substitution does not affect the oxyanion hole reorganization.

We hypothesize that the close location of N104A/T to the catalytic aspartate alters its local environment and pK_a value. We estimated the pK_a of Asp using the deep learning approach pKaI^[26] at each frame of an MD simulation and validated our predictions using constant pH MD simulations^[25] (Fig. 7). The catalytic aspartate in SABP2 was highly flexible causing changes to the local environment of aspartate and its pK_a . The predicted pK_a values for the catalytic aspartate varied between 2 and 7. For the HNL variants containing the N104A/T substitution, the pKaI predictions match constant pH MD simulations, both predicting a pK_a of ~5 for the catalytic aspartate. The catalytic aspartate is less flexible in these variants and the range of predicted pK_a values is narrower than that for SABP2. In contrast, HNL3V, which lacks the N104A substitution, has an estimated pK_a of <3. This shift in pK_a is consistent with the altered solvation and the lower catalytic activity. The catalytic aspartate maintained its hydrogen bond to the catalytic histidine throughout all the simulations in all cases.

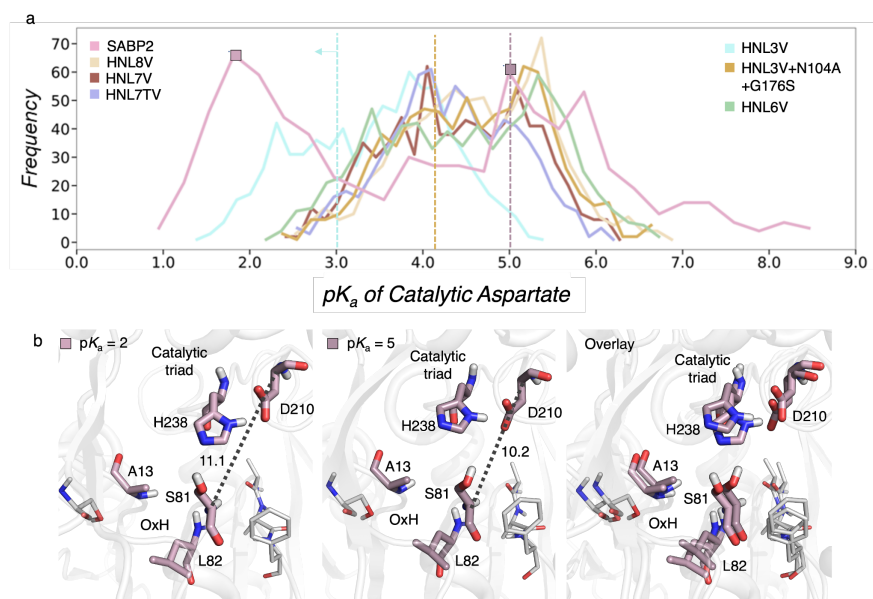


Fig. 7 | Estimation of the pK_a of the catalytic aspartate in SABP2 (pink) and HNL variants. a. Histogram of the pK_a values of the catalytic aspartate as predicted by pKaI at multiple frames of the MD simulation of SABP2 (pink), HNL3V (cyan), HNL3V N104A G176S (gold), HNL6V (green), HNL7V (brown), HNL7TV (purple), and HNL8V (light gold). The pK_a values estimated by constant pH MD simulations are marked with a vertical line for SABP2 (pink), HNL3V (cyan), and HNL3V N104A G176S (gold). b. Representative SABP2 conformation presenting a low pK_a value of ca. 2 (left panel), higher pK_a of ca. 5 (middle panel), and overlay of both conformations (right panel). The distances between the carbon alpha of the catalytic D210 and S81 are shown in Å.

X-ray structure of HNL6V reveals aspartate hydrogen bond network

To confirm the changes in OX1 N positioning and solvation of the catalytic aspartate, we solved the x-ray crystal structure of HNL6V, which contains three of the five substitutions that were added to HNL3V to create HNL8V. Substitutions C81L, N104A, and G176S are present in HNL6V, but substitutions V106F and S105A are missing.

The structure of HNL6V aligns closely with the structure of wild-type *HbHNL*. The catalytic domain (residues 1-114, 179-264) adopts the α/β -hydrolase fold and contains the catalytic triad and oxyanion hole residues. The lid or cap domain (residues 115-178) covers the active site to create a substrate binding pocket. The seven amino acid substitutions in HNL6V were in and around the active site, and none of the residues were on the outer protein

surface. Six of the substitutions are in the catalytic domain; only Gly176Ser is in the lid domain.

The OX1 N in HNL6V has moved closer to the position of SABP2, but remains intermediate between *HbHNL* and SABP2 (Fig. 8). The distance from OX1 N in HNL6V and OX1 N in SABP2 is 0.9 Å, while the corresponding distance between the three *HbHNL* structures and SABP2 is longer: 1.2 ± 0.2 Å. The internal distance between OX1 N and serine Ca is 6.2 Å in HNL6V, which is intermediate between that in SABP2 (5.8 Å) and the three *HbHNL* structures (6.8 ± 0.2 Å), see Fig. 2 above. Thus, the substitutions in HNL6V moved the main chain nitrogen (OX1 N) closer to its position in SABP2.

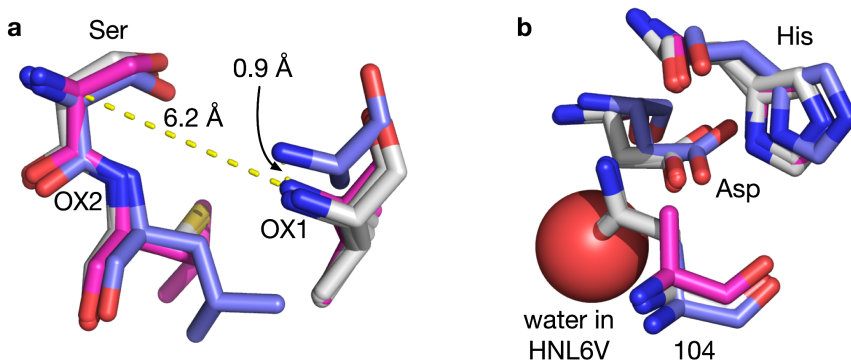


Fig. 8 | Changes in (a) the position of OX1 N and (b) the solvation of the catalytic aspartate in the x-ray structure of HNL6V (magenta carbons) as compared SABP2 (blue carbons) or *HbHNL* (three structures, white carbons). a The position of OX1 N in HNL6V has moved closer to the corresponding position in SABP2. The distance between OX1 N in SABP2 and HNL6V (0.9 Å) is shorter than the distance between SABP2 and the three *HbHNL* structures (1.2 ± 0.2 Å). The internal distance between OX1 N and serine Ca is 6.2 Å in HNL6V, which is intermediate between that in SABP2 (5.8 Å) and the three *HbHNL* structures (6.8 ± 0.2 Å). **b** The Asn104Ala substitution in HNL6V creates space for a water molecule (red sphere) near the catalytic aspartate. This space is blocked in *HbHNL* by the side chain of the asparagine. The alignment of the structures minimized the RMSD of the corresponding Ca atoms in the entire protein.

The Asn104Ala substitution in HNL6V created space for a water molecule that hydrogen bonds to the catalytic aspartate. In the *HbHNL* structures, the asparagine residue fills this region, but does not interact with the catalytic aspartate. Upon replacement of Asn104 with alanine, more space is available. A water occupies this region in the x-ray structure of HNL6V and contributes a hydrogen bond to the catalytic aspartate (O-O distance is 3.1 Å).

Discussion

Previous computational designs have required additional experimental optimization to reach catalytic efficiencies comparable to Nature's enzymes. Computational design of a luciferase using Rosetta combined with deep learning yielded impressive catalytic efficiencies of $10^6 \text{ M}^{-1}\cdot\text{s}^{-1}$ but this design also included experimental optimization of the ligand-binding pocket.^[27] A bioinformatics-based design of a hydroxynitrile lyase from an esterase yielded a catalytic efficiency of $3,300 \text{ M}^{-1}\cdot\text{s}^{-1}$.^[28] It required >120 substitutions, which make it difficult to explain how each substitution contributes to catalysis. The computational design of multi-step reactions has been less successful. A designed retroaldolase showed a k_{cat}/K_M of $0.2 \text{ M}^{-1}\cdot\text{s}^{-1}$ and reached a k_{cat}/K_M of $34,000 \text{ M}^{-1}\cdot\text{s}^{-1}$ only after multiple rounds of directed evolution.^[29, 30] Designed esterases based on a designed cysteine-histidine dyad and oxyanion hole yielded catalytic efficiencies of k_{cat}/K_M of $10\text{-}400 \text{ M}^{-1}\cdot\text{s}^{-1}$.^[12] One attempt at esterase design with serine-histidine-aspartate catalytic triads failed to complete a catalytic cycle. The enzymes could only react irreversibly with fluorophosphonate probes.^[9] A more recently designed esterase showed a catalytic activity 1000-times lower than commercially available esterases, but saturation kinetics were not reported.^[8] A Kemp eliminase also required eight substitutions outside the active site for high catalytic efficiency, but these were not rationally predicted, but found with experimental directed evolution.^[31] This inability to design efficient enzymes limits new applications of enzymes in medicines, non-polluting manufacture of fine chemicals and pharmaceuticals, food processing, and biodegradation of environmental contaminants.

The low catalytic efficiencies achieved by computational enzyme design have been associated with non-optimal arrangements of the catalytic residues for transition state(s) stabilization, the lack of a proper description of the conformational changes key for substrate binding and product release, and the limitation of introducing mutations in the active site pocket only.^[1] Our approach of identifying the correlated motions established by catalytic residues with second and third shell mutations has achieved catalytic efficiencies surpassing that of the reference SABP2 enzyme.

Comparison of the shortest path maps for HNL3V and SABP2 revealed differences in correlated movements in the two enzymes. To engineer increased esterase activity into HNL3V, we focused on correlated movements connected to the active site residues. HNL3V contained one movement associated with catalytic aspartate that was missing from SABP2. We hypothesized that this movement should be removed from HNL3V to increase esterase activity. HNL3V also lacked a correlated movement associated with an oxyanion hole residue that was present in SABP2. We hypothesized that this movement should be added to HNL3V to increase esterase activity. To add or remove movements, we replaced residues in HNL3V with the corresponding residues from SABP2. Only residues within the SPM of either HNL3V or SABP2 were changed. Since HNL3V and SABP2 share 45% sequence identity,

only five amino acid substitutions were required. Four of the five substitutions were outside of the active site demonstrating that the shortest path maps identify residues outside the active site that contribute to catalysis.

The resulting variant, HNL8V, showed a 25-fold increase in both catalytic rate (k_{cat}) and a 1.9-fold improvement in K_M for esterase catalysis demonstrating the value of the SPM-based predictions. The interactions between the substitutions showed negative cooperativity with respect to K_M , but positive cooperativity with respect to k_{cat} . Individually, the five substitutions showed modest improvements in K_M (mean of 2.5 ± 1.6 -fold improvement). If the improvements act additively, then HNL8V should show a 34-fold improvement in K_M , but it showed only a 1.9-fold improvement. The eighteen-fold lower observed value indicates negative cooperativity between the five substitutions with respect to K_M . For k_{cat} , four of the substitutions showed modest changes, but N104A showed a twelve-fold improvement (mean = 3.2 ± 4.9 -fold improvement). If the improvements act additively, then HNL8V should show a 8.1-fold improvement in k_{cat} , but it showed a 25-fold improvement. The 3.1-fold higher observed value indicates positive cooperativity between the five substitutions with respect to k_{cat} .

This cooperativity is consistent with the notion that cooperative movements cause the changes in esterase activity. The predicted SPM substitutions – C81L, V106F, and G176S – interact epistatically, Fig. 9. The effect of all three substitutions combined is more than twice the effect of the sum of the three individual substitutions. The conformational changes induced by each substitution suggest a mechanism for this epistasis. The replacement of a glycine by a serine at position 176 changes the backbone conformation, which properly positions the oxyanion hole residue Ile12 as (OX1)-Ala13 in SABP2. Fixing the orientation of the other oxyanion hole residue Cys81 (OX2-Leu82 in SABP2) requires both Cys81Leu and Val106Phe. The side chain of Leu81 can adopt two different conformations in HNL3V; one conformation hinders catalysis as it blocks access of the ester substrate to the active site pocket. Mutation Val106Phe restricts the side chain of Leu81 to the conformation that allows ester binding for catalysis. As noted in the results section, the effect of the mutation N104A/T is not connected to the oxyanion hole reorganization, but rather to the change of the electrostatic environment of the catalytic aspartate. Variants containing the key mutations for fixing the oxyanion hole (C81L, V106F, and G176S) together with N104A/T show the highest esterase catalytic efficiency.

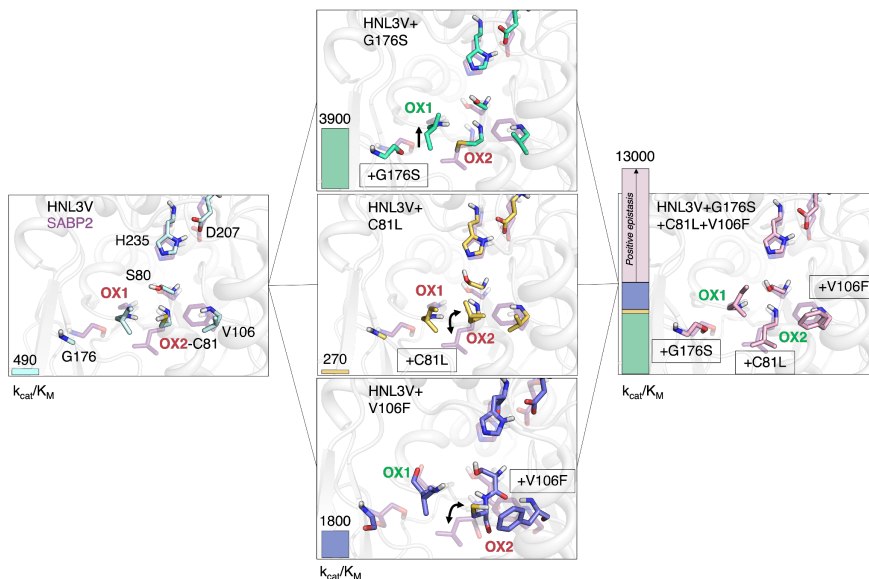


Figure 9 | Substitutions that enhance catalytic activity act cooperatively. Representative conformations for the HNL variants: starting variant HNL3V (left panel, light blue carbons), the singly mutated variants HNL3V+G176S (center panel, green carbons), HNL3V+C81L (gold carbons), HNL3V+V106F (dark blue carbons), and the triple variant (right panel, pink carbons). All panels include a representative conformation of the active site and oxyanion hole residues of SABP2 (purple carbons) and a bar at the left side indicating the catalytic efficiency (k_{cat}/K_M , $M^{-1}\cdot\text{min}^{-1}$) of each HNL variant. The combined effect of the three individual mutations is higher than their sum, due to positive sign epistasis highlighted in pink and with an arrow. The color of the labels of the oxyanion hole residues (OX1, OX2) indicate a proper (green) or bad (red) orientation as compared to SABP2. Double arrows mark the different conformations of the sidechain of C81 with respect to SABP2.

Combining the SPM analysis with a multiple sequence alignment of esterases yielded the best variant, HNL7TV. The substitution N104A yielded the largest increase in k_{cat} of all the substitutions tested. The multiple sequence alignment showed that most esterases contained threonine, not an alanine, at this position. The substitution N104T yielded the best variant with approximately twice the catalytic efficiency (k_{cat}/K_M) of SABP2.

Although the shortest path maps combined with sequence comparison identified the substitutions needed to improve catalytic activity, they did not identify why the substitutions increased catalytic activity. Gaining insight on why esterase activity increased required additional experiments combined with computational modeling. Molecular dynamics simulations support the notion that substitutions have repositioned the main chain of the oxyanion hole residue, Ile12. The 25-fold increase in k_{cat} from HNL3V vs. HNL7TV is

consistent with previous experiments that disrupted the oxyanion hole in enzymes using site-directed mutagenesis. Removing an N-H from the oxyanion hole in subtilisin lowered k_{cat} approximately 100-fold.^[32, 33] Experiments with ketosteroid isomerase^[34] and a decarboxylase^[35] gave similar estimates for the contributions of an oxyanion hole N-H to catalysis.

Both molecular dynamics simulations and an x-ray structure of HNL6V show changes in the local environment of the catalytic aspartate that might impact its pK_a . The mutation of the catalytic aspartate in trypsin by an asparagine lowers the pK_a of histidine by 1.5 pH units and dramatically reduces the catalytic activity.^[36, 37] One mechanism to explain the role of the catalytic aspartate in serine proteases is the formation of a low-barrier hydrogen bond between His-Asp, especially in the transition state or in enzyme-intermediate complexes.^[38, 39] This mechanism requires that both histidine and aspartate present similar pK_a values, which for aspartate was estimated to be around 6.7.^[40] The increase in catalytic efficiency observed for the variants containing N104A/T directly in contact with the catalytic aspartate suggest that this mutation impacts the pK_a of the catalytic aspartate. Our calculations predict higher pK_a values for aspartate especially in SABP2, and in the most efficient variants HNL7V, HNL7TV and HNL8V. This finding is in line with the requirement of matching pK_a values between His-Asp for the formation of a low-barrier hydrogen bond at the transition states and/or tetrahedral intermediates formed along the multistep esterase mechanism.

There must be additional contributions to esterase catalysis in SABP2 besides those identified here. The k_{cat} of the best variant, HNL7TV, is still about thirteen-fold lower than that for SABP2 indicating that additional substitutions are needed to fully match the k_{cat} of SABP2. Even more distant substitutions are likely required to modulate the conformational landscape of the enzymes and generate a SABP2-like environment of the catalytic His-Asp dyad for esterase catalysis. The higher flexibility of the catalytic aspartate in SABP2 identified in our MD simulations may contribute to its higher k_{cat} .

Methods

Chemicals were purchased from commercial suppliers and used without further purification.

Site-directed mutagenesis to create enzyme variants

The gene encoding the wild-type HNL from *Hevea brasiliensis* in the pSE420 plasmid^[41] was recloned into a pET21a(+) plasmid. *HbHNL* enzyme variants were constructed via inverse PCR^[42] using non-overlapping mutagenic primers in a sequential manner (Supplementary Table 3). Briefly, mutagenic primers anneal to the plasmid template in a back-to-back, outward-facing orientation and are amplified using New England Biolabs (NEB) Q5 HiFi polymerase (M0491S) to produce a linear, double-stranded DNA product containing the desired mutation(s). Primers were designed using the NEBaseChanger (<https://>

nebasechanger.neb.com) web tool and checked for secondary structure and self-dimer/heterodimer propensity with IDT's OligoAnalyzer Tool (<https://www.idtdna.com/pages/tools/oligoanalyzer>). Primers were purchased from Integrated DNA Technologies (Coralville, IA) and used without further purification. PCR was performed using a BioRad 2000 Thermal Cycler with the following conditions: initial denaturation at 98 °C for 30 sec, 30 cycles of denaturation (98 °C for 30 sec), annealing (calculated annealing temperature for 25 sec), and extension (72 °C for 150 sec), and a final extension step of 72 °C for 2 minutes. The PCR products were treated with a KLD enzyme mix (NEB M0554S), which phosphorylates the 5' ends of the linearized PCR products, ligates the phosphorylated ends, and degrades the original plasmid template. Five µl of the KLD product was used directly to transform chemically-competent *Escherichia coli* DH5α cells (NEB C2988) according to the manufacturer's protocol and plated on lysogeny broth (LB) plates containing 100 µg/ml carbenicillin. After overnight growth at 37 °C, individual colonies were picked and grown up overnight in LB media, and the plasmids were extracted via NEB Monarch Plasmid mini-prep kit (NEB T1010). Plasmid concentrations were measured spectrophotometrically at 260 nm via a Nanodrop 2000 (Thermo Scientific) and diluted to <1.0 OD units if necessary. Sanger sequencing from Genewiz/Azenta Life Sciences was used to confirm mutations. The sequence-confirmed plasmid was transformed into *Escherichia coli* strain BL21(DE3) chemically competent cells (NEB C2527) according to NEB's transformation protocol and plated on LB plates containing 100 µg/ml carbenicillin.

SABP2 A104N was constructed via isothermal assembly (also called Gibson assembly)^[43] using a gene fragment ordered from Twist Biosciences (San Francisco, CA). We used the NEB Gibson Assembly® Master Mix kit to perform the assembly under the following conditions: template DNA (pET21a(+)) plasmid containing SABP2 gene, enzyme master mix, and the synthesized gene fragment were incubated in a thermocycler for 15 minutes at 50°C. 2 µl of the assembly reaction mixture was used directly for transformation into NEB 5-alpha competent *E. coli* included in the assembly kit. All subsequent steps, i.e. expression, purification, and assays, are as described above. For detailed information on the protocol, including primer design, please see NEB's Gibson Assembly® Application Overview website (<https://www.neb.com/applications/cloning-and-synthetic-biology/dna-assembly-and-cloning/gibson-assembly>).

Protein expression and purification

LB media containing carbenicillin (100 µg/ml, 5 ml) was inoculated with a single bacterial colony from an agar plate and incubated in an orbital shaker at 37 °C and 240 rpm for 15 h to create a seed culture. A 1-L baffled flask containing terrific broth-amp media (250 ml) was inoculated with 2.5 mL of seed culture. The pre-induction culture was incubated at 37 °C and 240 rpm for 3–4 h until the absorbance at 600 nm reached 0.4–1.0. The culture was then transferred to ice for 30 minutes to cool. Isopropyl β-D-1-thiogalactopyranoside (0.75–1.0 mM final concentration) was added to induce protein expression, and cultivation was

continued for 20-24 h at 18°C. The cells were harvested by centrifugation (7000 rpm, 15 min at 4 °C), resuspended in NiNTA loading buffer (10 mM imidazole, 50 mM Tris pH 8.0, 500 mM NaCl, 4 ml/g of wet cells), and either directly sonicated or frozen for storage and later purification. Cells were flash frozen in liquid nitrogen or a dry ice-ethanol bath and stored at -80°C. Frozen cells were thawed at room temperature or in a room temperature water bath, and fresh/thawed cells were disrupted by sonication (400 W, 40% amplitude for 3 min). The cell lysate was centrifuged to pellet the cell debris (4 °C, 20,000 rcf for 20 min) and the supernatant was mixed with 1-2.5 ml of NiNTA resin (pre-equilibrated with 10 ml of NiNTA loading buffer) and incubated for 45 minutes at 4 °C with rotation (10 rpm). The resin/supernatant mixture was loaded onto a 25 ml column (Bio-Rad) and the resin was washed with 10 column volumes each of buffer containing increasing amounts of imidazole (25-50 mM imidazole, 50 mM Tris pH 8.0, 500 mM NaCl). The His-tagged protein was eluted with 10 column volumes of elution buffer (125 mM imidazole, 50 mM Tris pH 8.0, 500 mM NaCl) and collected in 1 ml fractions. The protein concentration of each elution fraction was determined from spectrophotometric measurements at 280 nm via Nanodrop 2000 (Thermo Scientific). The calculated extinction coefficient was determined using the ProtParam web tool (<https://web.expasy.org/protparam/>). Protein gels were used to check for the presence and purity of protein and run using sodium dodecyl sulfate polyacrylamide gradient gels (NuPage 4–12% Bis-Tris gel from Invitrogen) using the Precision Plus Dual Color protein standard (BioRad, 5 µl/lane), run for 50 min at 120V, stained with SimplyBlue Safe Stain (Thermo Fisher Scientific), and destained 2x with milliQ UltraPure H₂O. SDS-PAGE indicated a molecular weight of ~30 kDa in agreement with the predicted weight of 31.1 kDa. The imidazole-containing elution buffer was exchanged by addition of BES buffer (5 mM *N,N*-bis(2-hydroxyethyl)-2-aminoethanesulfonic acid, pH 7.2, 14 ml) followed by ultrafiltration (Amicon 15-ml ultrafiltration centrifuge filter, 10 kDa cutoff) to reduce the volume to ~250 µl. This addition of buffer and filtration was repeated four times. A 250-ml culture typically yielded 2-5 mg of protein.

Enzyme assays

Enzyme activity was monitored at room temperature (typically 22±2 °C) in triplicate for 10 min using a SpectraMax 384 Plus microplate reader. Ester hydrolysis activity was measured at 405 nm using *p*-nitrophenyl acetate (pNPAc), which releases the yellow *p*-nitrophenoxide. The reaction mixture (100 µL; path length 0.29 cm) contained 0.01-7.0 mM pNPAc, 6–8% v/v acetonitrile, 5 mM BES buffer, pH 7.2, and up to 15 µg enzyme. The slope of increase in absorbance versus time was measured in triplicate, fit to a line using linear regression, and corrected for spontaneous hydrolysis of pNPAc with blank reactions lacking protein, also measured in triplicate. The extinction coefficient used for calculations ($\epsilon_{405 \text{ nm}} = 11,588 \text{ cm}^{-1} \text{ M}^{-1}$) accounts for the incomplete ionization of *p*-nitrophenol at pH 7.2. For steady-state kinetic measurements, the enzyme concentration was determined by average absorbance at 280 nm measured in duplicate and normalized by subtracting a buffer blank. The enzyme

concentrations in the assay solution ranged from 50 nM to 5 μ M. k_{cat} and K_M were determined using a non-linear fit of the experimental data to the Michaelis-Menten equation using the solver program in Microsoft Excel or using the statistical program R (Huitema & Horsman, 2019).^[44]

An alternative assay protocol (the KP protocol) results in faster rates relative to the BES protocol described in the previous paragraph. The reaction mixture volume, substrate concentrations, amount of enzyme, and absorption wavelength are the same between both assays. In the KP protocol, the substrate is dissolved in methanol instead of acetonitrile and uses 100 mM KP buffer, pH 7.5, 1% v/v acetonitrile, and an extinction coefficient ($\epsilon_{405\text{ nm}} = 12,300\text{ cm}^{-1}\text{ M}^{-1}$) that accounts for the effect of the change in pH on absorbance.^[45] The faster rate is due to the differing organic solvents; increasing the acetonitrile concentration decreases the observed rate, as has been previously described.^[46] All kinetic parameters reported in this manuscript were obtained using the BES protocol.

Molecular modeling system preparation

The starting structures for the different enzymes (*HbHNL*, *HNL3V*, *HNL6V*, *HNL7V*, *HNL7TV*, *HNL8V*, *SABP2*) were generated with the predictions of the neural network AlphaFold 2 approach.^[47] The structures were prepared using the Python packages MDTraj,^[48] pytraj^[49] which is part of the cpptraj package,^[50] MDAnalysis,^[51] PyEMMA,^[52] and networkx.^[53]

Molecular dynamics simulation

The protocol applied for the MD equilibration phase was the one described by Roe and Brooks with small differences fine-tuned to our systems.^[54] For non-minimization steps, the bonds involving hydrogen are constrained by the SHAKE algorithm. Long-range electrostatic effects were modeled using the particle mesh-Ewald method.^[55] A 10 Å cut-off was applied to Lennard-Jones and electrostatic interactions. The MD protocol starts with the minimization phase of 1500 steps steepest descent method followed by 3500 steps of the conjugate gradient method with a positional restrain (*i.e.*, force constant of 5.0 kcal·mol⁻¹·Å⁻²) to the protein heavy atoms. Then, a heating phase is performed with increasing the temperature from 25 K to 300 K during 20 ps of MD simulation, a Langevin thermostat with a collision frequency of 5 ps⁻¹, and a positional restrain (*i.e.*, force constant of 5.0 kcal·mol⁻¹·Å⁻²) to the protein heavy atoms. The next step is the minimization and heating of all the atoms in the system. Starting with two minimization stages of 1000 steps steepest descent method followed by 1500 steps of the conjugate gradient method each with a positional restrain (*i.e.*, force constant of 2.0 kcal·mol⁻¹·Å⁻² in the first minimization and 0.1 kcal·mol⁻¹·Å⁻² in the second) to the protein heavy atoms. Then, a third minimization phase of 1500 steps steepest descent method followed by 3500 steps of the conjugate gradient method without any positional restraint is performed. Afterwards, the system is heated in the same way as previously defined. Finally, a five-round equilibration phase at the NPT ensemble with a constant pressure of 1 atm is performed: whereas the first four were done with the

Berendsen barostat, the fifth one with Monte-Carlo barostat. Langevin thermostat with a collision frequency of 1 ps^{-1} was used in the five equilibration rounds. The first two equilibration rounds of 5 ps had a positional restraint to the protein-heavy atoms with a force constant of 1.0 and $0.5 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$, respectively. A third round of 10 ps equilibration is followed with positional restraint to the backbone-heavy atoms with a force constant of $0.5 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$. The fourth equilibration round of 10 ps was performed without any restraint. The last equilibration round was of 1 ns without any restraint. The production runs were performed at the NVT ensemble with the Langevin thermostat with a collision frequency of 1 ps^{-1} during 250 ns. Finally, three replicas of equilibration and production runs were performed for each homodimer, reaching a total simulation time of 750 ns for *HbHNL*, *HNL3V*, *HNL3V_104A_176S*, *HNL6V*, *HNL7V*, *HNL8V*, *SABP2* systems, respectively. The MD trajectories were analyzed using the Python packages MDTraj,^[48] pytraj^[49] which is part of the cpptraj package,^[50] MDAAnalysis,^[51] PyEMMA,^[52] and networkx.^[53]

Constant pH Molecular dynamics simulation

Constant pH molecular dynamics simulations were done following the same protocol described in the molecular dynamics simulation section. Residues are allowed to change protonation state in the fifth equilibration and the production runs. All systems were simulated at pH values from 4.5 to 8.0 with a 0.5 spacing. The protonation state changes were attempted every 100 steps. The following 100 steps were used to relax the solvent after a successful attempt. A salt concentration of 0.1 was used. For SABP2 HID6, HID11, HID15, HIE32, HIE80, HIP113, HID158, ASP210, HID238, HID257 residues were selected to titrate, and for HNL variants HID5, HID10, HID14, HID20, HIE31, HIE79, HIP112, ASP207, HID235 were selected to titrate. Three replicas of equilibration and production run of 30 ns were performed for SABP2, HNL3V, HNL3V G176S N104A. pK_a values are estimated from the extrapolation of the sigmoidal function.

Molecular dynamics analysis

Shortest Path Map (SPM) calculations. The Shortest Path Map (SPM) analysis was performed using the MD simulations of SABP2 and HNL3V. For SPM calculation, the inter-residue mean distance and correlation matrices computed along the MD simulations need to be computed. From both matrices a simplified graph is drawn, in which only those pairs of residues displaying a mean distance shorter than 6 \AA along the MD simulation time are connected through a line. The edge connecting both residues is weighted to the Pearson correlation value ($d_{ij} = -\log |C_{ij}|$). Short lines will be drawn for those pairs of residues whose motions are more correlated. The generated graph is further simplified to identify the shortest path lengths. 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 SPM graph is then drawn on the 3D structure of the enzyme. More details about our SPM tool can be found in references 1 and 5.

X-ray crystal structure determination

HNL6V containing a C-terminal 6His tag was expressed from plasmid pET21a(+) in *Escherichia coli* BL21 (DE3). The protein was purified using nickel-affinity chromatography and concentrated to 9.3 mg/ml. Crystallographic screening was done using Phoenix crystallography dispenser from Art Robbins Instruments Inc. Sitting-drop vapor diffusion trays, the low profile INTELLI-PLATE® from Art Robbins Instruments Inc, were used for crystallization setup. All the setup and washing procedure was done through the Art Robbins Instruments software Phoenix. Each crystallization drop contained 0.1 μ l protein sample (9.3 mg/ml protein) and 0.1 μ l well solution. A total of 960 conditions were tested. Crystals appeared within one day from the Index HT screen from Hampton Research Inc, under the condition of 0.1 M Bis-Tris, pH 5.5, 2 M (NH₄)₂SO₄, and grew to the full size of 0.35 mm in three days. Two distinct crystals formed (Supplementary Fig. 5). The robotic screening crystals proved sufficient to refine the model so additional crystal screening trays were not needed.

The structural dataset was collected on beamline 24 ID-C (NE-CAT) at the Advanced Photon Source, Argonne National Laboratory (Supplementary Table 4). Crystals were transferred into cryoprotectant solutions consisting of well solution components and increasing concentration of sodium malonate. The final concentration of sodium malonate in the cryoprotectant solution was 1.2 M. Harvested crystals were flash frozen in liquid nitrogen. The datasets were collected at an oscillation angle of 0.2°. The crystal belonged to space group C222₁, with unit cell parameters a = 47.054, b = 106.378, c = 128.396 Å and, one molecule per asymmetric unit (Supplementary Table 5). The reconstructed ancestral hydroxynitrile lyase (PDB ID: 5tdx^[20]) with 75.85% sequence identity was used for molecular replacement, and refined to a 2.3 Å resolution model. This 2.3 Å initial model was used to aid in the refinement of the second and final model resulting in a 1.99 Å structure. HKL2000^[56] was used to process collected data and Phaser^[57] in Phenix^[58] was used for molecular replacement and refinement (Supplementary Table 6). Refinement modeling was performed using Coot.^[59] The structure was refined to R_{work} and R_{free} values of 0.1844 and 0.2376, respectively.

Structural refinement revealed 2F_O-F_C (blue) and F_O-F_C (red/green) electron density near the active site, located around the catalytic serine O_γ (Supplementary Fig. 6). This density suggested the presence of a bound molecule, perhaps in multiple orientations. Placing water, glycerol, malonate, or sulfate in this region did not improve the R-work and R-free statistics, nor did these placements satisfy the density. Placing water around the catalytic Serine O_γ in multiple orientations did not reduce the F_O-F_C density, suggesting the structure still required the addition of one or more molecules, and perhaps in sub-100% occupancies. Adding glycerol in multiple locations proved unsatisfactory due to the increase in red F_O-F_C around pieces of the molecule unable to fit in the 2F_O-F_C density properly. Placing glycerol in varying locations with occupancies summing to 100% in attempts to solve this issue

remained insufficient to satisfy the 2F_O-F_C density. When adding malonate, negative interactions with nearby amino acid residues became unavoidable, no matter the position or orientation of the malonate. Additionally, experimental placement of malonate with varying occupancy proved unfulfilling to the 2F_O-F_C density. Finally, the inability to find a placement of sulfate molecules that would fulfill the active site density without interaction with each other concluded attempts to place a ligand in the active site. As this electron density near catalytic serine O_γ remains unmodeled, this model should be considered a putative structure. Further refinement of other *HbHNL* structures could potentially aid in identifying the structure in the currently unresolved F_O-F_C density active site. Zuegg and coworkers^[13] also observed unidentified electron density near the active site during refinement of an x-ray crystal structure of wild-type *HbHNL*. The final model was deposited in the Research Collaboratory for Structural Bioinformatics Protein Data Bank (PDB ID: 8euo).

The three structures of wild-type hydroxynitrile lyase from *Hevea brasiliensis* without a bound ligand used for comparison have the following protein data bank IDs: 6yas,^[20] 3c6x,^[60] 2g4l.^[61] The structure of salicylic acid binding protein 2 from tobacco without a bound ligand used for comparison has the protein data bank ID 1y7h.^[14] PyMOL v.2.5.4 was used to overlay the structures using the align function and to create images of protein structures.^[62]

References

1. Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design. *WIREs Comp. Mol. Sci.* **11**, e1502 (2021).
2. Corbella, M., Pinto, G. P., & Kamerlin, S. C. L. Loop dynamics and the evolution of enzyme activity. *Nat. Rev. Chem.* **7**, 536–547 (2023).
3. Schenkmyerova, A. et al. Engineering the protein dynamics of an ancestral luciferase. *Nat. Commun.* **12**, 3616 (2021).
4. Casadevall, G., Duran, C., & Osuna, S. AlphaFold2 and deep learning for elucidating enzyme conformational flexibility and its application for design. *JACS Au* **3**, 1554–1562 (2023).
5. 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).
6. Boehr, D. D., Nussinov, R., & Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796.
7. Maria-Solano, M. A., Kinateder, T., Iglesias-Fernández, J., Sterner, R. & Osuna, S. *In silico* identification and experimental validation of distal activity-enhancing mutations in tryptophan synthase. *ACS Catal.* **11**, 13733–13743 (2021).
8. Li, G. et al. A de novo designed esterase with *p*-nitrophenyl acetate hydrolysis activity. *Molecules* **25**, 4658 (2020).
9. Rajagopalan, S. et al. Design of activated serine-containing catalytic triads with atomic-level accuracy. *Nat. Chem. Biol.* **10**, 386–391 (2014).
10. Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 14274–14279 (2001).
11. Moroz, Y. S. et al. New tricks for old proteins: single mutations in a nonenzymatic protein give rise to various enzymatic activities. *J. Am. Chem. Soc.* **137**, 14905–14911 (2015).

12. Richter, F. et al. Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *J. Amer. Chem. Soc.* **134**, 16197–16206 (2012).
13. Zuegg, J., Gruber, K., Gugganig, M., Wagner, U. G. & Kratky, C. Three-dimensional structures of enzyme-substrate complexes of the hydroxynitrile lyase from *Hevea brasiliensis*. *Protein Sci.* **8**, 1990–2000 (1999).
14. Forouhar, F. et al. Structural and biochemical studies identify tobacco SABP2 as a methyl salicylate esterase and implicate it in plant innate immunity. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1773–1778 (2005).
15. Ollis, D. L. et al. The α/β hydrolase fold. *Protein Eng. Des. Sel.* **5**, 197–211 (1992).
16. Bauer, T. L., Buchholz, P. C. F. & Pleiss, J. The modular structure of α/β -hydrolases. *FEBS J.* **287**, 1035–1053 (2020).
17. Rauwerdink, A. et al. Evolution of a catalytic mechanism. *Mol. Biol. Evol.* **33**, 971–979 (2016).
18. Devamani, T. et al. Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J. Am. Chem. Soc.* **138**, 1046–1056 (2016).
19. Nedrud, D. M. et al. Uncovering divergent evolution of α/β -hydrolases: a surprising residue substitution needed to convert *Hevea brasiliensis* hydroxynitrile lyase into an esterase. *Chem. Sci.* **5**, 4265–4277 (2014).
20. Jones, B. J. et al. Larger active site in an ancestral hydroxynitrile lyase increases catalytically promiscuous esterase activity. *PLoS One* **15**, e0235341 (2020).
21. Padhi, S. K. et al. Switching from an esterase to a hydroxynitrile lyase mechanism requires only two amino acid substitutions. *Chem. Biol.* **17**, 863–871 (2010).
22. Dadashpour, M., Fukuta, Y. & Asano, Y. Comparative expression of wild-type and highly soluble mutant His103Leu of hydroxynitrile lyase from *Manihot esculenta* in prokaryotic and eukaryotic expression systems. *Protein Express. Purif.* **77**, 92–97 (2011).
23. Park, S. W., Kaimoyo, E., Kumar, D., Mosher, S., & Klessig, D. F. Methyl salicylate is a critical mobile signal for plant systemic acquired resistance. *Science* **318**, 113–116 (2007).
24. Gisdon, F. J., Bombarda, E., & Ullmann, G. M. Serine and cysteine peptidases: So similar, yet different. How the active-site electrostatics facilitates different reaction mechanisms. *J. Phys. Chem. B* **126**, 4035–4048 (2022).
25. Hofer, F., Kraml, J., Kahler, U., Kamenik, A. S. & Liedl, K. R. Catalytic site pK_a values of aspartic, cysteine, and serine proteases: Constant pH MD simulations. *J. Chem. Inf. Model.* **60**, 3030–3042 (2020).
26. Reis, P. B. P. S. et al. A fast and interpretable deep learning approach for accurate electrostatics-driven pK_a predictions in proteins. *J. Chem. Theory Comput.* **18**, 5068–5078 (2022).
27. Yeh, A. H.-W. et al. (2023). De novo design of luciferases using deep learning. *Nature*, **614**, 774–780.
28. Nakano, S., & Asano, Y. Protein evolution analysis of S-hydroxynitrile lyase by complete sequence design utilizing the INTMSAlign software. *Sci. Rep.* **5**, 8193–10 (2015).
29. Jiang, L. et al. De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
30. Obexer, R. et al. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).

31. Broom, A. et al. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).
32. Bryan, P., Pantoliano, M. W., Quill, S. G., Hsiao, H. Y. & Poulos, T. Site-directed mutagenesis and the role of the oxyanion hole in subtilisin. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 3743–3745 (1986).
33. Brenner, C., Bevan, A. & Fuller, R. S. One-step site-directed mutagenesis of the Kex2 protease oxyanion hole. *Curr. Biol.* **3**, 498–506 (1993).
34. Kraut, D. A. et al. Testing electrostatic complementarity in enzyme catalysis: hydrogen bonding in the ketosteroid isomerase oxyanion hole. *PLoS Biol.* **4**, e99 (2006).
35. Desai, B. J. et al. Investigating the role of a backbone to substrate hydrogen bond in OMP decarboxylase using a site-specific amide to ester substitution. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15066–15071 (2014).
36. Sprang, S. et al. The three-dimensional structure of Asn¹⁰² mutant of trypsin: Role of Asp¹⁰² in serine protease catalysis. *Science* **237**, 905–909 (1987).
37. Craik, C. S., Roczniak, S., Largman, C., & Rutter, W. J. The catalytic role of the active site aspartic acid in serine proteases. *Science*, **237**, 909–913 (1987).
38. Cleland, W. W. & Kreevoy, M. M. Low-barrier hydrogen bonds and enzymic catalysis. *Science* **264**, 1887–1890 (1994).
39. Agback, P. & Agback, T. Direct evidence of a low barrier hydrogen bond in the catalytic triad of a serine protease. *Sci. Rep.* **8**, 10078 (2018).
40. Frey, P. A., Whitt, S. A. & Tobin, J. B. A low-barrier hydrogen bond in the catalytic triad of serine proteases. *Science* **264**, 1927–1930 (1994).
41. Hasslacher, M. et al. Molecular cloning of the full-length cDNA of (*S*)-hydroxynitrile lyase from *Hevea brasiliensis*. Functional expression in *Escherichia coli* and *Saccharomyces cerevisiae* and identification of an active site residue. *J. Biol. Chem.* **271**, 5884–5891 (1996).
42. Ochman, H., Gerber, A. S., & Hartl, D. L. Genetic applications of an inverse polymerase chain reaction. *Genetics*, **120**, 621–623 (1988).
43. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Meth.* **6**, 343–345 (2009)
44. Huitema, C. & Horsman, G. Analyzing enzyme kinetic data using the powerful statistical capabilities of R (preprint). BioRxiv. (2018) <https://doi.org/10.1101/316588>
45. Roa, A., Goble, M.L., Garcia, J.L., Acebal, C., & Virden, R. Rapid burst kinetics in the hydrolysis of 4-nitrophenyl acetate by penicillin G acylase from *Kluyvera citrophila*: Effect of mutation F360V on rate constants for acylation and de-acylation. *Biochem. J.* **316**, 409–412 (1996).
46. Peng, Y., Fu, S., Liu, H., & Lucia, L. A. Accurately determining esterase activity via the isobestic point of *p*-nitrophenol. *BioResources* **11**, 10099–111 (2016).
47. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
48. McGibbon et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
49. Nguyen, H., Roe, D. R., Swails, J., & Case, D. A. PYTRAJ: Interactive data analysis for molecular dynamics simulations. (<https://github.com/Amber-MD/pytraj>) (2016).

50. Roe, D. R., & Cheatham, T. E., III Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084-3095 (2013).
51. Gowers, R. J. et al. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In S. Benthall and S. Rostrup, editors, *Proceedings of the 15th Python in Science Conference*, pages 98-105, Austin, TX, 2016.
52. Scherer, M. K. et al. PyEMMA 2: A software package for estimation, validation, and analysis of Markov models, *J. Chem. Theory Comput.* **11**, 5525-5542 (2015).
53. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX, in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, G ael Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008
54. Roe, D. R., & Brooks, B. R. A protocol for preparing explicitly solvated systems for stable molecular dynamics simulations. *J. Chem. Phys.* **153**, 054123 (2020).
55. Darden, T., York, D., & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
56. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Method. Enzymol.* **276**, Macromolecular Crystallography, part A, 307-326 (1997).
57. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658–674 (2007).
58. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. *Acta Cryst.* **D75**, 861–877 (2019).
59. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Cryst.* **D66**, 486–501 (2010).
60. Schmidt, A., Kratky, C., Gruber, K., & Lamzin, V. S. Atomic resolution crystal structures and quantum chemistry meet to reveal subtleties of hydroxynitrile lyase catalysis. *J. Biol. Chem.* **283**, 21827–21836 (2008).
61. Mueller-Dieckmann et al. On the routine use of soft X-rays in macromolecular crystallography. Part IV. Efficient determination of anomalous substructures in biomacromolecules using longer X-ray wavelengths. *Acta Cryst.* **D63**, 366-380 (2007).
62. Schr odinger, LLC., 2022; <https://pymolwiki.org>

Acknowledgements

Funding for this research was provided by US National Science Foundation award CBET-2039039, National Institutes of Health/National Institute of General Medical Sciences (grant No. GM119483); NIH (grant No. NIGMS R35-GM118047); X-ray diffraction data were collected at the Northeastern Collaborative Access Team beamlines, which are funded by the U.S. National Institutes of Health (NIGMS P30 GM124165). The Pilatus 6M detector on the 24-ID-C beamline is funded by a NIH-ORIP HEI grant (S10 RR029205). This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. G.C. and S.O. thank the Generalitat de Catalunya for the consolidated group TCBioSys (SGR 2021 00487) and grant projects

PID2021-129034NB-I00 and PDC2022-133950-I00 funded by Spanish MICIN. 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, ERC-2022-POC-101112805, and ERC-2022-CoG-101088032), and the Human Frontier Science Program (HFSP) for project grant RGP0054/2020. G. C. was supported by a research grant from ERC-StG (ERC-2015-StG-679001) and HFSP RGP0054/2020. We thank Drenen Magee for help with the analysis of x-ray crystallography data.

Chapter 6:

Results and Discussion

This chapter will expose and discuss the three main publications included in this thesis. It will start by explaining the SPM webserver included in Chapter 3. Then, as shown in Chapter 4, a brief discussion of recent protein and enzyme design DL methods followed by a discussion of how to speed up the exploration of the TrpB enzyme conformational heterogeneity with a template-based AF2 approach. Finally, how to convert an HNL enzyme to an EST enzyme with only eight predicted mutations using the SPM tool and the rational insights that prove the remarkable improvement, including the epistatic effects learned with the intermediate variants are presented.

6.1 SPM Webserver for Computational Enzyme Design

Enzymes can be designed using a rational approach or by applying DE, always involving the selection of specific residues for mutagenesis and employing screening protocols to evaluate the enhancement of some targeted traits. Rational approaches often restrict alteration in the active site or tunnels, but DE studies reveal that mutations far away from those important regions can remarkably impact the catalytic activity.¹⁰⁸ Still, the challenge remains in rationally predicting which distal mutations can affect and regulate enzyme activity.

The already-introduced SPM tool (see Section 1.2.4) has been explored in different scenarios providing evidence for the potential of this tool for rationalizing DE mutations, deciphering allosterically important residues, and more recently for design.^{19,143} A webserver for the computation of SPM (SPMweb) has been released to open this tool to the scientific community.

Three files are needed to construct the final 3D graph using the SPMweb: the coordinates of the protein atoms in PDB format, and the distance and correlation matrices. These two matrices are obtained using the information from MD simulations and can be computed with different MD analysis software (*e.g.*, cpptraj or pytraj),^{115,116} usually considering $C\alpha$ or $C\beta$ positions.

For SPM construction, two thresholds must be defined. The distance threshold, which is by default set at 6\AA , is based on identifying the pair of residues whose mean distance between atoms ($C\alpha$ or $C\beta$) along the MD simulations is below the defined threshold. This information is obtained from the distance matrix. As a result, a small distance threshold value will only take into account residues that are close by, producing a very localized and constrained SPM graph.

The screenshot displays the 'Submit your Job' interface of the SPM Webserver. It features three upload boxes: 'Upload Distance Matrix' (Selected File: dist_0B2B.npy), 'Upload Correlation Matrix' (Selected File: corre_0B2B.npy), and 'Upload PDB File' (Selected File: 0B2B_MD.pdb). Below these are two sliders: 'SPM SIGNIFICANCE THRESHOLD' (Current: 0.3 - 1.0) and 'DISTANCE THRESHOLD' (Current: 6). A central area contains a 'Enter your token' input field and two buttons: 'Submit Job' and 'Download Pymol Script'. At the bottom, a 'Protein Visualization (from .pdb file)' section shows a 3D ribbon model of a protein structure with SPM highlighted as labeled spheres. Two buttons at the bottom of the visualization area are 'Toggle Fullscreen' and 'Toggle Labels (It can take several seconds to process)'.

Figure 6.1: **Main interface of the SPM Webserver.** Users can upload the required files, including distance matrix, correlation matrix, and PDB file from MD simulations, into the corresponding boxes. Users have the option to adjust key parameters such as the SPM significance threshold and distance threshold, with default values set at 0.3 and 6 Å respectively. A green box in the center of the interface allows the downloading of a PyMOL script, which can be used to visualize the SPM superposed on the protein structure within PyMOL software. Below, a 3D representation displays the protein with the SPM overlay, where SPM is highlighted with labeled spheres. Access the webserver directly at <https://spmosuna.com>.

The default value for the visualization/significance threshold is set to 0.3, being the significance value normalized to 1.0; bigger values will limit the number of connections to be visualized in the final SPM. Those threshold values were established based on experience in identifying distal mutations

6.2 AlphaFold2 and Deep Learning for Estimating Conformational Heterogeneity and Designing Proteins: The Case of Tryptophan Synthase

from DE variations.¹⁹ Nonetheless, having the chance to change these parameters allows the user to explore which default settings would work best in their particular case.

In the developed webserver the SPM is displayed in the provided enzyme 3D structure, where important residues appear as gray spheres labeled according to the residue number, and edges connecting pairs of residues are highlighted in black. The SPM is also displayed in 2D. The nodes' associated sphere sizes and the nodes' connecting edge widths are primarily qualitative. Moreover, a PyMol script is generated that can be executed in the software to represent the SPM in the loaded structure, thus allowing the user to change and customize the visualization settings (Fig. 6.1).

To show the functionalities of the webserver, the SPM has been computed to identify the key conformationally relevant positions and how they are connected in the case of 0B2-*Pf*TrpB, considering both monomeric and dimeric TrpB units. Therefore, information about how the active site is connected to distal sites and intersubunit communication between monomers can be obtained. The default parameters are first used, but also, distinctive threshold values have been tested to see how the SPM differs.

SPMweb is now freely accessible to academic users. Although it was initially developed for enzyme design, the potential applications are broad and can be further expanded by the scientific community ranging from enzyme design to cryptic pocket identification for drug discovery.

6.2 AlphaFold2 and Deep Learning for Estimating Conformational Heterogeneity and Designing Proteins: The Case of Tryptophan Synthase

As shown in Chapter 4 Review *AlphaFold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design*, many new methods have been revolutionizing the field of protein and enzyme design since 2020 with the appearance of AF2¹⁵ and RoseTTAFold.¹⁶ Many of these methods come from the Baker lab group, with ProteinMPNN¹⁴⁴ or, more recently, the diffusion-based model RoseTTAFold diffusion (RFDiffusion).¹⁴⁵ Other groups have made significant contributions, such as the team behind the ESM models at Facebook AI Research, which developed ESMFold,¹⁴⁶ or the team from the Profluent company with the ProGen2¹⁴⁷ model. Our group also worked on developing a pipeline for protein conformational exploration using AF2 and short-MDs. In this regard, we took advantage of the

6.2 AlphaFold2 and Deep Learning for Estimating Conformational Heterogeneity and Designing Proteins: The Case of Tryptophan Synthase

Trps enzyme, which is known to have catalytically relevant conformational dynamics.¹⁴⁸

TrpS is an enzyme with a well-studied allosteric communication network. This communication involves a conformational rearrangement of different regions of TrpA (e.g., loop 6) and TrpB (e.g., COMM domain). As described in a previous publication of the group,¹⁴⁹ the WT enzyme PfTrpB in the absence of its partner has restricted conformational heterogeneity, whereas the laboratory-evolved 0B2-PfTrpB can explore open (O), partially closed (PC), and closed (C) conformations of the COMM domain, resulting in a stand-alone variant. On the other hand, ancestrally reconstructed Anc3 and LBCA displayed opposite behavior between them, being the last found to display stand-alone activity (*i.e.*, TrpA is not needed for efficient catalysis).⁴⁵ In this line, SPM6 TrpB enzyme, based on Anc3 scaffold, was rationally designed.⁴⁶ In these previous studies, the conformational heterogeneity of all the above-mentioned enzymes was assessed through FEL reconstruction from the computationally demanding metadynamics simulations.¹⁴⁹

As mentioned in Section 1.2.5, AF2 is a DNN that performs extraordinarily well at predicting the lowest energy structure of a protein. Escaping from this static picture to understand enzyme function is not straightforward. Nonetheless, and as suggested in some studies, multiple conformations of the same protein can be predicted by fine-tuning AF2.^{138,139} In this study, we finally developed a template-based AF2 approach coupled to short nanosecond MD simulations to estimate the FEL of different TrpBs quickly.

The template-based AF2 pipeline consists of modifications to two important parts of the AF2 prediction procedure: the MSA depth and the templates used. The MSA depth is defined as the number of sequence co-evolutionary information features used for the prediction. As described in Section 1.2.5, the *max_msa_clusters* default value is 512, and the *max_extra_msa* changes between 1024 and 5120 based on the model used. In this approach, we defined the MSA depth to the *max_msa_clusters* and *max_extra_msa* used as features, being *max_msa_clusters* value half compared to *max_extra_msa* (*i.e.*, ranging from 1024/512 up to 32/12 *max_extra_msa*/*max_msa_clusters*), with the exception when *max_extra_msa* is set to 5120. The templates used in each calculation are set to just one, which can come from an X-ray or an MD frame structure. Another important modification is to set *num_recycle* and *num_ensemble* to 1. In this regard, just two different MSA samples are used. The model usage in this approach is reduced to only the *model_ptm_2*, as it is trained using less *max_extra_msa* (*i.e.*, 1024), and we got a better confidence score result, thus adapting better when lower MSA depth is used. To validate this template-based AF2 approach, we hypothesized that more information regarding the enzyme’s ability to adopt O, PC, and C conformations could be derived by fine-tuning the number of co-evolutionary and

6.2 AlphaFold2 and Deep Learning for Estimating Conformational Heterogeneity and Designing Proteins: The Case of Tryptophan Synthase

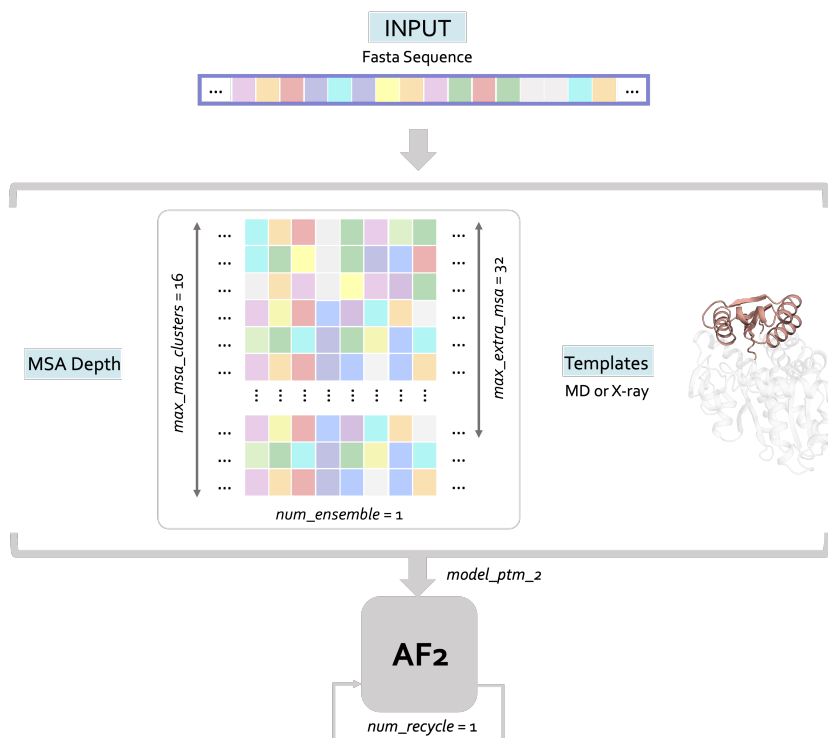


Figure 6.2: **Representation of the template-based AF2 approach diagram.** Key parameters that differ from default settings in the AF2 algorithm are highlighted. $Max_msa_clusters = 16$ and $max_extra_msa = 32$ are set to the minimum values used from the range of MSA Depth. Other modifications include setting $num_recycle$ and $num_ensemble$ to 1, with $model_ptm_2$ utilized as the model.

template features used. In this regard, assuming the AF2 has learned some FEL,¹⁴² lowering MSA depth (*i.e.*, lowering co-evolutionary information) will flatten the AF2's learned FEL and facilitate the prediction of other less probable structures. By adding templates, we incorporate 3D information that will bias the predicted structure to resemble the template.

First, to know how the templates bias the resulting prediction, the structures are predicted with different MSA depths and without templates. The results show that the most probable state for *Pf*TrpB and 0B2-*Pf*TrpB enzymes is the PC, with some O states at lower MSA depth. In the case of the LBCA and SPM6 enzymes, the most probable state of the COMM domain is C. However, the ancestral LBCA populates a wider range of states between PC and C. Instead, the SPM6 enzyme is less flexible, and the

6.2 AlphaFold2 and Deep Learning for Estimating Conformational Heterogeneity and Designing Proteins: The Case of Tryptophan Synthase

increase of co-evolutionary information forces AF2 to predict the closest COMM conformations among the systems (Fig. 6.3).

Structure prediction using the template AF2 approach with X-ray structures as templates was done with those available X-ray structures with sequence identities higher than 70% for all systems. When using C and PC X-ray templates, regarding 0B2-*Pf*TrpB and *Pf*TrpB, PC conformations of the COMM are mostly predicted. Although the differences between both systems are small, AF2 models suggest a slightly higher number of C conformations for 0B2-*Pf*TrpB. For the LBCA and SPM6 TrpB pair, C conformations of the COMM domain are mostly predicted, irrespective of the template structure and MSA depth used. However, when the MSA depth is low in all systems, the 3D information pushes the resulting structure to all possible COMM states (Fig. 6.3).

TrpB has multiple X-ray structures available, with different conformations of the COMM domain. Unfortunately, this is not the case for most systems. Therefore, instead of using X-ray structures as the input template for AF2 prediction, we use structures from MD simulations, which display C, PC, or O conformations of the COMM domain.

When MD-extracted O and PC conformations are used as templates, the predicted structures for 0B2-*Pf*TrpB and *Pf*TrpB present PC conformations, although a slightly higher ability to adopt C conformations can be predicted for 0B2-*Pf*TrpB (74% of the predicted structures adopt PC-C conformation, whereas 67% in the case of *Pf*TrpB). When using O conformations as templates, PC and C structures are predicted for LBCA (at high MSA depths), whereas for SPM6, more C conformations are obtained. This is in line with the higher conformational flexibility of LBCA TrpB. Herein, when the MSA depth is low in all systems, the MD structure also pushes the resulting structure to all possible COMM states (Fig. 6.3).

To further assess the potential application of this template-based AF2 approach for rapid estimation of the conformational heterogeneity, two replicas of short (*i.e.*, 10 ns) MD simulations were run starting from all AF2 structures obtained in X-ray template analysis. Although all systems explore productively the C conformation of the COMM, the FEL reconstruction of 0B2-*Pf*TrpB suggests an additional minimum at O conformations. It is worth mentioning that the O-to-C value (x-axis) at the C minima is ca. 9 for *Pf*TrpB, whereas ca. 10.5 for 0B2-*Pf*TrpB, suggesting a higher ability of the latter to adopt the catalytically productive C conformations.

6.2 AlphaFold2 and Deep Learning for Estimating Conformational Heterogeneity and Designing Proteins: The Case of Tryptophan Synthase

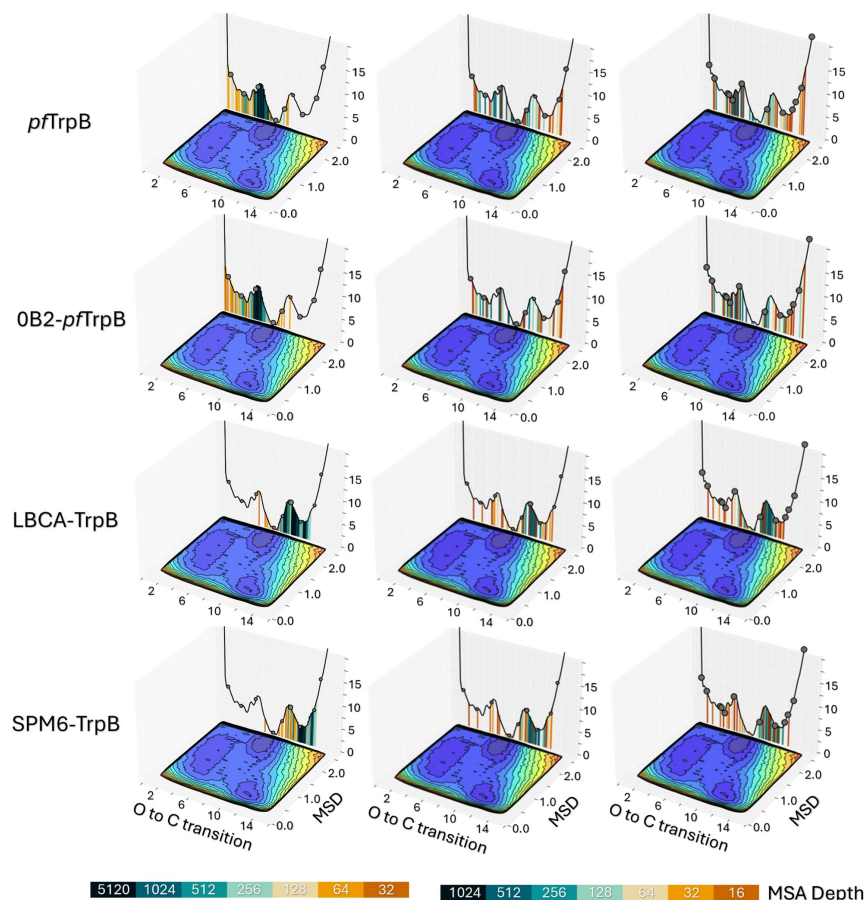


Figure 6.3: Representation of the previously reconstructed FEL of the **OB2-*pf*TrpB** variant.¹⁴⁹ The x-axis represents the open-to-closed (O-to-C) transition of the COMM domain with a scale from 1–5 for open (O), 6–10 for partially closed (PC), and 11–15 for closed (C) states. The y-axis measures the mean square deviation (MSD) from the pathway of O-to-C structural transitions. The color spectrum on the plots indicates the stability of conformations: blue represents the most stable regions, transitioning to red for higher energy states. The three columns depict different prediction methodologies: the first column shows AlphaFold2 (AF2) predictions without templates at varying multiple sequence alignment (MSA) depths, the second column utilizes X-ray structures in a template-based AF2 approach, and the third column uses MD simulation structures as templates in the AF2 predictions. The depth of the MSA is color-coded at the bottom of the figure, ranging from orange to dark blue, indicating increasing depth.

6.3 Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

Altogether, AF2 tends to predict structures close to the global minimum by increasing the co-evolutionary information from the MSA. By altering the MSA depth and using a template as an input (either X-ray structures or conformations taken from MD simulations), the conformational heterogeneity of TrpB can be rapidly estimated, thus visiting structures that escape from the global minimum learned by AF2. Some conformational changes induced by a reduced set of mutations (6 positions included by DE in 0B2-*Pf*TrpB) can be further captured utilizing short MD simulations (*i.e.*, 10 ns each replica). This template-based AF2 approach can therefore be potentially applied for assessing the conformational landscape of new enzyme variants at a rather reduced computational cost.

It is worth mentioning that the analysis of this template-based AF2 has been recently extended. Shortly, the 10 ns MD simulations were elongated up to 50 ns with the already used ff14SB and TIP3P water model, and additionally the ff19SB/OPC combination. To assess the length of the MD simulations and the force field/water model combination, the template-based AF2 approach was further coupled with the SPM analysis. Interestingly, the new SPM coming from the 50 ns MD simulations using the ff19SB/OPC combination, can capture two additional mutations (I68V and T292S) if compared to the results using the computationally demanding metadynamics simulations.⁸⁴

6.3 Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

Reaching nature-like enzyme efficiencies in enzyme design variants is still an unresolved problem, as engineered variants often perform much slower than natural enzymes (*i.e.*, being inefficient). The sophisticated catalytic mechanisms of enzymes involve the simultaneous optimization of substrate binding, TS stabilization, and product release. Enzymes are continuously moving, which is beneficial for precisely positioning the substrate and catalytic groups. Residues far away from the active site can affect catalysis. For all these reasons, it is difficult to predict which mutations will give rise to a shift in the conformational landscape favoring the catalytically active conformations. In this regard, by using the information from MD simulations, the SPM tool can predict the conformationally relevant positions that contribute to catalysis.

*Hb*HNL catalyzes cyanohydrin cleavage, while the homologous esterase SABP2 catalyzes ester hydrolysis. Both enzymes share 45% of sequence

6.3 Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

identity and are structurally very similar. For instance, they have the same serine-histidine-aspartate catalytic triad. *HbHNL* also catalyzes promiscuous hydrolysis of pNPAc, but much slower than SABP2 (*i.e.*, $k_{\text{cat}}/K_{\text{M}}$ of 83 and 61,000 $\text{M}^{-1} \text{min}^{-1}$, respectively). To improve the esterase activity of *HbHNL*, four positions in the active site were reverted towards the corresponding residue found in SABP2, leading to a (still) inefficient esterase named HNL3V (*i.e.*, $k_{\text{cat}}/K_{\text{M}}$ of 490 $\text{M}^{-1} \text{min}^{-1}$).

X-ray structures were overlaid to best fit the $\text{C}\alpha$ positions to compare the active site residues. Although a similar placement of the catalytic atoms was found, one residue from the oxyanion hole (OX1 N atom, the oxyanion residue that is not in the *nucleophilic elbow*) differs significantly between systems. The internal distance between the $\text{C}\alpha$ of Serine and OX1 N is significantly longer in *HbHNL* structures (1Å difference). Altogether, this indicates that *HbHNL* and SABP2 cannot form the same interactions, and thus display the same activity.

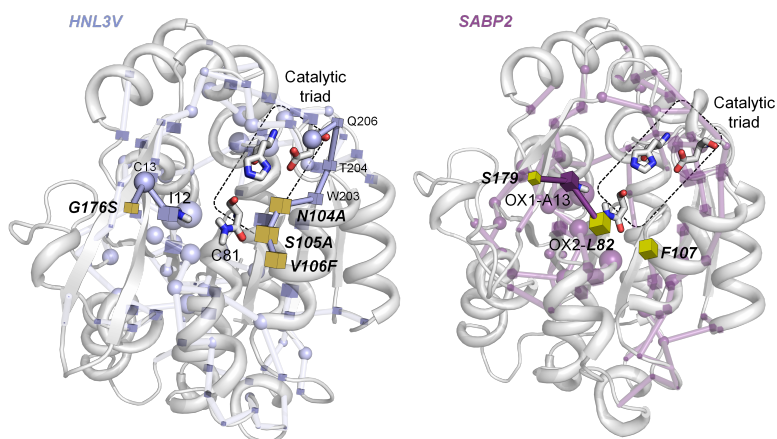


Figure 6.4: **Shortest Path Maps (SPM) of HNL3V (left) and SABP2 (right)**. Highlighted five key substitutions with yellow cubes, designed to enhance the catalytic dynamics of HNL3V to mimic those of SABP2. Significant modifications include Gly176Ser and Cys81Leu, which establish a connection to OX1 (Ala13 in SABP2, Ile12 in HNL3V) in SABP2 SPM. Additional significant substitutions are at residues 104–106, connected to the catalytic Asp207 in HNL3V and disconnected in SABP2 SPM. These substitutions collectively result in the HNL8V variant. Spheres denote conserved residues across both proteins and cubes denote residues that differ. Key catalytic residues and the amides forming the oxyanion hole are depicted with stick representations.

Taking HNL3V as a basis and SABP2 as a reference, MD simulations of both systems were performed, and the SPM was further constructed. Interestingly, some differences in the catalytic residues were revealed, thus

6.3 Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

resulting in two regions liable for mutagenesis. Five mutations were predicted with the SPM tool (*i.e.*, C81L-N104A-S105A-V106F-G176S set of five mutations) based on the network connectivity to the oxyanion hole and catalytic aspartate loop. These five positions were predicted to fix the oxyanion hole orientation and the correlated motions of the catalytic aspartate. The application of these five mutations in HNL3V yields a new variant named HNL8V, composed of the stabilization mutation H103V (*i.e.*, V from HNL8V coming from mutation H103V), the three obvious mutations (*i.e.*, Thr11Gly-Glu79His-Lys236Gly mutations), and the five mutations coming from the SPM tool. Moreover, HNL7V is also designed as an additional variant similar to HNL8V, excluding the Ser105Ala substitution, as it may be less important because of the similarity between residues (Fig. 6.4). Both variants proved to enhance the esterase activity. HNL7V and HNL8V showed a $k_{\text{cat}}/K_{\text{M}}$ of 25,000 and 23,000 $\text{M}^{-1} \text{min}^{-1}$, respectively. This corresponds to an improvement of 290-fold compared to *HbHNL*. However, those variants did not surpass the SABP2 catalytic efficiency of 61,000 $\text{M}^{-1} \text{min}^{-1}$. The incredible activity improvement of the single mutated HNL3V with N104A and the doubly mutated HNL3V with N104A-G176S (*i.e.*, $k_{\text{cat}}/K_{\text{M}}$ of 9,100 and 52,000 $\text{M}^{-1} \text{min}^{-1}$, respectively) highlighted the importance of residue position 104 (*i.e.*, position 105 in SABP2). In this regard, identification of the most conserved amino acid at this position among homologous esterases being 105T made it interesting to test the variant HNL7T (*i.e.*, HNL7 with A104T mutation). Surprisingly, the catalytic efficiency increased to 120,000 $\text{M}^{-1} \text{min}^{-1}$, which is almost twice the SABP2 enzyme's catalytic efficiency.

6.3 Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

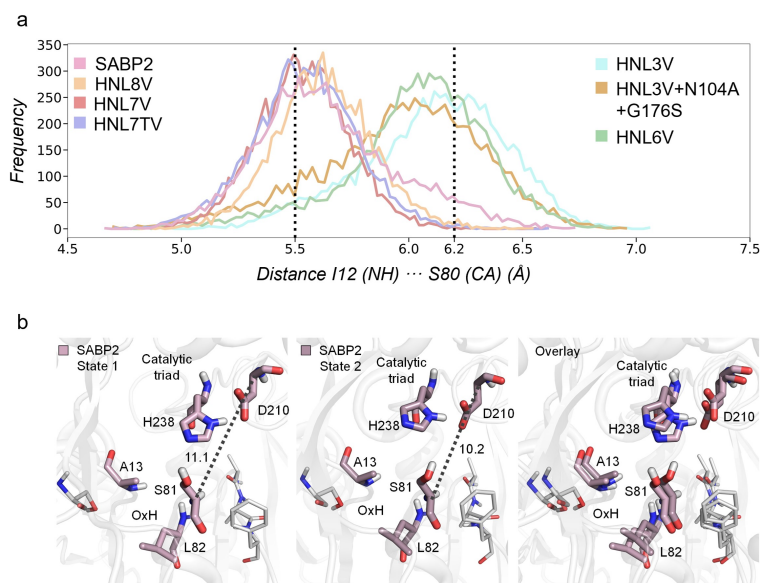


Figure 6.5: **a) Histogram of the distances (in Å)** between the amide backbone of OX1 N (Ile12 or Ala13) and C α of the catalytic serine (Ser80 or 81) for the following proteins: SABP2 (pink), HNL3V (blue), HNL3V N104A G176S (orange), HNL6V (green), HNL7V (dark pink), HNL7TV (purple), HbHNL8V (brown). **b) Representation of two SABP2's catalytic triad conformational states** presenting closer and longer distances between nucleophilic Ser C α and catalytic Asp C α , and the corresponding overlay.

The improvements in EST activity in *HbHNL* variants can be rationalized through MD simulations, which can be attributed to two important improvements concerning the SABP2 enzyme. The first is the oxyanion hole regeneration, where these active variants have a highly similar distance distribution between OX1 N and the serine C α to SABP2 enzyme, being of about ca. 5.5Å compared to the ca. 6.2Å of the bad variants (Fig. 6.5a). The second improvement corresponds to the distance distribution between the nucleophilic Ser amino acid and the catalytic triad His-Asp pair, where the *HbHNL* highly active variants resemble the SABP2 distribution of closer distances, being represented as SABP2 state 2 in Fig. 6.5b.

6.3 Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

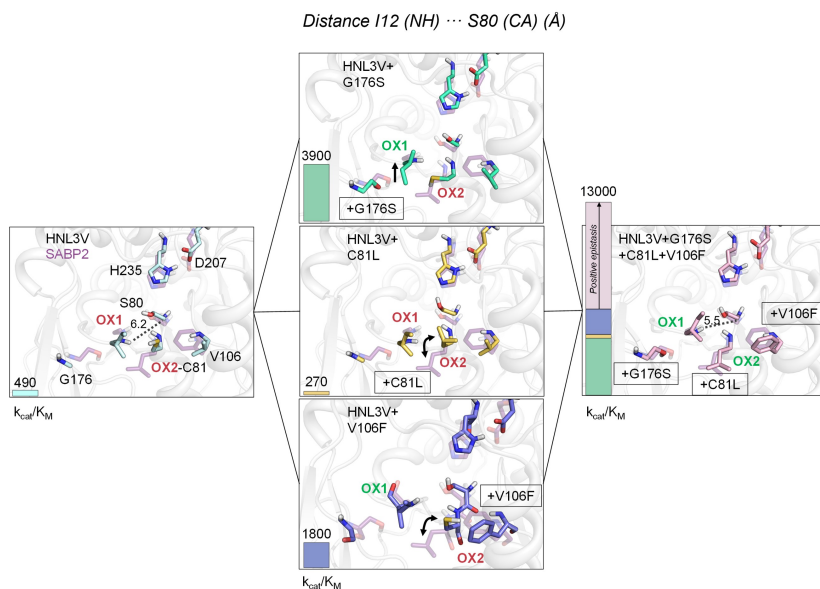


Figure 6.6: Cooperative interactions among mutations enhance catalytic activity. Represented are the conformations of HNL variants: HNL3V (left panel, light blue carbons) as the starting variant; the singly mutated variants HNL3V+G176S (center top panel, green carbons), HNL3V+C81L (center middle panel, gold carbons), and HNL3V+V106F (down middle panel, dark blue carbons); and the triple variant (right panel, pink carbons). All panels include a representative conformation of the active site and oxyanion hole residues of SABP2 (purple carbons) and a bar at the left side indicating the catalytic efficiency (k_{cat}/K_M , $M^{-1} \text{ min}^{-1}$) of each HNL variant. The synergistic effect of the three mutations exceeds their individual contributions, indicating a positive sign of epistasis, highlighted in pink with an arrow. The labels of the oxyanion hole residues (OX1, OX2) are color-coded to indicate correct (green) or incorrect (red) orientation relative to SABP2. Double arrows indicate the different conformations of the C81 sidechain compared to SABP2. Distances between the amide backbone of OX1 N Ile12 and $C\alpha$ of the catalytic serine are defined in HNL3V and the triple variant.

In this line, the regeneration of the oxyanion hole's functionality in the enzyme variant was elucidated through the epistatic effects among three specific mutations identified via SPM predictions: C81L, V106F, and G176S. These mutations collectively enhance the enzyme's EST catalytic efficiency by more than twice what would be expected if they were simply additive. Specifically, the cooperative interaction among these mutations leads to an optimal rearrangement of the oxyanion hole. The G176S mutation adjusts the enzyme's backbone conformation, effectively positioning residue Ile12 optimally for orienting the second oxyanion hole (*i.e.*, the OX12 N

6.3 Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

atom) (Fig. 6.6). The double mutations of C81L and V106F are crucial for correctly orienting the first oxyanion hole residue (*i.e.*, the OX11 N atom) and facilitating substrate access to the active site for catalysis. This demonstrates positive cooperativity among these mutations and underscores their critical roles in enhancing the catalytic function by structurally and dynamically reconfiguring key elements of the enzyme's active site.

Chapter 7:

Conclusions

In this thesis, we have extensively explored the utility of computational tools in advancing the field of enzyme design. The development and application of the SPM webserver tool and a template-based AF2 approach have demonstrated significant improvements in our understanding and capability to design enzymes with desired functionalities and explore the conformational heterogeneity of proteins. The major conclusions drawn from this thesis organized by chapter are as follows:

1. **In Chapter 3**, the SPM tool has proven to be a groundbreaking approach for identifying key dynamic residues essential for enzyme function beyond the active site. This tool uses MD simulations to build an extensive map of connections within the enzyme, showing how distal mutations impact enzyme dynamics and activity. The release of SPMweb extends the accessibility and applicability of this tool, promoting wider use in the scientific community for research and enzyme design. The user-friendly interface allows the user to change the key parameters to not be limited by the default settings.
2. **In Chapter 4**, the advances in DL methods and AF2 for protein design are shown, and the template-based AF2 pipeline is presented, which consists of modifying the MSA depth and template inputs in the AF2 protocol to promote conformational exploration of enzymes. Utilizing four TrpB enzymes with their already computed FEL of states from O to C of the COMM domain, and a huge amount of X-ray structures, we could prove the benefits of tuning the AF2 pipeline with the balance of co-evolutionary and physical information to obtain the widest conformational search of each system. With the introduction of short-MD in the pipeline, we could compare the benefits of this approach to the expensive metadynamics calculation for recovering the FEL. The low cost of this approach helps to get insights that will be highly beneficial for the fast rational design of enzymes or for discovering the complexity of protein dynamics.
3. **In Chapter 5**, the rational design of an HNL enzyme into an efficient EST through the introduction of eight predicted mutations utilizing the SPM tool highlights the power of including the dynamics of enzymes are an important part of the enzyme design process. The rational identification of the properties that were gained by the *HbHNL* variants through oxyanion regeneration and catalytic triad repositioning, combined with the epistatic effects needed for the final improvement, showcases the utility of MD simulations to get insights for further enzyme design campaigns. Thanks to a deeper understanding of these enzymes we can further use this enzyme to design new variants with novel functionalities.

The findings from this thesis underline the complex connection between enzyme structure, its dynamic behavior, and catalytic function. The ability to manipulate enzyme activity through strategic mutations reported by detailed dynamic maps and structurally validated through cheap conformational sampling pipelines offers a powerful method for enzyme design. The potential to apply these computational strategies to the tailored design of industrially relevant enzymes presents an exciting frontier for research and development.

References

- (1) Mai, V. Q.; Meere, M. *Mathematics* **2021**, *9*, 2315.
- (2) Pinney, M. M.; Mokhtari, D. A.; Akiva, E.; Yabukarski, F.; Sanchez, D. M.; Liang, R.; Doukov, T.; Martinez, T. J.; Babbitt, P. C.; Herschlag, D. *Science* **2021**, *371*, eaay2784.
- (3) Bangaru, A.; Sree, K. A.; Kruthiventi, C.; Banala, M.; Shreya, V.; Vineetha, Y.; Shalini, A.; Mishra, B.; Yadavalli, R.; Chandrasekhar, K.; Reddy, C. N. In *Bio-Clean Energy Technologies: Volume 1*, Chowdhary, P., Khanna, N., Pandit, S., Kumar, R., Eds.; Springer Nature Singapore: Singapore, 2022, pp 81–112.
- (4) Pang, C.; Yin, X.; Zhang, G.; Liu, S.; Zhou, J.; Li, J.; Du, G. *Systems Microbiology and Biomanufacturing* **2021**, *1*, 24–32.
- (5) Saravanan, A.; Kumar, P. S.; Vo, D.-V. N.; Jeevanantham, S.; Karishma, S.; Yaashikaa, P. *Journal of Hazardous Materials* **2021**, *419*, 126451.
- (6) Oyewole, O. A.; Idris, A. D.; Bello, A. B.; Yakubu, J. G.; Saidu, M. M. In *Ecological Interplays in Microbial Enzymology*, Maddela, N. R., Abiodun, A. S., Prasad, R., Eds.; Springer Nature Singapore: Singapore, 2022, pp 189–213.
- (7) Gurung, N.; Ray, S.; Bose, S.; Rai, V., et al. *BioMed research international* **2013**, *2013*, DOI: 10.1155/2013/329121.
- (8) Chandra, P.; Enespa; Singh, R.; Arora, P. K. *Microbial cell factories* **2020**, *19*, 1–42.
- (9) Zuk, R. F.; Ginsberg, V. K.; Houts, T.; Rabbie, J.; Merrick, H.; Ullman, E. F.; Fischer, M. M.; Sizto, C. C.; Stiso, S. N.; Litman, D. J. *Clinical Chemistry* **1985**, *31*, 1144–1150.
- (10) May, K. *American Journal of Obstetrics and Gynecology* **1991**, *165*, 2000–2002.
- (11) Aydin, S. *Peptides* **2015**, *72*, 4–15.

REFERENCES

- (12) Blair, H. A. *Drugs* **2023**, *83*, 739–745.
- (13) Bornscheuer, U. T.; Huisman, G.; Kazlauskas, R.; Lutz, S.; Moore, J.; Robins, K. *Nature* **2012**, *485*, 185–194.
- (14) Sarai, N. S.; Fulton, T. J.; O'Meara, R. L.; Johnston, K. E.; Brinkmann-Chen, S.; Maar, R. R.; Tecklenburg, R. E.; Roberts, J. M.; Reddel, J. C. T.; Katsoulis, D. E.; Arnold, F. H. *Science* **2024**, *383*, 438–443.
- (15) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A., et al. *Nature* **2021**, *596*, 583–589.
- (16) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D., et al. *Science* **2021**, *373*, 871–876.
- (17) Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. *Proceedings of the National Academy of Sciences* **2020**, *117*, 1496–1503.
- (18) Arnold, F. H. *Angewandte Chemie (International Ed. in English)* **2018**, *57*, 4143.
- (19) Osuna, S. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2021**, *11*, e1502.
- (20) Bell, E. L.; Finnigan, W.; France, S. P.; Green, A. P.; Hayes, M. A.; Hepworth, L. J.; Lovelock, S. L.; Niikura, H.; Osuna, S.; Romero, E., et al. *Nature Reviews Methods Primers* **2021**, *1*, 1–21.
- (21) Bryliakov, K. P. *Chemical Reviews* **2017**, *117*, 11406–11459.
- (22) Sun, W.; Sun, Q. *Accounts of Chemical Research* **2019**, *52*, 2370–2381.
- (23) Larson, V. A.; Battistella, B.; Ray, K.; Lehnert, N.; Nam, W. *Nature Reviews Chemistry* **2020**, *4*, 404–419.
- (24) Dobson, C. M. *Nature* **2003**, *426*, 884–890.
- (25) Pauling, L. *Chem. Eng. News* **1946**, *24*, 1375–1377.
- (26) Warshel, A. *Journal of Biological Chemistry* **1998**, *273*, 27035–27038.
- (27) Martí, S.; Roca, M.; Andrés, J.; Moliner, V.; Silla, E.; Tuñón, I.; Bertrán, J. *Chemical Society Reviews* **2004**, *33*, 98–107.
- (28) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. *Chemical reviews* **2006**, *106*, 3210–3235.
- (29) Hennefarth, M. R.; Alexandrova, A. N. *Current opinion in structural biology* **2022**, *72*, 1–8.
- (30) Fried, S. D.; Bagchi, S.; Boxer, S. G. *Science* **2014**, *346*, 1510–1514.

-
- (31) Wu, Y.; Boxer, S. G. *Journal of the American Chemical Society* **2016**, *138*, 11890–11895.
- (32) Vaissier Welborn, V.; Head-Gordon, T. *Chemical reviews* **2018**, *119*, 6613–6630.
- (33) Wolfenden, R. *Nature* **1969**, *223*, DOI: 10.1038/223704a0.
- (34) Li, C. M.; Tyler, P. C.; Furneaux, R. H.; Kicska, G.; Xu, Y.; Grubmeyer, C.; Girvin, M. E.; Schramm, V. L. *nature structural biology* **1999**, *6*, 582–587.
- (35) Dagen, M.; Patrick, G. L. In *Antimalarial Agents*; Elsevier: 2020, pp 513–546.
- (36) Schramm, V. L. *Journal of Biological Chemistry* **2007**, *282*, 28297–28300.
- (37) Fischer, E. *Berichte der deutschen chemischen Gesellschaft* **1894**, *27*, 2985–2993.
- (38) O'Brien, P. J.; Herschlag, D. *Chemistry & biology* **1999**, *6*, R91–R105.
- (39) Nam, H.; Lewis, N. E.; Lerman, J. A.; Lee, D.-H.; Chang, R. L.; Kim, D.; Palsson, B. O. *Science* **2012**, *337*, 1101–1104.
- (40) Jensen, R. A. *Annual review of microbiology* **1976**, *30*, 409–425.
- (41) Koshland Jr, D. E. *Proceedings of the National Academy of Sciences* **1958**, *44*, 98–104.
- (42) Ma, B.; Kumar, S.; Tsai, C.-J.; Nussinov, R. *Protein engineering* **1999**, *12*, 713–720.
- (43) Gardner, J. M.; Biler, M.; Risso, V. A.; Sanchez-Ruiz, J. M.; Kamerlin, S. C. *ACS Catalysis* **2020**, *10*, 4863–4870.
- (44) Buller, A. R.; Brinkmann-Chen, S.; Romney, D. K.; Herger, M.; Murciano-Calles, J.; Arnold, F. H. *Proceedings of the National Academy of Sciences* **2015**, *112*, 14599–14604.
- (45) Schupfner, M.; Straub, K.; Busch, F.; Merkl, R.; Sterner, R. *Proceedings of the National Academy of Sciences* **2020**, *117*, 346–354.
- (46) Maria-Solano, M. A.; Kinateder, T.; Iglesias-Fernández, J.; Sterner, R.; Osuna, S. *ACS catalysis* **2021**, *11*, 13733–13743.
- (47) Busch, F.; Rajendran, C.; Heyn, K.; Schlee, S.; Merkl, R.; Sterner, R. *Cell chemical biology* **2016**, *23*, 709–715.
- (48) Crawford, I. P. *Biochimica et Biophysica Acta* **1960**, *45*, 405–407.
- (49) Romney, D. K.; Murciano-Calles, J.; Wehrmüller, J. E.; Arnold, F. H. *Journal of the American Chemical Society* **2017**, *139*, 10769–10776.

REFERENCES

- (50) Watkins-Dulaney, E. J.; Dunham, N. P.; Straathof, S.; Turi, S.; Arnold, F. H.; Buller, A. R. *Angewandte Chemie* **2021**, *133*, 21582–21587.
- (51) Romney, D. K.; Sarai, N. S.; Arnold, F. H. *ACS catalysis* **2019**, *9*, 8726–8730.
- (52) Holmquist, M. *Current Protein and Peptide Science* **2000**, *1*, 209–235.
- (53) Nardini, M.; Dijkstra, B. W. *Current opinion in structural biology* **1999**, *9*, 732–737.
- (54) Ollis, D. L.; Cheah, E.; Cygler, M.; Dijkstra, B.; Frolow, F.; Franken, S. M.; Harel, M.; Remington, S. J.; Silman, I.; Schrag, J., et al. *Protein Engineering, Design and Selection* **1992**, *5*, 197–211.
- (55) Bauer, T. L.; Buchholz, P. C.; Pleiss, J. *The FEBS Journal* **2020**, *287*, 1035–1053.
- (56) Rauwerdink, A.; Kazlauskas, R. J. *ACS catalysis* **2015**, *5*, 6153–6176.
- (57) Xu, A.; Zhou, J.; Blank, L. M.; Jiang, M. *Trends in Microbiology* **2023**, DOI: 10.1016/j.tim.2023.04.002.
- (58) Wajant, H.; Förster, S. *Plant Science* **1996**, *115*, 25–31.
- (59) Du, H.; Klessig, D. F. *Plant Physiology* **1997**, *113*, 1319–1327.
- (60) Du, H.; Klessig, D. F. *Plant Physiology* **1997**, *113*, 1319–1327.
- (61) Lieberei, R.; Selmar, D.; Biehl, B. *Plant Systematics and Evolution* **1985**, *150*, 49–63.
- (62) Poulton, J. E. *Plant physiology* **1990**, *94*, 401–405.
- (63) Hickel, A.; Hasslacher, M.; Griengl, H. *Physiologia Plantarum* **1996**, *98*, 891–898.
- (64) Forouhar, F.; Yang, Y.; Kumar, D.; Chen, Y.; Fridman, E.; Park, S. W.; Chiang, Y.; Acton, T. B.; Montelione, G. T.; Pichersky, E., et al. *Proceedings of the National Academy of Sciences* **2005**, *102*, 1773–1778.
- (65) Soares, J. M.; Weber, K. C.; Qiu, W.; Mahmoud, L. M.; Grosser, J. W.; Dutt, M. *Plant Cell Reports* **2022**, *41*, 2305–2320.
- (66) Devamani, T.; Rauwerdink, A. M.; Lunzer, M.; Jones, B. J.; Mooney, J. L.; Tan, M. A. O.; Zhang, Z.-J.; Xu, J.-H.; Dean, A. M.; Kazlauskas, R. J. *Journal of the American Chemical Society* **2016**, *138*, 1046–1056.
- (67) Zhao, Y.; Chen, N.; Wang, C.; Cao, Z. *ACS Catalysis* **2016**, *6*, 2145–2157.
- (68) Nedrud, D. M.; Lin, H.; Lopez, G.; Padhi, S. K.; Legatt, G. A.; Kazlauskas, R. J. *Chemical science* **2014**, *5*, 4265–4277.

- (69) Jones, B. J.; Evans III, R. L.; Mylrea, N. J.; Chaudhury, D.; Luo, C.; Guan, B.; Pierce, C. T.; Gordon, W. R.; Wilmot, C. M.; Kazlauskas, R. J. *PLoS One* **2020**, *15*, e0235341.
- (70) Bixon, M.; Lifson, S. *Tetrahedron* **1967**, *23*, 769–784.
- (71) Levitt, M. *Nature structural biology* **2001**, *8*, 392–393.
- (72) McCammon, J. A.; Gelin, B. R.; Karplus, M. *nature* **1977**, *267*, 585–590.
- (73) Warshel, A.; Levitt, M. *Journal of molecular biology* **1976**, *103*, 227–249.
- (74) Warshel, A.; Weiss, R. M. *Journal of the American Chemical Society* **1980**, *102*, 6218–6226.
- (75) Levitt, M.; Sharon, R. *Proceedings of the National Academy of Sciences* **1988**, *85*, 7557–7561.
- (76) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *Journal of chemical theory and computation* **2015**, *11*, 3696–3713.
- (77) Tian, C.; Kasavajhala, K.; Belfon, K. A.; Raguette, L.; Huang, H.; Miguez, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q., et al. *Journal of chemical theory and computation* **2019**, *16*, 528–552.
- (78) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *Journal of computational chemistry* **2004**, *25*, 1157–1174.
- (79) Ren, P.; Ponder, J. W. *The Journal of Physical Chemistry B* **2003**, *107*, 5933–5947.
- (80) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A., et al. *The journal of physical chemistry B* **2010**, *114*, 2549–2564.
- (81) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. *Chemical Reviews* **2021**, *121*, 10142–10186.
- (82) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *The Journal of chemical physics* **1983**, *79*, 926–935.
- (83) Izadi, S.; Anandakrishnan, R.; Onufriev, A. V. *The journal of physical chemistry letters* **2014**, *5*, 3863–3871.
- (84) Duran, C.; Casadevall, G.; Osuna, S. *Faraday Discuss.* **2024**, DOI: 10.1039/D3FD00156C.
- (85) Becke, A. D. *The Journal of chemical physics* **1992**, *96*, 2155–2160.
- (86) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

REFERENCES

- (87) Tomasi, J.; Mennucci, B.; Cammi, R. *Chemical reviews* **2005**, *105*, 2999–3094.
- (88) Schutz, C. N.; Warshel, A. *Proteins: Structure, Function, and Bioinformatics* **2001**, *44*, 400–417.
- (89) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *Journal of molecular graphics and modelling* **2006**, *25*, 247–260.
- (90) Singh, U. C.; Kollman, P. A. *Journal of computational chemistry* **1984**, *5*, 129–145.
- (91) Besler, B. H.; Merz Jr, K. M.; Kollman, P. A. *Journal of computational chemistry* **1990**, *11*, 431–439.
- (92) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *The Journal of Physical Chemistry* **1993**, *97*, 10269–10280.
- (93) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *Journal of the American Chemical Society* **1993**, *115*, 9620–9631.
- (94) Frisch, M. J. et al. Gaussian 16 Revision C.01, Gaussian Inc. Wallingford CT, 2016.
- (95) Jessen, L. E.; Hoof, I.; Lund, O.; Nielsen, M. *Nucleic acids research* **2013**, *41*, W286–W291.
- (96) Khersonsky, O.; Lipsh, R.; Avizemer, Z.; Ashani, Y.; Goldsmith, M.; Leader, H.; Dym, O.; Rogotner, S.; Trudeau, D. L.; Prilusky, J., et al. *Molecular cell* **2018**, *72*, 178–186.
- (97) Fleishman, S. J.; Leaver-Fay, A.; Corn, J. E.; Strauch, E.-M.; Khare, S. D.; Koga, N.; Ashworth, J.; Murphy, P.; Richter, F.; Lemmon, G., et al. *PloS one* **2011**, *6*, e20161.
- (98) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. *Angewandte Chemie International Edition* **2013**, *52*, 5700–5725.
- (99) Tantillo, D. J.; Jiangang, C.; Houk, K. N. *Current opinion in chemical biology* **1998**, *2*, 743–750.
- (100) Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Röthlisberger, D.; Baker, D. *Protein Science* **2006**, *15*, 2785–2794.
- (101) Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D. *Journal of molecular biology* **2003**, *332*, 449–460.
- (102) Kiss, G.; Röthlisberger, D.; Baker, D.; Houk, K. N. *Protein science* **2010**, *19*, 1760–1773.
- (103) Wijma, H. J.; Floor, R. J.; Bjelic, S.; Marrink, S. J.; Baker, D.; Janssen, D. B. *Angewandte Chemie International Edition* **2015**, *54*, 3726–3730.

- (104) Prokop, Z.; Gora, A.; Brezovsky, J.; Chaloupkova, R.; Stepankova, V.; Damborsky, J. *Protein engineering handbook* **2012**, *3*, 421–464.
- (105) 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 Computational Biology* **2012**, *8*, e1002708.
- (106) Magdziarz, T.; Mitusińska, K.; Goldowska, S.; Pluciennik, A.; Stolarczyk, M.; Lugowska, M.; Góra, A. *Bioinformatics* **2017**, *33*, 2045–2046.
- (107) Corbella, M.; Pinto, G. P.; Kamerlin, S. C. *Nature Reviews Chemistry* **2023**, *7*, 536–547.
- (108) Currin, A.; Swainston, N.; Day, P. J.; Kell, D. B. *Chemical Society Reviews* **2015**, *44*, 1172–1239.
- (109) 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. *Nature chemical biology* **2014**, *10*, 431–436.
- (110) Ghislieri, D.; Green, A. P.; Pontini, M.; Willies, S. C.; Rowles, I.; Frank, A.; Grogan, G.; Turner, N. J. *Journal of the American Chemical Society* **2013**, *135*, 10863–10869.
- (111) Campbell, E.; Kaltenbach, M.; Correy, G. J.; Carr, P. D.; Porebski, B. T.; Livingstone, E. K.; Afriat-Jurnou, L.; Buckle, A. M.; Weik, M.; Hollfelder, F., et al. *Nature chemical biology* **2016**, *12*, 944–950.
- (112) Bell, E. L.; Smithson, R.; Kilbride, S.; Foster, J.; Hardy, F. J.; Ramachandran, S.; Tedstone, A. A.; Haigh, S. J.; Garforth, A. A.; Day, P. J., et al. *Nature Catalysis* **2022**, *5*, 673–681.
- (113) Shi, L.; Liu, P.; Tan, Z.; Zhao, W.; Gao, J.; Gu, Q.; Ma, H.; Liu, H.; Zhu, L. *Angewandte Chemie International Edition* **2023**, *62*, e202218390.
- (114) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. *Proceedings of the National Academy of Sciences* **2009**, *106*, 6620–6625.
- (115) Roe, D. R.; Cheatham III, T. E. *Journal of chemical theory and computation* **2013**, *9*, 3084–3095.
- (116) Nguyen, H.; Roe, D. R.; Swails, J.; Case, D. A. pytraj, <https://github.com/Amber-MD/pytraj>, (in preparation), 2015.
- (117) Csardi, G.; Nepusz, T. *InterJournal* **2006**, *Complex Systems*, 1695.
- (118) Romero-Rivera, A.; Garcia-Borras, M.; Osuna, S. *ACS catalysis* **2017**, *7*, 8524–8532.

REFERENCES

- (119) Levinthal, C. In *Mossbauer Spectroscopy in Biological Systems, Proceedings of a Meeting held at Allerton House*, ed. by Debrunner, P.; Tsibris, J.; Münck, E., University of Illinois Press: Urbana, 1969, p 22.
- (120) Karplus, M. *Folding and design* **1997**, *2*, S69–S75.
- (121) Wedemeyer, W. J.; Welker, E.; Scheraga, H. A. *Biochemistry* **2002**, *41*, 14637–14644.
- (122) Baldwin, R. L. *Proceedings of the National Academy of Sciences* **2017**, *114*, 8442–8443.
- (123) Bryngelson, J. D.; Wolynes, P. G. *The Journal of Physical Chemistry* **1989**, *93*, 6902–6915.
- (124) Leopold, P. E.; Montal, M.; Onuchic, J. N. *Proceedings of the National Academy of Sciences* **1992**, *89*, 8721–8725.
- (125) Dill, K. A.; Chan, H. S. *Nature structural biology* **1997**, *4*, 10–19.
- (126) Anfinsen, C. B.; Haber, E.; Sela, M.; White Jr, F. *Proceedings of the National Academy of Sciences* **1961**, *47*, 1309–1314.
- (127) Anfinsen, C. B. *Science* **1973**, *181*, 223–230.
- (128) Hirata, F.; Sugita, M.; Yoshida, M.; Akasaka, K. *The Journal of Chemical Physics* **2018**, *148*, DOI: 10.1063/1.5013104.
- (129) Eddy, S. R. *PLoS computational biology* **2011**, *7*, e1002195.
- (130) Johnson, L. S.; Eddy, S. R.; Portugaly, E. *BMC bioinformatics* **2010**, *11*, 1–8.
- (131) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. *Nature methods* **2012**, *9*, 173–175.
- (132) Richardson, L.; Allen, B.; Baldi, G.; Beracochea, M.; Bileschi, M. L.; Burdett, T.; Burgin, J.; Caballero-Pérez, J.; Cochrane, G.; Colwell, L. J., et al. *Nucleic Acids Research* **2023**, *51*, D753–D759.
- (133) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; Consortium, U. *Bioinformatics* **2015**, *31*, 926–932.
- (134) Mirdita, M.; Von Den Driesch, L.; Galiez, C.; Martin, M. J.; Söding, J.; Steinegger, M. *Nucleic acids research* **2017**, *45*, D170–D176.
- (135) Steinegger, M.; Söding, J. *Nature communications* **2018**, *9*, 2542.
- (136) Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S. J.; Söding, J. *BMC bioinformatics* **2019**, *20*, 1–15.
- (137) Zhang, Y.; Skolnick, J. *Proteins: Structure, Function, and Bioinformatics* **2004**, *57*, 702–710.
- (138) Del Alamo, D.; Sala, D.; Mchaourab, H. S.; Meiler, J. *Elife* **2022**, *11*, e75751.

-
- (139) Stein, R. A.; Mchaourab, H. S. *PLOS Computational Biology* **2022**, *18*, e1010483.
- (140) Wayment-Steele, H. K.; Ojoawo, A.; Otten, R.; Apitz, J. M.; Pitsawong, W.; Hömberger, M.; Ovchinnikov, S.; Colwell, L.; Kern, D. *Nature* **2024**, *625*, 832–839.
- (141) Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. *Bioinformatics* **2013**, *29*, 2722–2728.
- (142) Roney, J. P.; Ovchinnikov, S. *Physical Review Letters* **2022**, *129*, 238101.
- (143) Calvó-Tusell, C.; Maria-Solano, M. A.; Osuna, S.; Feixas, F. *Journal of the American Chemical Society* **2022**, *144*, 7146–7159.
- (144) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I.; Courbet, A.; de Haas, R. J.; Bethel, N., et al. *Science* **2022**, *378*, 49–56.
- (145) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F., et al. *Nature* **2023**, *620*, 1089–1100.
- (146) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y., et al. *Science* **2023**, *379*, 1123–1130.
- (147) Nijkamp, E.; Ruffolo, J. A.; Weinstein, E. N.; Naik, N.; Madani, A. *Cell systems* **2023**, *14*, 968–978.
- (148) Casadevall, G.; Duran, C.; Estévez-Gay, M.; Osuna, S. *Protein Science* **2022**, *31*, e4426.
- (149) Maria-Solano, M. A.; Iglesias-Fernández, J.; Osuna, S. *Journal of the American Chemical Society* **2019**, *141*, 13049–13056.



Appendix

Supporting Information of Chapter 4

Estimating conformational heterogeneity of tryptophan synthase with a template-based AlphaFold2 approach

Guillem Casadevall,¹ Cristina Duran,¹ Miquel Estévez-Gay,¹ Sílvia Osuna^{1,2*}

[1] CompBioLab group, Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, Carrer Maria Aurèlia Capmany 69, 17003 Girona, Spain

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

Supporting Information

Molecular dynamics simulations. System preparation. The starting structures for the four enzymes (*pf*TrpB, OB2-*pf*TrpB, LBCA-TrpB, and SPM6-TrpB) were generated with the predictions of the X-ray template-based AF2 approach. The AF2 models simulated were the ones without backbone steric clashes and with the predicted LDDT-C α score (pLDDT) higher than 89. Finally, from the 63 predicted AF2 structures for each system, we simulated a total of 60, 59, 62, and 59 structures for *pf*TrpB, OB2-*pf*TrpB, LBCA-TrpB, and SPM6-TrpB systems, respectively. To generate the TrpB homodimer enzyme of *pf*TrpB and OB2-*pf*TrpB AF2 predictions, we used the TrpB crystal structure from *Pyrococcus furiosus* (PDB: 5DW0) as a template to superpose the AF2 structure to each TrpB monomer. Whereas, for the LBCA-TrpB and SPM6-TrpB AF2 predictions, we used the TrpB homodimer of the LBCA TrpS complex (PDB: 5EY5) to generate the TrpB homodimer. The Q₂ intermediate was placed in the TrpB subunits through superposition to the Aex2 intermediate of the engineered TrpB X-ray structure with PDB accession code 6AM8. Also, to avoid clashes with the Q₂ intermediate, the chi2, chi3, and chi4 torsion angles of the catalytic lysine (i.e., Lys84) were switched to the ones in the 6AM8 X-ray structure. Likewise, the chi4 torsion angle of the Arg375 sidechain was switched to maintain the Arg-Arg interaction at the dimer interface. The structures were prepared using the Python packages MDTraj, pytraj, MDAnalysis, PyEMMA, and networkx.¹⁻⁴

The water molecules added to each homodimer were selected from the DBSCAN clusterization^{5,6} algorithm implemented in the scikit-learn Python library,⁷ of all X-ray TrpB monomers with a sequence identity greater than 70% for the four systems *i.e.*, order by PDB accession number and chain: 5IXJ_A, 5IXJ_B, 5IXJ_C, 5IXJ_D, 5DW3_A, 5DW3_B, 5DW3_C, 5DW3_D, 5E0K_B, 5E0K_D, 5E0K_F, 5E0K_H, 5E0K_J, 5E0K_L, 5DW0_A, 5DW0_B, 5DW0_C, 5DW0_D, 1V8Z_A, 1V8Z_B, 1V8Z_C, 1V8Z_D, 5T6M_A, 5T6M_B, 5T6M_C, 5T6M_D, 1WDW_B, 1WDW_D, 1WDW_F, 1WDW_H, 1WDW_J, 1WDW_L, 5DVZ_A, 5DVZ_B, 5DVZ_C, 5DVZ_D, 6AMH_A, 6AMH_B, 6AMH_C, 6AMH_D, 6AMI_A, 6AMI_B, 6AMI_C, 6AMI_D, 6AMC_A, 6AMC_B, 6AMC_C, 6AMC_D, 5VM5_A, 5VM5_B, 5VM5_C, 5VM5_D, 6AM8_A, 6AM8_B, 6AM8_C, 6AM8_D, 6AM9_A, 6AM9_B, 6AM9_C, 6AM9_D, 6AM7_A, 6AM7_B, 6AM7_C, 6AM7_D, 6CUV_A, 6CUV_B, 6CUV_C, 6CUV_D, 6CUT_A, 6CUT_B, 6CUT_C, 6CUT_D, 6CUZ_A, 6CUZ_B, 6CUZ_C, 6CUZ_D, 5EY5_B, 5EY5_D. Additionally, three conserved sodium ions in the X-ray structures were added to all structures located at the dimer interface and in each monomer close to the active site.

The MD parameters for Q₂ intermediate were generated with the antechamber and parmchk2 modules of AMBER20⁸ using the 2nd generation of the general amber force-field (GAFF2).^{9,10} The Q₂ intermediate was optimized at the B3LYP/6-31G(d) level of theory including Grimme's dispersion correction with Becke-Johnson Damping (D3-BJ) and the polarizable conductor model (PCM) (dichloromethane, $\epsilon = 8.9$) as an estimation of the dielectric permittivity in the enzyme active site.¹¹ The partial charges (RESP model)¹² were set to fit the electrostatic potential generated at the HF/6-31G(d) level of theory. The charges were calculated according to the Merz-Singh-Kollman^{13,14} scheme using the Gaussian16 software package.¹⁵ The protonation states were predicted using PROPKA.^{16,17} However, the protonation state of the catalytic residue Lys84 was neutral (i.e., LYN84), as is described in the mechanism at the Q₂ intermediate. All the histidine residues were neutral and protonated in the epsilon nitrogen (i.e., Hie), excluding His334 in *pf*TrpB and OB2-*pf*TrpB systems, His257 in SPM6-TrpB, and His350 in the four systems that are protonated just in the delta position (i.e., HID). The enzyme structures were solvated in a pre-equilibrated truncated octahedral box of 11 Å edge distance using the TIP3P water model, resulting in the addition of ca. 24,000 water molecules, and neutralized by the addition of explicit counterions (i.e., Na⁺) using the AMBER20 leap module. All MD simulations were performed using a modification of the amber99 force field (ff14SB).¹⁸

MD simulation details. The protocol applied for the MD equilibration phase was the one described by Roe and Brooks with small differences fine-tuned to our systems.¹⁹ During the non-production phase, a distance harmonic restraint was applied between the Q₂ intermediate and the catalytic Lys84 to maintain a catalytic pre-organized conformation and a time step of 1 fs (i.e.,

excluding the fourth and fifth equilibration rounds that the time step is 2 fs), to allow potential inhomogeneities to self-adjust. For non-minimization steps, the bonds involving hydrogen are constrained by the SHAKE algorithm. Long-range electrostatic effects were modelled using the particle mesh-Ewald method.²⁰ A 10 Å cut-off was applied to Lennard-Jones and electrostatic interactions. The MD protocol starts with the minimization phase of 1500 steps steepest descent method followed by 3500 steps of the conjugate gradient method with a positional restraint (*i.e.*, force constant of 5.0 kcal·mol⁻¹·Å⁻²) to the protein-heavy atoms. Then, a heating phase is performed with increasing the temperature from 25 K to 300 K during 20 ps of MD simulation, a Langevin thermostat with a collision frequency of 5 ps⁻¹, and a positional restraint (*i.e.*, force constant of 5.0 kcal·mol⁻¹·Å⁻²) to the protein-heavy atoms. The next step is the minimization and heating of all the atoms in the system. Starting with two minimization stages of 1000 steps steepest descent method followed by 1500 steps of the conjugate gradient method each with a positional restraint (*i.e.*, force constant of 2.0 kcal·mol⁻¹·Å⁻² in the first minimization and 0.1 kcal·mol⁻¹·Å⁻² in the second) to the protein-heavy atoms. Then, a third minimization phase of 1500 steps steepest descent method followed by 3500 steps of the conjugate gradient method without any positional restraint is performed. Afterwards, the system is heated in the same way as was previously defined. Finally, a five-round equilibration phase at the NPT ensemble with a constant pressure of 1 atm is performed: whereas the first four were done with the Berendsen barostat, the fifth one with Monte-Carlo barostat. Langevin thermostat with a collision frequency of 1 ps⁻¹ was used in the five equilibration rounds. The first two equilibration rounds of 5 ps had a positional restraint to the protein-heavy atoms with a force constant of 1.0 and 0.5 kcal·mol⁻¹·Å⁻², respectively. A third round of 10 ps equilibration is followed with positional restraint to the backbone-heavy atoms with a force constant of 0.5 kcal·mol⁻¹·Å⁻². The fourth equilibration round of 10 ps was performed without any restraint. The last equilibration round was of 500 ps without any restraint. The production runs were performed at the NVT ensemble with the Langevin thermostat with a collision frequency of 1 ps⁻¹ during 10 ns. Finally, two replicas of equilibration and production runs were performed for each homodimer, reaching a total simulation time of 1200 ns, 1180 ns, 1240 ns, and 1180 ns for *pf*TrpB, *OB2-pf*TrpB, *LBCA*-TrpB, and *SPM6*-TrpB systems, respectively.

Table S1. Mean and standard deviation of the O-to-C values of the AF2 predicted structures for all systems at different MSA depths (section 1 in results and discussion).

<i>MSA</i>	<i>pf</i> TrpB	<i>OB2</i> -TrpB	<i>LBCA</i> -TrpB	<i>SPM6</i> -TrpB
32	5,1 ± 1,6	3,2 ± 2,1	8,8 ± 1,5	8,0 ± 1,5
64	5,1 ± 2,1	5,2 ± 2,2	9,9 ± 1,1	10,5 ± 1,0
128	4,5 ± 2,0	4,9 ± 1,5	10,2 ± 1,1	11,4 ± 0,8
256	5,0 ± 1,3	4,8 ± 1,0	10,6 ± 0,8	11,9 ± 0,5
512	5,0 ± 0,7	5,1 ± 0,8	10,5 ± 0,9	11,5 ± 0,5
1024	5,1 ± 0,7	5,1 ± 0,6	10,0 ± 0,9	11,1 ± 0,7
5120	5,1 ± 0,6	5,0 ± 0,5	10,2 ± 0,8	11,3 ± 0,6

Table S2. Representation of the sequence identity between the different analyzed systems and the available X-ray structures, as well as the selected X-rays for developing the template-based AF2 approach in section 2.

Identity % for *LBCA*-TrpB ALL X-rays. MAX: 100.0, MIN: 69.6

Identity % for *LBCA*-TrpB SELECTED X-rays. MAX: 100.0, MIN: 69.6

Identity % for 0B2-*pf*TRPB ALL X-rays. MAX: 99.7, MIN: 70.4
 Identity % for 0B2-*pf*TRPB SELECTED X-rays. MAX: 99.0, MIN: 70.4

Identity % for *pf*TRPB ALL X-rays. MAX: 100.0, MIN: 70.9
 Identity % for *pf*TRPB SELECTED X-rays. MAX: 99.7, MIN: 70.9

Identity % for SPM6 TrpB ALL X-rays. MAX: 91.1, MIN: 68.3
 Identity % for SPM6 TrpB SELECTED X-rays. MAX: 91.1, MIN: 68.3

Table S3. Percentage of **O-PC** and **PC-C** structures as predicted by the X-ray and MD template-based AF2 for the different analyzed systems. The percentages are provided taking into account all predicted structures with the complete ensemble of templates (named X-ray templates and MD templates), and also by analyzing separately the results with templates with only **O** or **C** conformations. The mean and standard deviation of the **O**-to-**C** values of the AF2 predicted structures for all systems is also provided.

	X-ray templates				MD templates			
	<i>pf</i> TrpB	0B2-TrpB	LBCA-TrpB	SPM6-TrpB	<i>pf</i> TrpB	0B2-TrpB	LBCA-TrpB	SPM6-TrpB
% O-PC	50,8	49,2	16,1	14,8	57,1	52,6	28,6	24,7
% PC-C	49,2	50,8	83,9	85,2	42,9	47,4	71,4	75,3
(O -to- C) \pm std	7,2 \pm 3,0	7,4 \pm 3,2	9,2 \pm 2,2	9,6 \pm 2,2	7,0 \pm 3,3	7,4 \pm 3,7	8,6 \pm 2,3	9,1 \pm 2,5

	X-ray OPEN templates				MD OPEN templates			
	<i>pf</i> TrpB	0B2-TrpB	LBCA-TrpB	SPM6-TrpB	<i>pf</i> TrpB	0B2-TrpB	LBCA-TrpB	SPM6-TrpB
% O-PC	78,6	78,6	35,7	33,3	85,7	85,3	51,4	45,7
% PC-C	21,4	21,4	64,3	66,7	14,3	14,7	48,6	54,3

	X-ray CLOSED templates				MD CLOSED templates			
	<i>pf</i> TrpB	0B2-TrpB	LBCA-TrpB	SPM6-TrpB	<i>pf</i> TrpB	0B2-TrpB	LBCA-TrpB	SPM6-TrpB
% O-PC	28,6	25,7	0,0	0,0	33,3	26,2	9,5	7,1
% PC-C	71,4	74,3	100,0	100,0	66,7	73,8	90,5	92,9

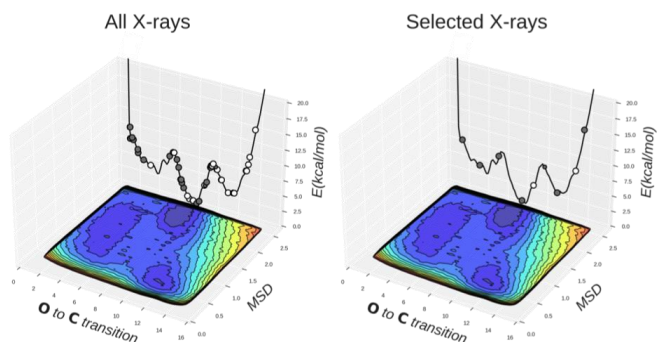


Figure S1. On the left, projection of the available X-ray structures presenting a sequence identity higher than 70% with respect to 0B2-*pf*TrpB. X-ray structures deposited after the training of AF2 network was made are highlighted in white. Multiple structures have been deposited presenting

different levels of closure of the COMM domain, as shown by the x axis that denotes the open-to-closed transition of the COMM domain, which ranges from 1-5 (open, **O**), 6-10 (partially-closed, **PC**), to 11-15 (closed, **C**). The y axis is the mean square deviation (MSD) deviation from the path of **O**-to-**C** structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.²¹ Most stable conformations are shown in blue, whereas higher in energy regions in red. On the right, representation of the 9 X-ray structures used as templates in section 2.

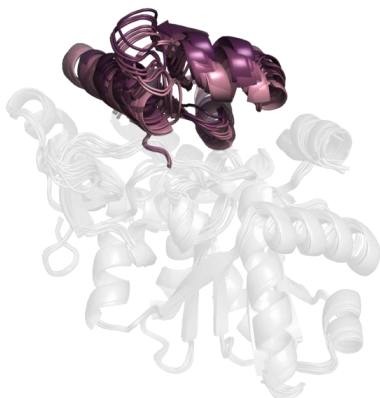


Figure S2. Overlay of the 9 X-ray structures used as templates that display different conformations of the COMM domain.

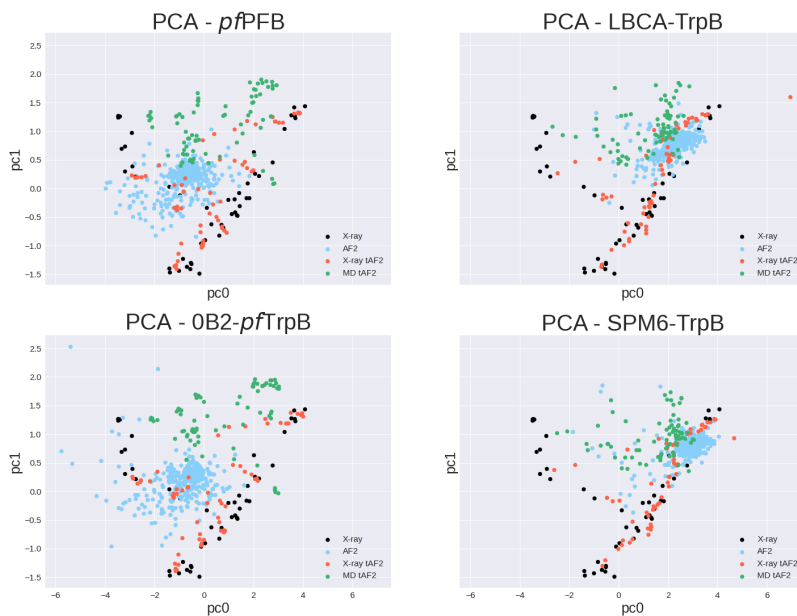


Figure S3. Representation of the principal component analysis (PCA) space created using carbon alpha distances between conserved residues all around the protein for the available X-ray structures presenting a sequence identity higher than 70% with respect to OB2-*pfTrpB* (the different X-rays are represented with black dots). All generated structures with AF2 pipelines are projected into this experimentally-based PC space. AF2 models of the different systems generated without the use of any templates but altering the MSA depths are shown in light blue (named AF2 in the legend), the structures predicted with X-ray template-based AF2 in red (named X-ray tAF2) and MD template-based AF2 in green (MD tAF2).

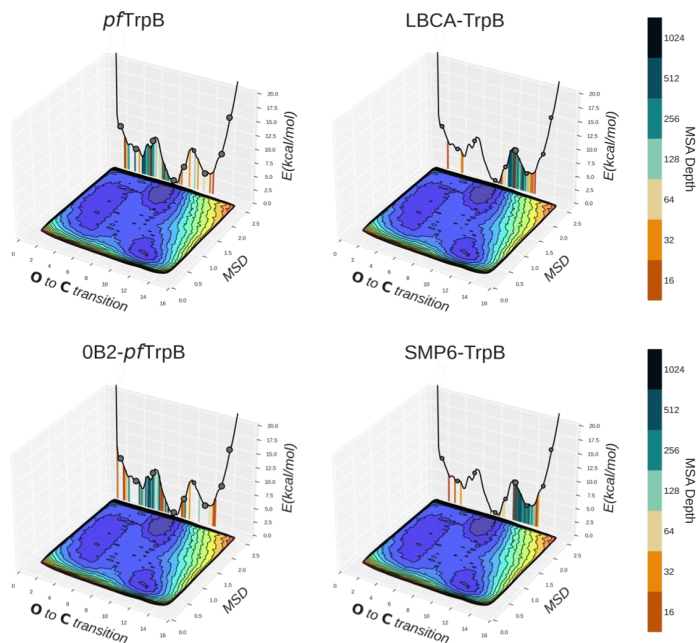


Figure S4. Representation of the previously reconstructed free energy landscape (FEL) of the OB2-*pf*TrpB variant.²¹ The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1-5 (open, **O**), 6-10 (partially-closed, **PC**), to 11-15 (closed, **C**), the y axis is the mean square deviation (MSD) deviation from the path of O-to-C structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.²¹ The predictions of the **X-ray template-based AF2 approach** for the different analyzed systems are represented on the 2D-FEL representation using vertical lines colored from orange to dark blue depending on the MSA depth: AF2 predictions obtained with a 32 MSA depth are shown with a vertical orange line, 64 in light orange, 128 in light brown, 256 in light cyan, 512 in cyan, 1024 in teal, and 5120 in dark blue. Black dots indicate the used X-ray structures as input templates, and the size of the spheres is proportional to the sequence identity of the X-ray with respect to the studied TrpB system. **No side-chain information was included in the provided set of X-ray templates.**

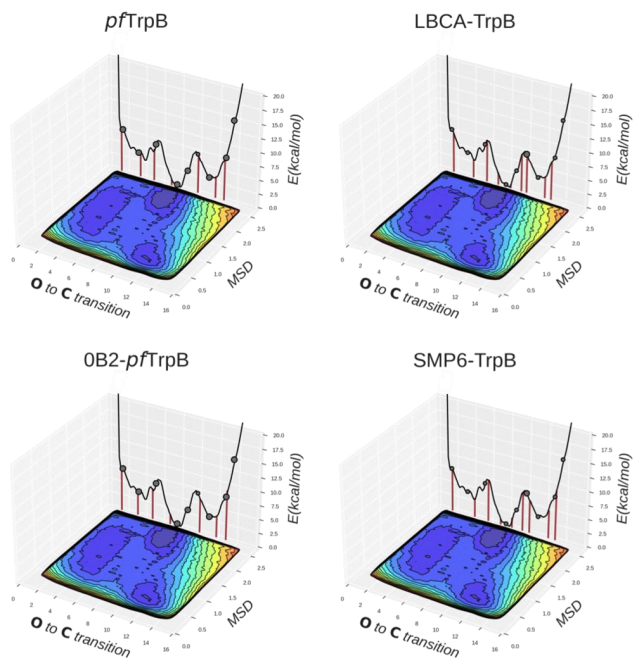


Figure S5. Representation of the previously reconstructed free energy landscape (FEL) of the OB2-*pf*TrpB variant.²¹ The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1-5 (open, **O**), 6-10 (partially-closed, **PC**), to 11-15 (closed, **C**), the y axis is the mean square deviation (MSD) deviation from the path of O-to-C structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.²¹ The predictions of the **X-ray template-based AF2 approach** for the different analyzed systems are represented on the 2D-FEL representation using vertical lines with a **MSA depth of 1**. Black dots indicate the used X-ray structures as input templates, and the size of the spheres is proportional to the sequence identity of the X-ray with respect to the studied TrpB system. **No side-chain information was included in the provided set of X-ray templates.**

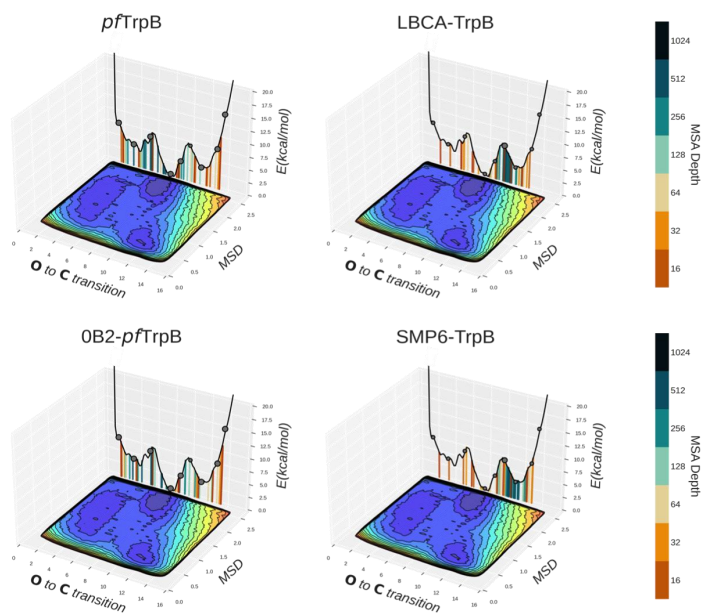


Figure S6. Representation of the previously reconstructed free energy landscape (FEL) of the OB2-*pf*TrpB variant.²¹ The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1-5 (open, **O**), 6-10 (partially-closed, **PC**), to 11-15 (closed, **C**), the y axis is the mean square deviation (MSD) deviation from the path of O-to-C structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.²¹ The predictions of the **X-ray template-based AF2 approach** for the different analyzed systems with vertical lines colored from orange to dark blue depending on the MSA depth: AF2 predictions obtained with a 32 MSA depth are shown with a vertical orange line, 64 in light orange, 128 in light brown, 256 in light cyan, 512 in cyan, 1024 in teal, and 5120 in dark blue. Black dots indicate the used X-ray structures as input templates, and the size of the spheres is proportional to the sequence identity of the X-ray with respect to the studied TrpB system. **Side-chain information was included in the provided set of X-ray templates.**

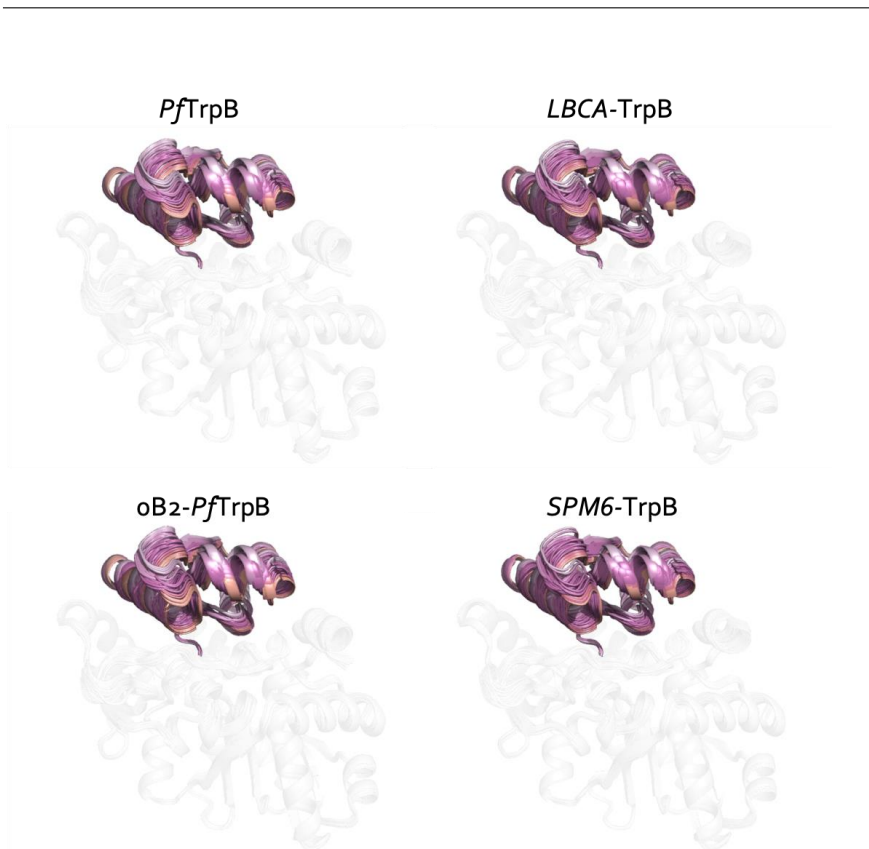


Figure S7. Overlay of the different COM domain conformational states (O highlighted in lilac, PC in pink, and C in light brown) of the predictions obtained with X-ray template-based AF2 approach for the different analyzed systems.

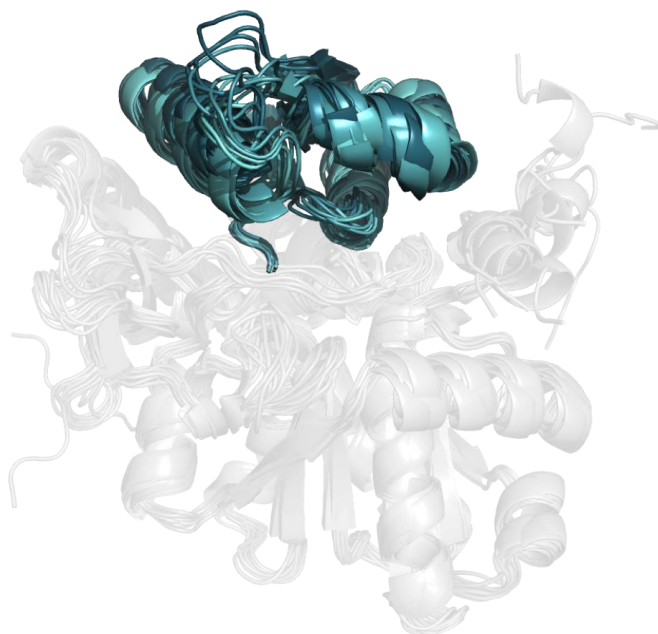


Figure S8. Overlay of the 11 MD conformations used as templates, and extracted from the previously reconstructed free energy landscape (FEL) of the OB2-*pf*TrpB variant.²¹ These 11 structures were used as templates and display different conformations of the COMM domain.

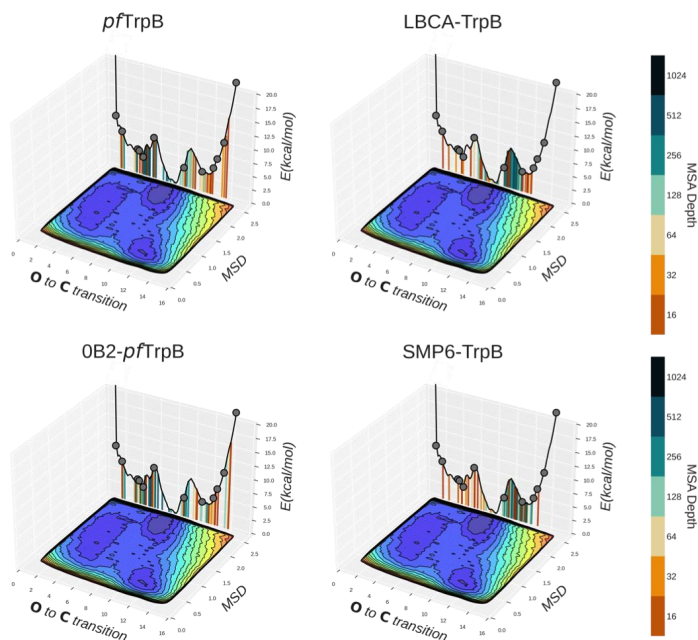


Figure S9. Representation of the previously reconstructed free energy landscape (FEL) of the OB2-*pfTrpB* variant.²¹ The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1-5 (open, **O**), 6-10 (partially-closed, **PC**), to 11-15 (closed, **C**), the y axis is the mean square deviation (MSD) deviation from the path of O-to-C structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.²¹ The predictions of the Molecular Dynamics (MD) extracted template-based AF2 approach for the different analyzed systems with vertical lines colored from orange to dark blue depending on the MSA depth: AF2 predictions obtained with a 32 MSA depth are shown with a vertical orange line, 64 in light orange, 128 in light brown, 256 in light cyan, 512 in cyan, 1024 in teal, and 5120 in dark blue. Black dots indicate the used representative MD conformations as input templates. **No side-chain information was included in the provided set of MD-based templates.**

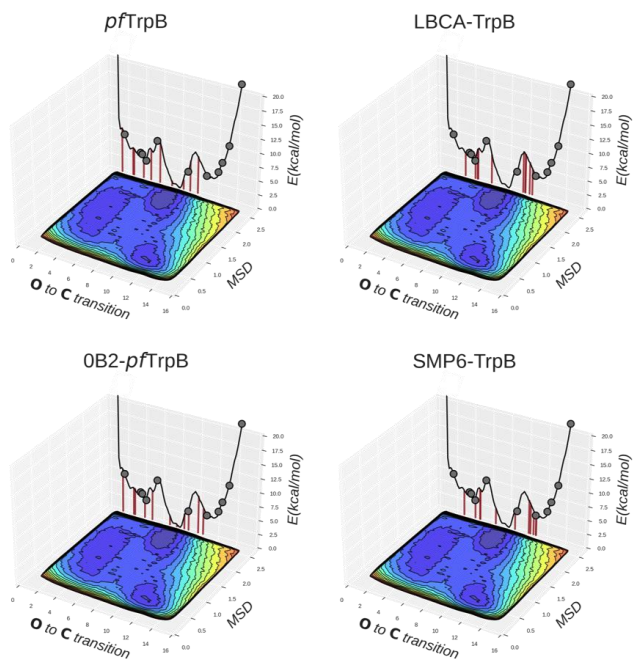


Figure S10. Representation of the previously reconstructed free energy landscape (FEL) of the OB2-*pf*TrpB variant.²¹ The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1-5 (open, **O**), 6-10 (partially-closed, **PC**), to 11-15 (closed, **C**), the y axis is the mean square deviation (MSD) deviation from the path of O-to-C structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.²¹ The predictions of the Molecular Dynamics (MD) extracted **template-based AF2 approach** for the different analyzed systems with vertical lines with a **MSA depth of 1**. Black dots indicate the used representative MD conformations as input templates. **No side-chain information was included in the provided set of MD-based templates.**

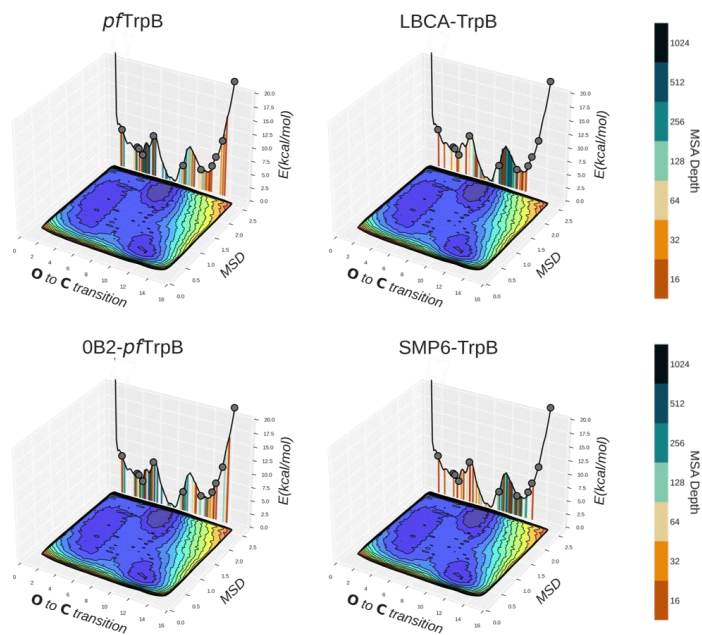


Figure S11. Representation of the previously reconstructed free energy landscape (FEL) of the OB2-*pfTrpB* variant.²¹ The x axis denotes the open-to-closed transition of the COMM domain, which ranges from 1-5 (open, **O**), 6-10 (partially-closed, **PC**), to 11-15 (closed, **C**), the y axis is the mean square deviation (MSD) deviation from the path of O-to-C structures generated. Most stable conformations are shown in blue, whereas higher in energy regions in red.²¹ The predictions of the Molecular Dynamics (MD) **extracted template-based AF2 approach** for the different analyzed systems with vertical lines colored from orange to dark blue depending on the MSA depth: AF2 predictions obtained with a 32 MSA depth are shown with a vertical orange line, 64 in light orange, 128 in light brown, 256 in light cyan, 512 in cyan, 1024 in teal, and 5120 in dark blue. Black dots indicate the used representative MD conformations as input templates. **Side-chain information was included in the provided set of MD-based templates.**

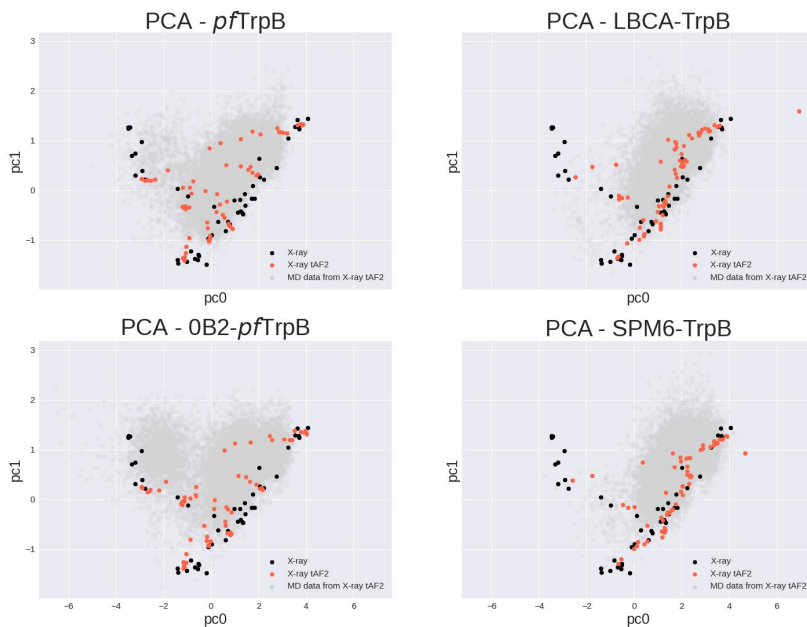


Figure S12. Representation of the principal component analysis (PCA) space created using carbon alpha distances between conserved residues all around the protein for the available X-ray structures presenting a sequence identity higher than 70% with respect to 0B2-pfTrpB (the different X-rays are represented with black dots). AF2 models of the different systems generated with the X-ray template-based AF2 approach are shown in red (named X-ray tAF2), whereas conformations sampled in the multiple replica short nanosecond timescale MD simulations starting from the different AF2 predictions are shown in gray (MD data from X-ray tAF2).

References:

- (1) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domanski, J.; Dotson, D. L.; Buchouz, S.; Kenney, I. M.; Beckstein, O. MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations. *Proc. of the 15th python in science conf.* **2016**, 98-105.
- (2) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. *Proc. of the 7th Python in Science Conf. (SciPy2008)* **2008**, 11-15.

-
- (3) McGibbon, Robert T.; Beauchamp, Kyle A.; Harrigan, Matthew P.; Klein, C.; Swails, Jason M.; Hernández, Carlos X.; Schwantes, Christian R.; Wang, L.-P.; Lane, Thomas J.; Pande, Vijay S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **2015**, *109*, 1528-1532.
- (4) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013**, *9*, 3084-3095.
- (5) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. of 2nd International Conference on Knowledge Discovery and*, 1996; pp 226-231.
- (6) Jukič, M.; Konc, J.; Gobec, S.; Janežič, D. Identification of Conserved Water Sites in Protein Structures for Drug Design. *J. Chem. Inf. Model.* **2017**, *57*, 3094-3103.
- (7) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **2011**, *12*, 2825-2830.
- (8) D.A. Case, K. B., I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F. Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York and P.A. Kollman: AMBER 2020. University of California, San Francisco, 2020.
- (9) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comp. Chem.* **2004**, *25*, 1157-1174.
- (10) 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.*
- (11) Schutz, C. N.; Warshel, A. What are the dielectric “constants” of proteins and how to validate electrostatic models? *Proteins* **2001**, *44*, 400-417.
- (12) 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. *J. Phys. Chem.* **1993**, *97*, 10269-10280.
- (13) Besler, B. H.; Merz Jr., K. M.; Kollman, P. A. Atomic charges derived from semiempirical methods. *J. Comp. Chem.* **1990**, *11*, 431-439.
- (14) Singh, U. C.; Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comp. Chem.* **1984**, *5*, 129-145.
- (15) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J.: *Gaussian 16 Rev. C.01.* Wallingford, CT, 2016.
- (16) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J Chem Theory Comput* **2011**, *7*, 525-537.

-
- (17) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J Chem Theory Comput* **2011**, *7*, 2284-2295.
- (18) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015**, *11*, 3696-3713.
- (19) Roe, D. R.; Brooks, B. R. A protocol for preparing explicitly solvated systems for stable molecular dynamics simulations. *J. Chem. Phys.* **2020**, *153*, 054123.
- (20) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089-10092.
- (21) 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.

Supporting Information of Chapter 5

Supplementary Information

Designing Efficient Enzymes: Eight Predicted Mutations Convert a Hydroxynitrile Lyase into an Efficient Esterase

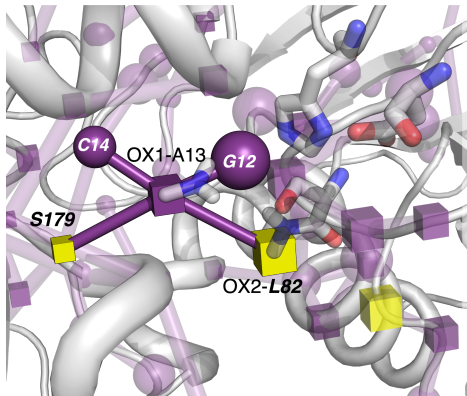
Guillem Casadevall, Colin Pierce, Bo Guan, Javier Iglesias-Fernandez, Huey-Yee Lim, Lauren R. Greenberg, Meghan E. Walsh, Ke Shi, Wendy Gordon, Hideki Aihara, Robert L. Evans III, Romas Kazlauskas, Sílvia Osuna

Supplementary Figures.....	s2
Supplementary Tables.....	s5
Supplementary References.....	s13

Supplementary Figures

SABP2	1	MKEGKHFVLVHGACHGGWSYKLLKPLLEAAGHKVTALDLAASGTDLRKIE	50
	 : : . . . :	
<i>HbHNL</i>	1	-MAFAHFVLIHTICHGAWIWHKLLKPLLEALGHKVTALDLAASGVDPQRQIE	49
SABP2	51	ELRTLVDYDYLPLMELMESLSADEKVIILVGHSLGGMNGLLAMEKYPQKIYA	100
		: : : . : . . : : : : . . : : . .	
<i>HbHNL</i>	50	EIGSFDEYSEPLLTFLLEALPPGEKVIILVGESCGGLNIAIAADKYCEKIAA	99
SABP2	101	AVFLAAFMPDSVHNSFFVLEQYNERTPAENWLDLQFLPYGSPPEEPLTSMF	150
	 : : : : : : : . .	
<i>HbHNL</i>	100	AVFHNSVLPDTEHCPSYVVDKLMVEVFP--DWKDTTYFTYTKDGKEITGLK	147
SABP2	151	FGPKFLAHKLYQLCSPEDLALASSLVRPSSLFMEEDLSKAKYFTDERFGSV	200
	 : : . . : . . : :	
<i>HbHNL</i>	148	LGFTLLRENLYTLTGPEEYELAKMLTRKGSLSLQNILAKRPFPTKEGYGSI	197
SABP2	201	KRVYIVCTEDKGIPEEFQRWQIDNIGVTEAIEIKGADHMAMLCEPQKLCA	250
		: : : . . . : : : : : : 	
<i>HbHNL</i>	198	KKIYVWTDQDEIFLPEFQLWQIENYKPKVKYVEGGDHKLQLTKTKEIAE	247
SABP2	251	SLEIAHKYN	260
		. . : . .	
<i>HbHNL</i>	248	ILQEVADTYN	257

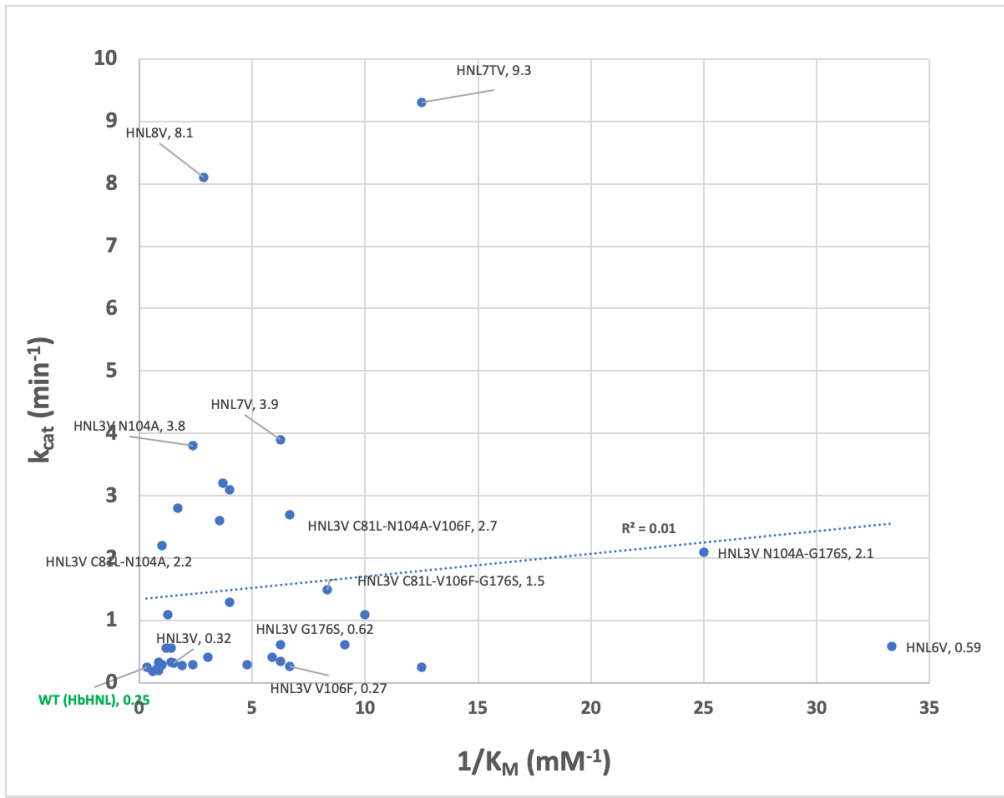
Supplementary Fig. 1 | Pairwise amino acid sequence alignment of SABP2 (UniProt Q6RYA0) and *HbHNL* (UniProt P52704) using the Needleman-Wunsch algorithm.^[1] The comparison over 260 positions identified 114 (44%) identical positions (marked by '|'), 161 (62%) similar positions (identical plus those marked by ':') and 3 (1.2%) gaps (marked by a blank and a '-' in the *HbHNL* sequence). The default settings (EBLOSUM62 matrix, gap penalty of 10, extend penalty of 0.5) of the web tool EMBOSS Needle (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) were used to create the alignment.



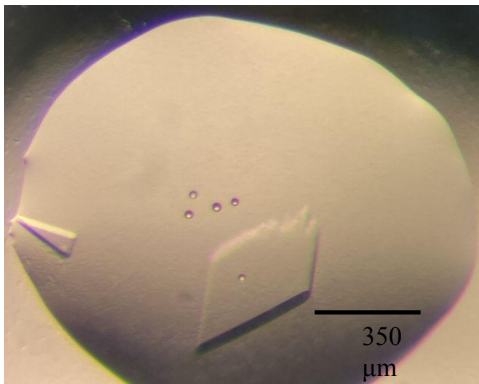
Supplementary Fig. 2 | Zoom and slight rotation relative to Fig. 3 of the SPM of SABP2 showing the correlated motions of OX1 (Ala13) and residues Cys14, Gly12, Ser179, and OX2 (Leu 82). Cys14 and Gly12 are conserved between SABP2 and HNL3V (and are shown as spheres), whereas Ser179 and Leu82 (shown as cubes) correspond to Gly176 and Cys81 in HNL3V. The catalytic residues and amides of the oxyanion hole residues are shown in sticks. C14 and G12 were hard to see in Fig. 3.



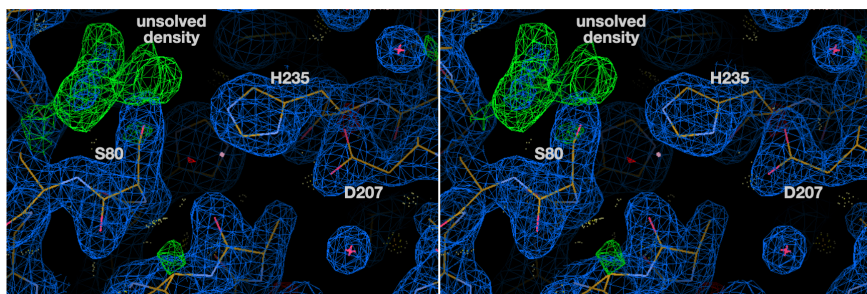
Supplementary Fig. 3 | **Relative frequencies of amino acids at positions 104-107 among 889 homologs of SABP2.** The height of the letters indicates the relative frequency of the amino acid. The positions correspond to positions 103-106 in *HbHNL*. The most common amino acid at SABP2 position 105 is threonine; the second most common is alanine. The multiple sequence alignment was generated using Consensus Finder^[2] (<http://kazlab.umn.edu/>) using default settings and the UniProt SABP2 sequence Q6RYA0. Consensus Finder searches for homologs using BLAST, reduces sequence redundancy by clustering all sequences into groups with 90% sequence identity using CD-HIT and retains only one representative sequence from each cluster. Finally, Clustal W creates the multiple sequence alignment. The image was generated with WebLogo^[3] (<https://weblogo.berkeley.edu/>).



Supplementary Fig. 4 | Improvements in catalytic turnover and binding are independent. Linear regression of k_{cat} (y-axis) vs. $1/K_M$ (x-axis) values for HbHNL variants shows no correlation ($R^2 = 0.01$). k_{cat} values are shown for selected variants.



Supplementary Fig. 5 | HNL6V crystal used for data collection prior to extraction or soaking. Growth conditions: 0.1 M BIS TRIS, pH 5.5, 2.0 M ammonium sulfate. The HNL6V crystals measured 90x175 μm (left crystal) and 350x300 μm (right crystal). The right crystal was harvested for data collection. The five circular “bubbles” are optical artifacts of the microscope.



Supplementary Fig. 6 | Cross-eyed stereo view of the unmodeled density (green mesh) of HNL6V near the active site at a 3.0 Å contour level. The electron density associated with the catalytic triad (S80-D207-H235) is labeled.

Supplementary Tables

Supplementary Table 1 | Steady-state kinetic parameters for hydrolysis of *p*-nitrophenyl acetate of all enzyme variants. See Materials and Methods for experimental conditions and details.

Enzyme	k_{cat} (min ⁻¹)	K_M (mM)	k_{cat}/K_M (M ⁻¹ *min ⁻¹)
SABP2	130±3.7	2.2±0.17	61,000
WT (HbHNL)	0.25±0.02	3±0.4	84
HNL3 = WT T11G-E79H-K235M	0.33±0.02	0.71±0.1	460
HNL3V = HNL3 H103V	0.32±0.02	0.65±0.1	490
HNL3V C81A	0.29±0.04	0.42±0.3	690
HNL3V N104A	3.8±0.23	0.42±0.08	9,100
HNL3V G176S	0.62±0.02	0.16±0.03	3,900
HNL3V I12A	0.28±0.03	0.53±0.17	530
HNL3V C81L	0.28±0.02	1±0.18	270
HNL3V F54L	0.35±0.01	0.16±0.03	2,200
HNL3V V106F	0.27±0.01	0.15±0.05	1,800
HNL3V I209G	0.61±0.07	0.11±0.08	5,500
HNL3V I209G-F210I	0.23±0.04	1.3±0.56	180
HNL3V N104A-G176S	2.1±0.06	0.04±0.01	52,000
HNL3V C81L-N104A	2.2±0.08	0.99±0.13	2,200

Enzyme	k_{cat} (min⁻¹)	K_M (mM)	k_{cat}/K_M (M⁻¹*min⁻¹)
HNL3V N104A-V106F	3.2±0.22	0.27±0.08	12,000
HNL3V V106F-G176S	3.1±0.21	0.25±0.06	13,000
HNL3V C81L-V106F	1.3±0.03	0.25±0.03	5,300
HNL3V C81A-G176S	0.42±0.02	0.17±0.04	2,500
HNL3V C81L-F54L	0.41±0.02	0.33±0.06	1,200
HNL3V C81L-F54L-V106F	1.1±0.05	0.8±0.13	1,400
HNL3V C81L-G176S	0.3±0.02	1±0.23	300
HNL3V C81L-F54L-G176S	0.25±0.01	0.08±0.01	3,100
HNL3V C81L-N104A-V106F	2.7±0.18	0.15±0.05	18,000
HNL3V N104A-V106F-G176S	2.6±0.21	0.28±0.09	9,400
HNL3V C81L-V106F-G176S	1.5±0.02	0.12±0.01	13,000
HNL3V C81L-I12A-V106F	0.56±0.08	0.84±0.41	670
HNL3V C81L-I12A-G176S	0.3±0.01	0.21±0.05	1,400
HNL3V C81L-F54L-V106F-G176S	1.1±0.02	0.1±0.01	11,000
HNL3V C81L-I12A-G176S-V106F	0.56±0.06	0.71±0.26	790
HNL3V C81L-G176S-V106F-I209G-F210I	0.34±0.03	1.2±0.32	290
HNL3V C81L-G176S-V106F-I209G-F210I-L121Y-F125T	0.2±0.01	1.2±0.16	170

Enzyme	k_{cat} (min ⁻¹)	K_M (mM)	k_{cat}/K_M (M ⁻¹ min ⁻¹)
HNL3V I209G-F210I-L121Y-F125T	0.7 ± 0.11	0.97 ± 0.12	720
HNL6V (HNL3V C81L-N104A-G176S)	2.3±0.02 ^a	0.13±0.01 ^a	18,000 ^a
HNL7V (HNL3V C81L-N104A-V106F-G176S)	3.9±0.45	0.16±0.07	25,000
HNL7TV (HNL3V C81L-N104T-V106F-G176S)	9.3±0.33	0.08±0.04	120,000
HNL8V (HNL3V C81L-N104A-S105A-V106F-G176S)	8.1±0.38	0.35±0.06	23,000
SABP2 A104N	29±1.6	3±0.16	9,600

^a Activity was measured at a higher temperature (29°C) relative to other variants (22 ±2 °C). Increased temperature correlates with higher observed reaction rates.

Supplementary Table 2 | Rank of *HbHNL* variants ordered by k_{cat} and K_M

Ordered by k_{cat}			Ordered by K_M	
Variant	k_{cat} (min ⁻¹)	Rank order	K_M (mM)	Variant
HNL7TV	9.34	1	0.04	HNL3V N104A-G176S
HNL8V	8.13	2	0.08	HNL7TV
HNL7V	3.92	3	0.08	HNL3V C81L-F54L-G176S
HNL3V N104A	3.81	4	0.1	HNL3V C81L-F54L-V106F-G176S
HNL3V N104A-V106F	3.21	5	0.11	HNL3V I209G
HNL3V V106F-G176S	3.14	6	0.12	HNL3V C81L-V106F-G176S
HNL3V C81L-N104A-V106F	2.72	7	0.13	HNL3V C81L-N104A-G176S (aka HNL6V)

Ordered by k_{cat}			Ordered by K_M	
Variant	k_{cat} (min^{-1})	Rank order	K_M (mM)	Variant
HNL3V N104A-V106F-G176S	2.63	8	0.15	HNL3V C81L-N104A-V106F
HNL6V (HNL3V C81L-N104A-G176S)	2.3	9	0.15	HNL3V V106F
HNL3V C81L-N104A	2.2	10	0.16	HNL7V
HNL3V N104A-G176S	2.08	11	0.16	HNL3V G176S
HNL3V C81L-V106F-G176S	1.52	12	0.16	HNL3V F54L
HNL3V C81L-V106F	1.33	13	0.17	HNL3V C81A-G176S
HNL3V C81L-F54L-V106F	1.12	14	0.21	HNL3V C81L-I12A-G176S
HNL3V C81L-F54L-V106F-G176S	1.08	15	0.25	HNL3V V106F-G176S
HNL3V G176S	0.62	16	0.25	HNL3V C81L-V106F
HNL3V I209G	0.61	17	0.27	HNL3V N104A-V106F
HNL3V C81L-I12A-V106F	0.56	18	0.28	HNL3V N104A-V106F-G176S
HNL3V C81L-I12A-G176S-V106F	0.56	19	0.33	HNL3V C81L-F54L
HNL3V C81A-G176S	0.42	20	0.35	HNL8V
HNL3V C81L-F54L	0.41	21	0.42	HNL3V N104A
HNL3V F54L	0.35	22	0.42	HNL3V C81A
HNL3V C81L-G176S-V106F-I209G-F210I	0.34	23	0.53	HNL3V I12A

Ordered by k_{cat}			Ordered by K_M	
Variant	k_{cat} (min^{-1})	R a n k order	K_M (mM)	Variant
HNL3	0.33	24	0.65	HNL3V
HNL3V	0.32	25	0.71	HNL3V C81L-I12A-G176S-V106F
HNL3V C81L-G176S	0.3	26	0.71	HNL3
HNL3V C81L-I12A-G176S	0.3	27	0.8	HNL3V C81L-F54L-V106F
HNL3V C81A	0.29	28	0.84	HNL3V C81L-I12A-V106F
HNL3V I12A	0.28	29	0.99	HNL3V C81L-N104A
HNL3V C81L	0.28	30	1	HNL3V C81L-G176S
HNL3V V106F	0.27	31	1.04	HNL3V C81L
HNL3V C81L-F54L-G176S	0.25	32	1.16	HNL3V C81L-G176S-V106F-I209G-F210I-L121Y-F125T
HNL3V I209G-F210I	0.23	33	1.17	HNL3V C81L-G176S-V106F-I209G-F210I
HNL3V C81L-G176S-V106F-I209G-F210I-L121Y-F125T	0.2	34	1.3	HNL3V I209G-F210I
HNL3V I209G-F210I-L121Y-F125T	0.19	35	1.7	HNL3V I209G-F210I-L121Y-F125T

Supplementary Table 3 | Mutagenic primers for site-directed mutagenesis

Primer/sequence name	Primer sequence
I208G-F209I Rev	tcggtccacacataaatcttc
I208G-F209I Fwd	ccaagacgaaggtattttacctgaatttcaactctgg

I208G Fwd	ccaagacgaaggtttttacctgaattcaac
F209I Fwd	ccaagacgaaataatttacctgaattc
C81L-L54F-G176S-H103V Fwd	gattggctcatttgatgagtattc
C81L-L54F-G176S-H103V Rev	tctcaattgccttggg
81L-103V-104A-106F-176S Fwd	tgtttcgtcgcgtcattttgccagacac
81L-103V-104A-106F-176S Rev	gcagctgcaatctttcac
103V-104A Fwd	tgtttcgtcgcgtcagtattgccagacac
103V-104A Rev	gcagctgcaatctttcac
103V-81L-104A-106F Fwd	tgtttcgtcgcgtcattcttgccagacac
103V-81L-104A-106F Rev	gcagctgcaatctttcac
103V-C81L-N104T-V106F-G176S Fwd	tgtttcgtcacctcattcttgccag
103V-C81L-N104T-V106F-G176S Rev	gcagctgcaatctttcac
H103V_C81L_N104A_Fwd	ttcgtcgcgtcagtattgccagacacc
H103V_C81L_N104A_Rev	acagcagctgcaatctttcacagtattatcagc
103V_N104A_V106F_Fwd	gtcattcttgccagacaccgagcac
103V_N104A_V106F_Rev	gcgacgaaaacagcagctgcaatctttc
C81L Fwd	cagcctgggaggactcaatatagcaattg
C81L Rev	tggccaaccagaatcacctttccccc

SABP2 A104N Fwd	tgtttcttgaacgcttcatgcctg
SABP2 A104N Rev	gcagcatagatcttttgtg
SABP2 A104N Gibson vector Fwd	ttaaaggctgatcacatggcaatgctatg
SABP2 A104N Gibson vector Fwd	taaccttctcatctgctgaaagagattcc
SABP2 A104N Gibson gene Fwd	ttcagcagatgagaaggtatattagtggg
SABP2 A104N Gibson gene Rev	catgtgatcagcaccttaatctctattgc
C81A Fwd	attctggttgccatagcgcctggaggactcaatagc
C81L Fwd	tggccacagcctgggaggactcaatag
C81L Rev	accagaatcacctttcc
H103V-N104A-S105A-V106F Fwd	tgccagacaccgagcactgccat
H103V-N104A-S105A-V106F Rev	agaatgccgcgacgaaaacagcagc
F54L Fwd	gattggctcactggatgagtattc
F54L Rev	tcctcaattgccttggg
G176S Fwd	gacaaggaagagctcattattcaaaatatttagc
G176S Rev	aacatcttegccagtcatattc
H103V Fwd	gattgcagctgctgttttctcaattcagattgccagac
H103V Rev	gtctggcaataactgaattgacgaaaacagcagctgcaatc
I12A Fwd	tattcatggcgcgtgccacgggtc
I12A Rev	agaacaaaatgagcgaatg

L121Y F125T Fwd	tatatggaggtgacccccgactgaaagacacc
L121Y F125T Rev	cttatccacgacgtaagatgggc
V106F Fwd	ccacaattcattttgccagacac
V106F Rev	aaaacagcagctgcaatc

Supplementary Table 4 | Crystallization conditions for the x-ray structure determination of HNL6V

Method	Vapor diffusion, sitting drop
Plate type	CrystalMation Intelli-Plate 96-3 low-profile
Temperature (K)	293
Protein concentration (mg ml ⁻¹)	9.3
Buffer composition of protein solution	5 mM BES, pH 7.2
Composition of reservoir solution	0.1 M Bis-Tris, pH5.5, 2 M (NH ₄) ₂ SO ₄
Volume and ratio of drop (nl)	200, 1:1 (protein:screen solution)
Volume of reservoir (μl)	50

Supplementary Table 5 | Data collection and processing for the x-ray structure determination of HNL6V.

Values in parentheses are for the outer shell.

X-ray source	APS BEAMLINE 24-ID-C
Wavelength (Å)	0.979 Å
Detector	DECTRIS EIGER2 S 16M
Exposure Time (s)	0.2
Crystal-to-detector distance (cm)	230
Angle increment (°)	0.2
Resolution Range (Å)	43.03 -1.99 (2.04-1.99)
Space Group	C222 ₁
a, b, c (Å)	47.054, 106.378, 128.396
α, β, γ (°)	90, 90, 90
Matthews coefficient (Å ³ Da ⁻¹)	2.74
Solvent Content (%)	55.07

Total reflections	102923 (7877)
Unique Reflections	20181 (1606)
Multiplicity	5.1
Mosaicity (°)	0.2
Completeness (%)	89.43 (89.95)
(I/σ(I))	13.9 (2.3)
Wilson B Factor (Å ²)	28.09
Rmerge	0.063 (0.703)
Rmeas	0.077 (0.861)
Rp.i.m	0.032 (0.490)
CC _{1/2}	0.998 (0.689)

Supplementary Table 6 | Structure refinement for the x-ray structure determination of HNL6V. Values in parentheses are for the outer shell.

Reflections used in refinement	20178 (2005)
Reflections used for R _{free}	2000 (199)
R _{work}	0.1849 (0.2285)
R _{free}	0.2374 (0.2603)
No. of non-H atoms	
total	2162
Macromolecules	2030
Ligands	0
Solvent	123
No. of protein residues	255
R.m.s.d, bonds (Å)	0.009
R.m.s.d, angles (°)	1.03
Ramachandran favored (%)	96.44
Ramachandran preferred (%)	96.44
Ramachandran allowed (%)	3.56

Ramachandran outliers (%)	0.00
---------------------------	------

Supplementary References

1. Needleman S. B. & Wunsch C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453 (1970).
2. Jones, B. J., Kan, C. N. E., Luo, C. & Kazlauskas, R. J. Consensus Finder web tool to predict stabilizing substitutions in proteins. *Method. Enzymol.* **643**, Enzyme Engineering and Evolution: General Methods 129–148 (2020).
3. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A sequence logo generator, *Genome Res.* **14**, 1188-1190 (2004).