UVIC
UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL
DE CATALUNYA

Development of a framework for the computational design and evolution of enzymes

Thesis for the Doctoral degree in Bioinformatics

Facultat de Ciències i Tecnologia

November 2021

Martin Floor Pilquil

Supervisors: Jordi Villà i Freixa and Luis Agulló Rueda

In his house at R'lyeh, dead Cthulhu waits dreaming

— H. P. Lovecraft

Acknowledgements

I want to thank everyone who has been a part of this journey.

Thank you, Jordi, for believing in me and giving me the opportunity and the support over these years. Without you, my luck, and probably my destiny, would have been entirely different. So, thank you for being the great guy you are! I've learned a lot by working at your side, and it has been an honour.

Thank you, Luís, for being my lab mate from the beginning. Thanks for all the time and chats we have shared while figuring out where this work was going. Thanks for all the car rides, meals, trekkings, laughter, and of course, your friendship. It has been just great having you around!

I also want to thank Pau, the last addendum to this body that is the CBBL. Thanks for the coffee, the chats, the great sense of humour, and especially for sharing your office! It has been great working with you.

I want to thank Lynn and Jenn, who graciously hosted me in their labs. I remember you and all your people with whom I shared space and time very fondly.

I want to thank everyone who has made me feel at home again.

Thank you, Elisabet, for accompanying me throughout these years of adaptation and learning. You have become an essential part of this journey. I love the time we spend together; it truly has been a precious epoch of mutual discovery.

I also profit from this opportunity to thank your family, as they have embraced me as one of their own. Special thanks to Rosa, who has welcomed me unquestionably to her family, and to Arturo, who has always shared everything.

Thanks, Leu, for feeling me as a part of your family. This journey has been an excellent opportunity to know each other, and it has been great to have you nearby and get to know you a little bit better.

Thanks to all my new Catalan friends with whom I have shared many great moments throughout these years.

I want to thank everyone who was not close in space.

Thanks, mom, for accepting me the way I am. For all the love you have given me, and for all the love I know we have not been able to share. Because of you, I am the way I am. I love you.

Thank you to all my family and friends back in Chile. You have been an essential part of my life, and you still are. I truly miss you.

I want to thank everyone who is not close in time.

I want to thank my father for believing in his dreams.

I finally want to thank everyone who supported me.

I thank the UVic-UCC for the PhD fellowship.

I thank the Spanish Subprograma Estatal de Generación del Conocimiento, BIO2017-83650-P project, for partially funding the work of this thesis.

I thank the Barcelona Supercomputing Centre for the computing time provided from activities: BCV-2020-3-0019, BCV-2020-1-0001, and QCM-2018-1-0038.

Resum

El disseny enzimàtic es troba al cor de la biotecnologia moderna i, cada cop més de les anomenades química fina i química verda. Dissenyar un enzim per a aplicar-lo en contextos industrials o de bioremediació, per exemple, implica disposar d'un coneixement profund dels sistemes enzimàtics, per tal de poder proposar canvis racionals que millorin les seves propietats catalítiques. En els darrers anys, s'ha desenvolupat un gran nombre de mètodes computacionals amb l'objectiu de dissenyar o millorar nous enzims. No obstant això, aconseguir, mitjançant aquests mètodes, que les prediccions de nous enzims assoleixin el poder dels enzims naturals és encara un repte científic no assolit.

En aquesta tesi, proposem la combinació de dues metodologies robustes per idear un marc computacional de disseny i evolució d'enzims. D'una banda, una metodologia exitosa de disseny de proteïnes, l'entorn de treball Rosetta, i, de l'altra, un mètode eficient per l'avaluació de reactivitat química basat en simulacions moleculars, el mètode de l'Empirical Valence Bond. Ambdues eines —treballant col·lectivament— són una proposta atractiva per afrontar reptes capdavanters en el camp del disseny enzimàtic.

Després d'aplicar la nostra metodologia en un sistema químic habitualment utilitzat com a prova de concepte, la reacció catalítica de la Kemp eliminasa, hem trobat un seguit d'obstacles que cal abordar abans de crear un marc reeixit per al disseny computacional i l'evolució d'enzims. En aquest treball, explorem aquests reptes en profunditat i suggerim noves direccions per millorar diferents aspectes de la metodologia proposada. En concret, per una banda fem una dissecció acurada de les energies d'interacció que proporciona Rosetta, aspecte clau per a una millor predicció de marcs (o scaffolds) estructurals sobre els quals construir nous dissenys enzimàtics. Per una altra, proposem una nova implementació pràctica d'un model de simulació basat en estructura en el paquet OpenMM de simulacions moleculars. Ambdós elements són un pas de gran rellevància en la consecució de l'objectiu de disposar d'una "caixa d'eines" eficient i robusta per a l'exploració del mapa estructura-funció dels enzims dissenyats.

Abstract

Enzymatic design is at the heart of modern biotechnology and, increasingly so, the so-called fine chemistry and green chemistry. Designing enzymes for applications in industrial or bioremediation contexts, for example, involves having a deep knowledge of enzymatic systems to propose rational changes that improve their catalytic properties. In recent years, a large number of computational methods have been developed to design or improve new enzymes. However, achieving enzymatic predictions through these methods to reach the power of natural enzymes is still an unattained scientific challenge.

In this thesis, we propose the combination of two robust methodologies to devise a computational framework for enzyme design and evolution. On the one hand, a successful protein design methodology– the Rosetta modelling environment– and on the other, an efficient method for evaluating chemical reactivity based on molecular simulations– the Empirical Valence Bond method. Both tools, working collectively, are an attractive proposition for tackling state-of-the-art challenges in the field of enzymatic design.

After applying our methodology in a proof-of-concept chemical system, the catalytic reaction of Kemp eliminase, we found a series of obstacles that need to be addressed before creating a successful framework for the computational design and evolution of enzymes. This work explores these challenges in-depth and suggests new directions to improve different aspects of the proposed methodology. Specifically, on the one hand, we make a careful dissection of the interaction energies provided by Rosetta, a key aspect for a better prediction of structural frames (or scaffolds) on which to build new enzymatic designs. On the other hand, we propose a new practical implementation of a structure-based simulation model in the OpenMM package of molecular simulations. Both elements are a critical step in achieving an efficient and robust "toolbox" for exploring the structure-function map of designed enzymes.

Content

Abbreviations

ANK - Ankyrin repeat
CA - Alpha carbon
CC - Correlation coefficient
DE - Directed evolution
EVB - Empirical valence bond
FEEC - Fold-Excluded Evolutionary Conservation
GNCA4 - Gram-negative bacteria beta-lactamase class A
GPU - Graphical processing unit
IC50 - Half-maximum inhibitory concentration
ITS - Implied timescale
IκBα - Nuclear factor kappa-light-chain-enhancer of activated B cells inhibitor, alpha
KE - Kemp elimination
LRA - Linear response approximation
MD - Molecular dynamics
MHC - Major histocompatibility complex
MHC-I - MHC class I
MHCp - MHC-peptide
MM - Molecular Mechanics
MSM - Markov State Model
NF-κB - Nuclear factor kappa-light-chain-enhancer of activated B cells
NLS - Nuclear localisation signal
PCC - Pearson correlation coefficient
PCCA - Perron-cluster cluster analysis
QC - Quantum chemical
QM - Quantum mechanical
REU - Rosetta energy units
RHR - Rel Homology Regions
RMSD - Root-mean square deviation
RMSF - Root-mean square fluctuations
RS - Reactant state
SASAh - Hidden solvent-accessible surface area upon complexation
SBM - Structure-based model
SC - Sequence conservation
TICA - Time-structure Independent Components Analysis
TS - Transition state
TSA - Transition state analogue
VAMP - Variational approach to the Markov process
WCN - Weighted contact number

Introduction

Enzymes are Nature's biological catalysts to accelerate biochemical change. At the most basic level, the need for living cells to oppose entropic decay constrains biochemical reactions to be carried out at increased speeds. This requirement means that all necessary chemical processes not occurring spontaneously in an aqueous solution at physiological temperature compel a biological catalyst, positioning enzymes at the heart of all cellular metabolic processes.

In addition to their fundamental role in cellular biochemistry, enzymes have an important place in human applications. From their ancient use in food fermentation to their more recent applications in industrial endeavours,[1] enzymatic activity is still finding new ways to be applied in modern chemical processes. They offer clear advantages to using standard organic chemical synthesis by leveraging the use of high temperatures, pressures, and concentrations of organic solvents. Enzymatic chemistry can be carried out in an aqueous solution at environmental temperatures with high selectivity and specificity. Moreover, enzymes can also be adapted to work at drastically different temperatures, pH, and even in low concentrations of organic solvents,[2] making them adaptable systems for a diverse set of applications. All these features make enzymes attractive targets for developing new highly efficient green-chemistry operations.[3]

In the physicochemical landscape of biophysical behaviour, the mechanisms by which enzymes catalyse reactions have been the subject of a large amount of scientific literature. A sizable chapter of experimental and theoretical studies has shed light upon this question; however, there is no definitive consensus on how enzymes physically work.[4] The pioneering proposition of Linus Pauling[5] devised enzymes to work by stabilising the reactions' transition state (TS) to a greater extent than its ground state (i.e., substrate), leading to diminished activation energies and thus to accelerated chemical reactions. The discovery that transition state analogues (TSA) can act as specific enzymatic inhibitors[6] or be used as haptens to elicit catalytic antibodies for target reactions[7] has supported this idea. However, despite this apparently simple rationalisation of enzymatic action, connecting enzymatic structure with its catalytic activity has proven an enormous challenge.[8–14]

Enzymes catalyse diverse types of reactions, therefore, a plethora of experimental results has led to a broad set of interpretations on the molecular origins of their catalytic effect. Examinations to explain enzymatic activity should include all aspects of the physical process, encompassing structural and dynamical facets of the enzymatic molecular system, and, perhaps, more importantly, a thorough characterisation of their different thermodynamics contributions. In this regard, computer simulations have aided in interpreting the complex enzymatic behaviour by connecting fundamental physical theory with the kinetics of enzymatic processes.[15] This interpretation has been of fundamental importance to advance our understanding of chemical catalysis since there are no yet physical or spectroscopic methods capable of observing the structure of the short-lived TS in enzymatic reactions.

In the last decades, rational enzyme design has become the holy grail search for proof on how enzyme catalysis works. The field got to a turning point when artificial enzymes were computationally designed for a series of chemical reactions with unknown natural counterparts. Although this did not come without controversy,[16,17] enzymes for retro-aldol reaction, Kemp elimination (KE), Diels-Alder, among others,[18–22] were successfully expressed and tested in many proof-of-concept experiments. The success was mainly based on applying optimisation methods using a knowledge-based score function[23] that had already performed successfully on protein folding prediction experiments[24] and de novo scaffold design[25] but now applied to improve the protein interactions with a virtual TS model. The whole method was later deemed the inside-out approach,[26] and it is still a popular method for active-site design, especially when attempting to endow proteins with *de-novo* catalytic activities.

Despite the success of computational methods in giving active enzymes for reactions not known to be catalysed by any natural enzyme, they are still considered flawed regarding attained catalytic efficiencies.[27] The first attempts screened tens of proposals, but only a few turned out to be active.[26] Paradoxically, the experimental structures of many active designs were highly close to their computational models; however, catalytic rates were orders of magnitude below what could be expected when considering the activities of natural enzymes.[28,29] Further studies later revealed that there were important flaws in the prediction of bound ligands,[30] which could explain the low success rates obtained.[18–20] These results led to questioning the catalytic hypothesis on how enzymatic catalysis could be achieved by these computational design methods and opened

the field into an exciting pursuit to understand the physical aspects that contribute most to the catalytic phenomena.

Nobel laureate Arieh Warshel's work made essential contributions to characterising the physical contributors to enzyme catalysis and deemed electrostatic preorganisation one of the main contributors to the process.[31] The idea was supported by computer simulations that queried many natural enzymatic reactions by comparing how the reactions behaved in water and the solvated protein system.[9,31–34] In the enzymatic active site, electrostatics played an essential role in lowering the activation energy since protein electrostatic dipoles were pre-oriented towards stabilising the target reaction's TS. On the contrary, water dipoles were not pre-oriented towards binding the reaction's TS, incurring high reorganisational energy as the reaction progresses from substrate to TS. Notably, the estimated activation energy differences between the water and enzyme simulations agreed with the observed catalytic effects of natural enzymes. These comparisons could not have been possible by traditional quantum chemical (QC) methods since they are still too costly to explore the conformational dynamical aspects of the solvated enzymatic system. However, the empirical valence bond (EVB) approach,[35] devised by Warshel in these studies, was able to simulate the protein and reacting system dynamics, using parameterized energy functions as low-cost proxies for the QC interactions, allowing better exploration of protein and water conformations throughout the reaction coordinate.

The EVB method’s success in predicting activation energies of natural enzymatic processes has positioned it as an appealing method for screening computational designs. Several attempts have been used to explore the catalytic effect of mutations, either by rational choice[36] or by mutational scanning of active site residues.[37] Despite this, few reported enzymatic optimisations or new designs using these methodologies to obtain improved variants have been published. One reason could be that single-point mutation approaches, such as alanine scanning strategies, rarely significantly impact catalytic activity; multiple coordinated mutations are usually needed to improve enzymatic activity.[38] On the other hand, most molecular dynamics (MD) methods are still too costly to sample the protein conformational landscape in a convergent manner, limiting the explored reacting trajectories to conformations near the native structure, thus, sometimes, overlooking relevant competing configurations that enzymes could adopt.[12]

Despite the failure of current computational methods in delivering high catalytic rates for artificial enzymes, once activity can be measured on a new scaffold, directed evolution (DE) methods can increase these activities further, giving sometimes rise to full-fledged enzymes, capable of catalytic rates comparable to the power of natural enzymes.[39] Most of what we know about how enzyme catalysis can be improved towards native-like activities comes from these experiments, in which several rounds of mutagenesis (and sometimes recombination) and selection are iterated to screen variants with improved activity. Despite attaining great success, the DE method is not fully rational since there is no clear principle about how these mutations improve the enzyme's catalytic activity, and, often, computational methods must be used to assess their physical origin.[27,38] Notably, some DE evolutionary trajectories have been structurally recorded by crystallizing relevant variants along the full optimisation path, giving meaningful, although still small, datasets to study the enzymatic structure/activity relationship.[40–42]

There are many intriguing aspects of enzymatic activity optimisation through DE. First, many of the selected mutations are not in direct contact with the substrates and some are very distant from the active site. This finding has led to speculation about the effect of distant mutations over catalytic rates. The most conspicuous hypothesis seems to be allosteric effects over the preorganisation of active site residues,[12] while others also propose that electrostatic fields could help boost transition-state stabilisation.[11] It is not clear how allosteric effects can be propagated to the active site; however, MD and crystallographic studies have shown that a network of interactions could help to propagate and constraint interactions to pre-organise active site residues.[43–45]

Other works in computational enzyme design have suggested that the optimisation potential of active sites depends upon the properties of the selected scaffolds. In this regard, reconstructed phylogenetic nodes, representing ancient evolutionary enzymes, have been deemed "more evolvable" for protein optimisation given their increased thermostability, which could allow them to accept an increasing number of mutations before losing their characteristic folds.[46] As an applied case, the best artificial Kemp Eliminase was designed in a TIM barrel fold, which is deemed exceptionally well suited for the evolvability of enzymatic activity.[45]

A pivotal point in understanding the importance of these catalytic factors occurring in enzymatic

reactions is to be able to guide their rational design and refinement. Despite the great advances made through experimental mutational studies, a quantitative understanding of these factors would not be possible without the use of computer simulation experiments. In this regard, guiding the computational design of enzymatic reactions, from first principles and using a defined catalytic hypothesis, would be a major establishment of their importance.

This work focuses its first effort on combining two state-of-the-art methodologies, each successful in enzymatic studies and design. On the one hand, the EVB method, which by MD simulations of the catalytic system, can address the effect that a particular environment has over the free energy of a specific chemical reaction. On the other hand, the Rosetta approach to protein design, that helps optimise and suggest putatively improved variants towards high catalytic activities.

Both methods are complementary since they address the design protocol at different stages. First, designs are suggested based on active-site optimisation to create variants that stabilise a TS model of the target reaction. Second, the catalytic activity of the suggested variants is evaluated using EVB simulations that can rank the simulated variants and pinpoint which residues are responsible for the catalytic improvements or mechanisms. Therefore, a fully-computational iterative approach could be devised in which the idiosyncrasies learned through simulations can be transferred to the protein design protocol to suggest variants with an improved likelihood of increasing the catalytic activity of designed systems.

The working hypothesis of this thesis relies on the computational studies of enzymatic reactivity carried out with the EVB framework, in which electrostatic preorganisation has been defined as the driving force of the catalytic activity of natural proteins. However, this hypothesis is not the central question of this work, since the main objective is setting up a computational framework to deliver variants for enzymatic systems with improved catalytic constants. Nonetheless, this hypothesis is a good starting place to address the interpretations derived from applying simulations to assess the catalytic activity of designed enzymatic variants.

From the initial hypothesis, though, the work has been deriving towards the exhaustive characterisation of the potential energy surface by means of simplified structure based models

and proper balance between minimisation and molecular simulations protocols, in an attempt to build some of the tools that hopefully will ultimately lead to a complete protocol for rational design of enzymatic scaffolds for non-naturally occurring chemical reactions. Both the objectives and the structure of the thesis below reflect the critical steps in achieving a proper protocol for enzyme design.

Hypothesis

"Differential atomic charge distributions developed over the reaction coordinate accounts for the major differences observed in ground-state versus transition-state stabilisation upon enzyme binding. It is possible to exploit this differential stabilisation using the EVB model through an exhaustive sampling of the available conformational space to optimise computationally designed enzymes."

Objectives

Main goals

- ➢ To combine complementary methodologies to set up a fully-computational evolutionary scheme for enzymatic design.
- ➢ To identify theoretical and practical challenges derived from applying the proposed computational enzyme design evolutionary scheme.
- ➢ To circumvent the challenges arising from the implementation of the computational enzyme design evolutionary scheme.

Specific Goals

- ➢ To test the combined implementation of the Rosetta enzymatic design protocol with EVB evaluations using a redesign strategy.
    - ○ To validate the Rosetta energy function to model the target system interactions.
    - ○ To test different enzyme design protocols to generate enzyme variants with improved catalytic properties.
    - ○ To rank the designed models using a rapid and systematic assessment of their catalytic capabilities.

- To validate the EVB simulations of the target reaction for delivering proper catalytic parameters.
- To apply EVB simulations to assess activation free energies for a ranked subset of the designed variants.

➢ To understand the catalytic effect of the original and improved enzymatic systems over the target reaction.
- To run single-residue-uncharged simulations to assess the catalytic effect at the residue level.
- To explore through an LRA analysis, between the substrate and the transition state region, the catalytic effects of individual residues.
- To assess the effect of enzymatic positional constraints over the prediction of activation free energies.

After addressing the previous specific objectives, we identified theoretical and practical challenges for improving the enzyme design framework, and the following additional specific objectives were derived:

➢ To validate the Weighted Contact Number (WCN) as a fast metric to predict protein residue-level dynamics.
- To validate the use of the WCN for predicting experimental dynamic profiles of protein systems.
- To explore the effect of WCN in a binding prediction test case.

➢ To validate the use of binding free energies, derived with the Rosetta score function, as a tool to predict binding activities.
- To compile a binding dataset for a statistically-significant validation of the binding free energy metric.
- To validate the binding free energy metric to explain experimental binding data.
- To understand the physical origin of the predicted binding free energies.

➢ To create a library for running structure-based simulations on the OpenMM platform

- To create a Python-based library to complement the OpenMM API to write SBM force fields.
- To validate the library in generating the correct force field energies.
- To validate the application of SBM to explore native basin simulations.
- To demonstrate the use of the library in a biophysical question.

Structure

This thesis is divided into four chapters, each describing different points at which the work was focused.

The first chapter deals with applying the Rosetta design methodology in conjunction with EVB simulations to evaluate the activity of the designed variants. The starting point is a reconstructed beta-lactamase ancestral fold in which Kemp Eliminase activity was built in a secondary active site. The new active site was created by a few mutations, achieving important initial catalytic activity. This activity is still below the value of other KE enzymes, however, and given the properties of reconstructed ancestral folds, there is still room for improving this secondary active site in a search for higher catalytic activity. The results obtained point to several shortfalls of the methodology, which should be addressed before successfully applying the evolutionary protocol to improve enzyme activity. Among them, there is the need of:

1. Create better filtering metrics for the designed variants before moving to the most costly screening steps (chapter 2).
2. Create an optimisation target function able to capture the preorganisation of active site residues (chapter 3).
3. Find better and faster methods to sample the protein configurational space (Chapter 4).

As indicated, subsequent chapters deal with most of these challenges and provide possible ways to solve these problems.

The second chapter uses a knowledge-based score function (e.g., the Rosetta score function) to predict binding free energies using the Major Histocompatibility Complex I (MHC-I) receptor and

a dataset of bound peptides as a model system. This validation is relevant for the step of quickly screening designs before being evaluated by more costly MD simulations.

The third chapter addresses the use of the WCN metric, capable of capturing the relationship between structure, dynamics, and evolution. The metric is validated to predict dynamics profiles from protein structures and in predicting protein-protein interaction sites based on evolutionary and chemical analysis. This metric is promising to develop a score capturing the preorganisation of active site residues.

Finally, in chapter four, a Structure-Based Model (SBM) simulation package, the SBMOpenMM library, is implemented for the fast exploration of protein conformational landscapes. The method is validated and exemplified in the study of a protein folding reaction. The method can be helpful in efficiently sampling enzyme conformations.

An appendix containing brief descriptions of the methodologies employed in this thesis is included at the end for consultation.

Chapter 1 - Enzyme optimisation: the Kemp Elimination case

Improving the catalytic performance of a low-activity variant is still a big challenge for computational enzyme design methods. It requires simultaneously predicting the effect of multiple mutations over the scaffold stability and the activation free energy of the target reaction. Scaffold stability design has been addressed by optimising the sequence/structure space using energy functions capable of discriminating between native and non-native conformations and giving agreeable energies for protein mutational changes. The Rosetta score function[47] has been developed towards this goal and employed to create new scaffolds and other protein design applications.[48,49] However, it is still unclear how to use it successfully to optimise and select enzymatic variants with improved catalytic activities.

Most enzyme design strategies that employ the Rosetta energy function attempt to create active-site complementarity to bind the target reaction TS model. The optimisation target is the system's total energy, or, sometimes, the interface score (interaction energy) between the TS-ligand and the protein system. These optimisation parameters are used as targets to improve the catalytic activity during the design algorithm; therefore, they are critical players in the generation of new enzymatic proposals.

After the design optimisation is carried out, the produced designs are filtered and ranked by using a set of metrics such as total score (enzyme complex energy), ligand-binding interface score, hydrogen bonding, active-site geometry, packing scores of the active-site cavity, among others.[26] Many of these metrics complement the Rosetta score function; however, they do not directly account for catalytic activity and are guided by the designer's chemical intuition about enzymatic structures.

Given the difficulty in predicting catalytic activities for the designed variants, all computational activity predictions, if any, are usually carried out after the design process has ranked and selected a set of optimised variants or, as has been previously done, they are assessed after being tested experimentally.[30] A standard tool to explore catalytic activity, either directly or indirectly, are MD simulations of the enzymatic system. They aid in showing the dynamical

progression of the system as a whole and can shed light upon the stability of particular conformations.

The main limitation of these MD studies is that they do not directly evaluate the kinetic barriers, instead, they use a series of geometric cues to classify conformations into catalytically competent or otherwise. An alternative to classic MD simulations is the EVB method that can directly query the activation free energy of the simulated chemical reaction. In this fashion, designed enzyme models could be simulated and ranked upon the correct physicochemical parameter being pursued.

We explored combining the Rosetta design methodology[50] with EVB simulations for the screening step in an enzymatic redesign strategy to suggest improved catalytic variants (see Appendix 1 - Methodologies for more details on these methods). We employed as a model system a reconstructed beta-lactamase ancestral fold with initial Kemp Eliminase activity.
The scaffold has its enzymatic activity built upon a Precambrian beta-lactamase resurrected through ancestral sequence reconstruction.[51]

The strategy to design this catalytic system was a single amino acid substitution of a conserved tryptophan (W) to an aspartate (D) residue. The objective of this substitution was two-fold; first, the W side-chain is quite similar to the KE substrate (Figure 1B); second, the D side chain is small and has proton abstraction ability to act as the catalytic base in the KE reaction (Figure 1A). Of course, in any given scaffold substituting a W with a D will not immediately create an enzyme with KE catalytic activity; however, reconstructed enzymes have different dynamical properties than their modern counterparts. Ancestral enzymes have greater flexibility and deformability and, therefore, can adopt alternative pocket conformations, giving rise to promiscuous activities.[46] Evidence for this is suggested by the fact that the W to D substitution strategy worked only on resurrected scaffolds, but not in modern beta-lactamase variants.[51]

Despite the simple strategy of a single amino acid substitution to come up with a *de novo* protein catalyst, the designed Kemp eliminase has a high catalytic activity, comparable to more complex computational enzyme design strategies, which involved a more significant number of mutations on their protein scaffolds in creating their *de novo* activity.[20] However, its catalytic activity is still

magnitudes lower than other artificial enzymatic counterparts obtained through DE strategies.[42] Because of this, it is reasonable to hypothesise that this designed enzyme have still room for further improvements.

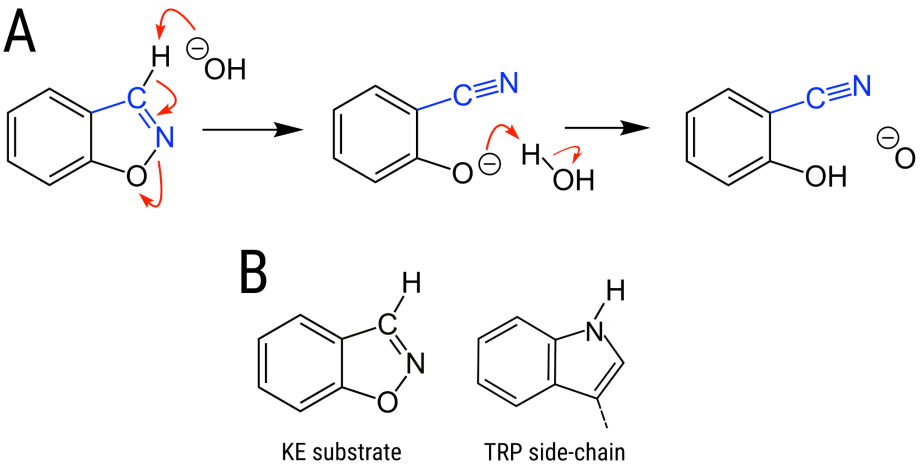

Figure 1. A) the mechanism of the KE reaction. B) The KE substrate (left) in comparison with the Tryptophan side chain (right).

Results

Validating the Rosetta score function to model KE enzymatic interactions

The *de novo* Kemp eliminase variant is based on the W229D mutation over a last common ancestor of Gram-negative bacteria beta-lactamase class A (GNCA4) scaffold. The structure of the GNCA4-W229D variant is available in complex with the KE reaction TSA (Figure 2). Besides the mutation W229D, in which D was incorporated as the catalytic base, H291, makes a critical interaction to stabilise the oxyanion hole developed at the TS region.

Before attempting an optimisation process to find TS-stabilising mutations in the active site, it was essential to validate the score function employed to describe the crystallographic ligand pose as a minimum energy conformation. We tested this by self-docking the TSA of the KE reaction upon the GNCA4-W229D structure (Figure 3). The interface score between the KE reaction TSA and the protein was calculated for each docked conformation and was plotted together with the ligand root-mean-square deviation (RMSD) to the crystallographic ligand conformation (Fig. 3A).

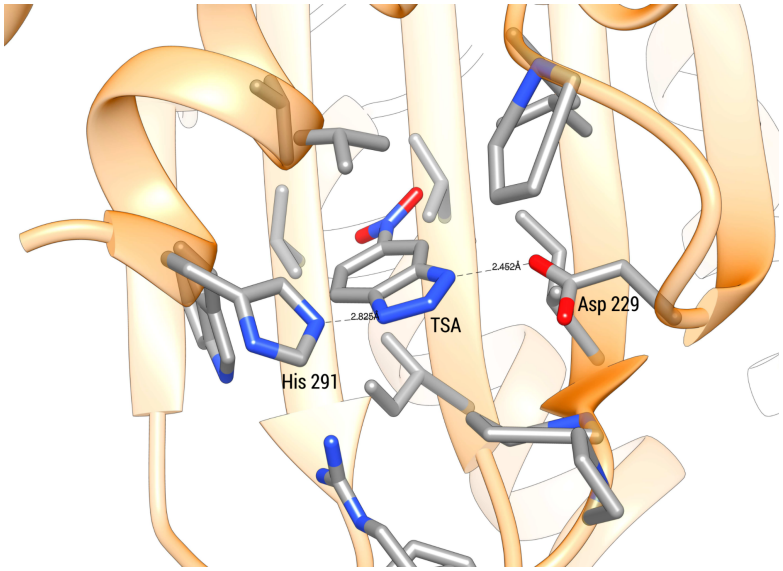

Figure 2. Active site structure of the GNCA4-W229D variant with Kemp eliminase activity in complex with the KE reaction TSA (6-nitrobenzotriazole). Interactions between the catalytic base (D229) and stabilizing the oxyanion hole (by residue H291) are shown as dotted black lines. The side chains of other residues surrounding the TSA in the active site are also depicted.

The self-docking test describes a low-RMSD population of conformations that does not have the best interface scores. The conformations with the best interface score are at 4 to 5 Å RMSD from the TSA's crystallographic conformation. The interface score is a helpful metric to calculate interaction energies between the ligand and the protein system and is widely adopted in docking calculations to select the best-docked conformations.[52] However, this metric is not a predictor of the most populated conformation of the system. Therefore, to better describe the sampled conformational landscape of the docked system, we plotted the system total Rosetta Score to map where these low-interface-score docked conformations lay in the potential energy surface (Figure 3B).

Despite having good interactions between the ligand and the protein system, the minimum interface score conformations were not necessarily the minimum-energy conformations of the system. In these cases, the overall system sacrifices stabilising interactions to adopt ligand conformations with increased protein-ligand interactions, making these conformations less likely (although not inconsequential) to be adopted in a thermodynamic ensemble. The RMSD to the native pose (colour bar in Figure 3B) shows that indeed the minimum-energy conformations correspond to the lowest RMSD structures. This result confirms the ability of the score function

to model protein-ligand interactions to describe the native ligand pose as the minimum energy conformation, and suggest caution when using or validating intermolecular interaction metrics for selecting poses in ligand-docking ensembles.

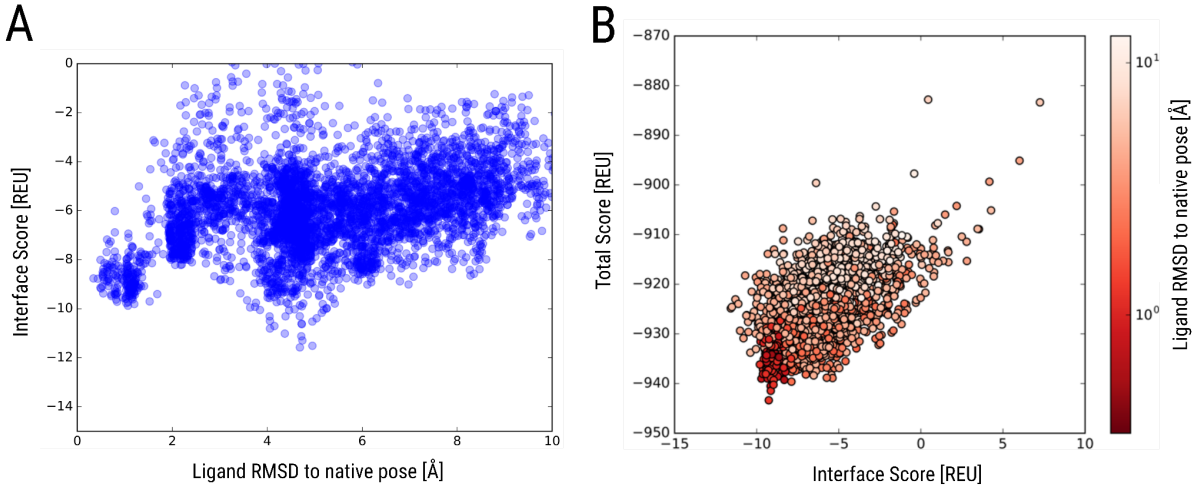

Figure 3. Self-docking test of the KE reaction TSA over the GNCA4-W229D active site. This docking test docks the TSA over the same protein conformation that binds the TSA to generate decoys evaluated by the energy function. The docking ligand conformational landscape is shown as the interface score vs the ligand RMSD to the crystallographic pose (A) or as the complex (GNCA4-W229D + TSA) total score vs the interface score of the bound TSA for each docked pose, with the ligand RMSD shown in a coloured dimension (colour bar) as a logarithmic scale for easy visualization (B).

Generating newly redesigned variants for the Kemp Elimination catalysis

The strategy to suggest new variants is based on the simultaneous optimisation of backbone, ligand, and side-chain (sequence-space) degrees of freedom. This optimisation was carried out upon the highest activity variant available (GNCA4-W229D-F290W) of our model system.[51] The method uses catalytic restraints to ensure the catalytic base D229 is positioned at a proton abstracting distance and also defines different distance cut-offs from the ligand to select which protein residues will be designable (i.e., change their amino acid identity) or repackable (i.e., treated as flexible but maintaining their identity). Original versions of the enzyme design algorithm used a fixed (or minimally flexible) backbone approach to propose new designs.[53] Here, we have selected a fully-flexible-backbone approach that expands the repertoire of scaffold conformations and, thereby, is expected to generate an increased number of different proposals. We employed the two design protocols to test this hypothesis: EnzDes,[53] a rigid backbone

approach, and FastDes,[54] a flexible-backbone approach (for details on these protocols, see Methods).

The starting scaffold for the optimisation protocol is the GNCA4-W229D-F290W variant including the F290W mutation over the wild type background, which stabilises the KE reaction ligand-system through a face-to-edge interaction made by the tryptophan side chain.[51] Each method produced ten thousand design trajectories that were analysed by their ligand RMSD to the starting ligand position, total score, and interface score energies (Figure 4).

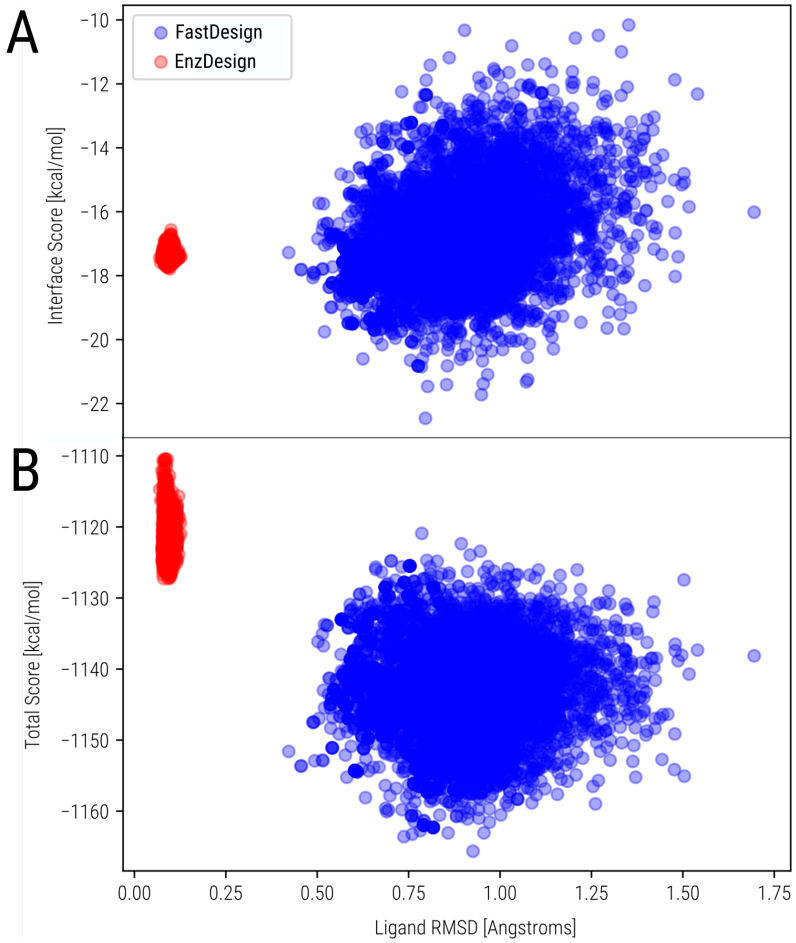

Figure 4. Potential and binding energy landscape of the enzymatic designs based on the GNCA4-W229D-F290W for the KE reaction. The interface score (up) or the total score (down) was used to visualise the distribution of ligand RMSDs for each design trajectory. The EnzDes strategy (red dots) generates structures with low RMSD and low variability in the energy scores. On the contrary, the FastDes strategy creates a high dispersion in total and interface scores, with increased, although not too high, RMSD values.

The EnzDes protocol produces a much narrower distribution of total and interface scores and very low ligand RMSD values than the FastDes protocol. On the other hand, this latter protocol allows an exploration of lower total and interface scores, and a more diverse set of ligand conformations, in agreement with the idea that backbone flexibility is required to make scaffold conformations more compatible with TS binding.

However, it is unclear if lower total scores produced by the FastDesign protocol came from mutations very far from the active site. To confirm the behaviour of the design algorithm, we evaluated the different amino acid identities explored by each design protocol at each protein position (Figure 5). On the one hand, the rigid backbone protocol changed very little the amino acid identity of the protein during optimisation. This result validates the score function in recognising the protein native sequence as the optimal sequence for the input conformation but also renders the rigid backbone protocol purposeless for enzymatic redesign strategies unless, previously, compatible backbone conformations were sampled, and an ensemble of precomputed structures were used as input for this protocol.

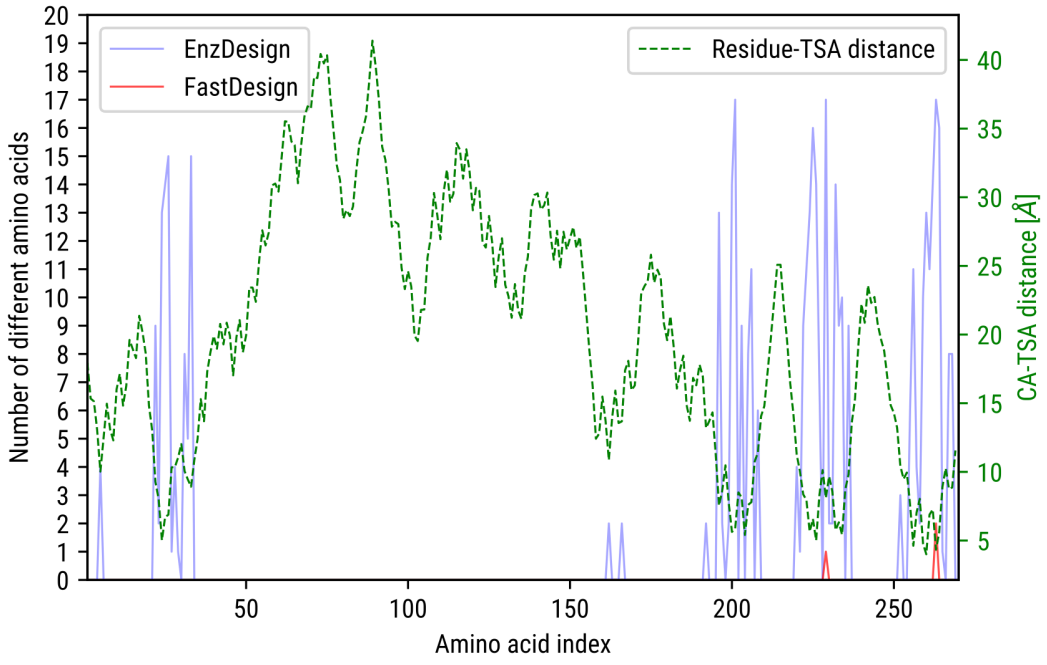

Figure 5. Mutational frequency of each amino acid position in the GNCA4-W229D-F290W scaffold. Few positions were changed with the EnzDes method (blue line), while numerous did when the FastDes (red line) protocol was used. The atomic distance between the residues' alpha carbon (CA) atoms and the TSA closest atom (green dashed line) is shown to control how far mutations were produced relative to the ligand starting position.

On the other hand, the flexible-backbone protocol explored the sequence landscape more richly when searching for optimal protein sequences to stabilise alternative protein backbone and ligand conformations. As is expected, there are regions not covered by the sequence search since they are far away from the ligand, confirming the protocol's restriction to avoid changes in their identities (Figure 5). The fixed-backbone protocol only introduced one, and rarely, two mutations per design, while the flexible-backbone protocol introduced a minimum of four and a maximum of twenty-four mutations, with an average of twelve mutations per design.

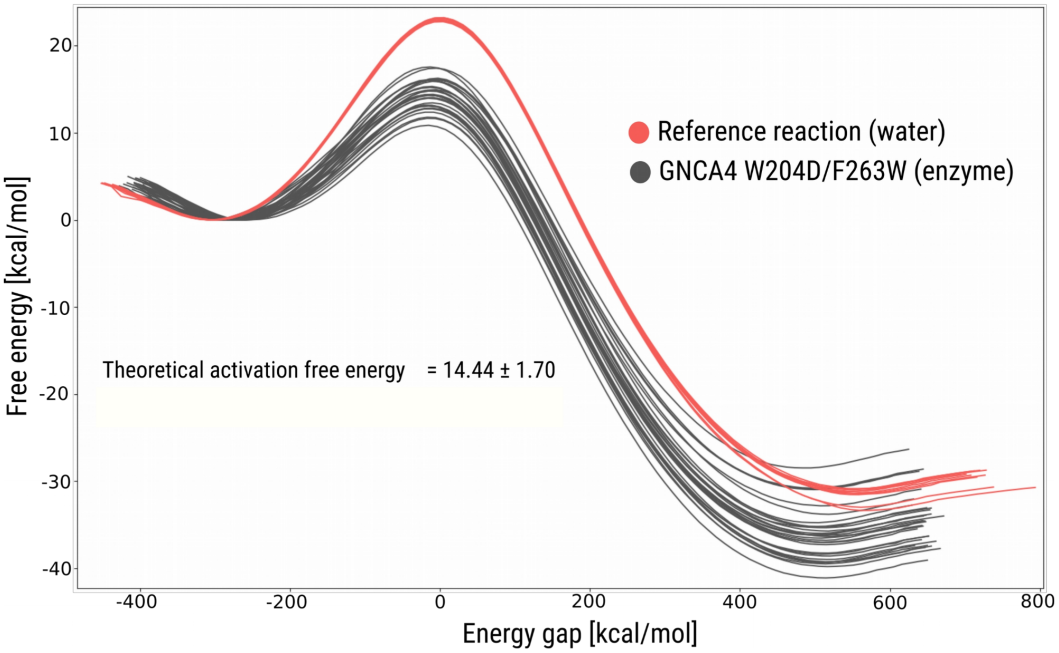

Figure 6. EVB free energy of reactive trajectories for the KE reaction. Trajectories for the reference reaction in water (red curves) were parameterized to match an activation free energy of 21.2 kcal/mol. The same parameters were used to analyse the enzymatic reaction occurring in the GNCA4-W229D-F290W variant (black curves).

Validation of the Kemp elimination EVB simulations

We set up EVB simulations to query the catalytic activity of a selection of FastDesign-produced models. We first parameterized a reference simulation in aqueous solution by adjusting the EVB parameters to match the *ab initio* activation free energy, as published in other works.[38] Using the same parameters obtained for adjusting the reference water reaction, we ran the KE simulations in the active site of the GNCA4-W229D-F290W variant. We obtained a 14.44 ± 1.70 kcal/mol value

for this EVB simulation, which agrees with the experimental value for this enzymatic variant of 16.62 ± 0.12 kcal/mol[51] (Figure 6). These results show that the EVB simulations for the KE enzymatic reaction have the correct trend and could, in principle, be applied to screen the catalytic effect of other enzymatic variants over the KE reaction.

Relationship between Rosetta interface scores and EVB activation free energies

Querying activation free energies with MD simulations is computationally costly. So, to maximize the fraction of good designs to be tested with EVB simulations, it is essential to find metrics to filter out designs with low chances of being catalytically active. Therefore, we ran EVB simulations of selected enzymatic designs to find a possible relationship between the EVB-calculated activation free energies and the Rosetta TS interface scores.

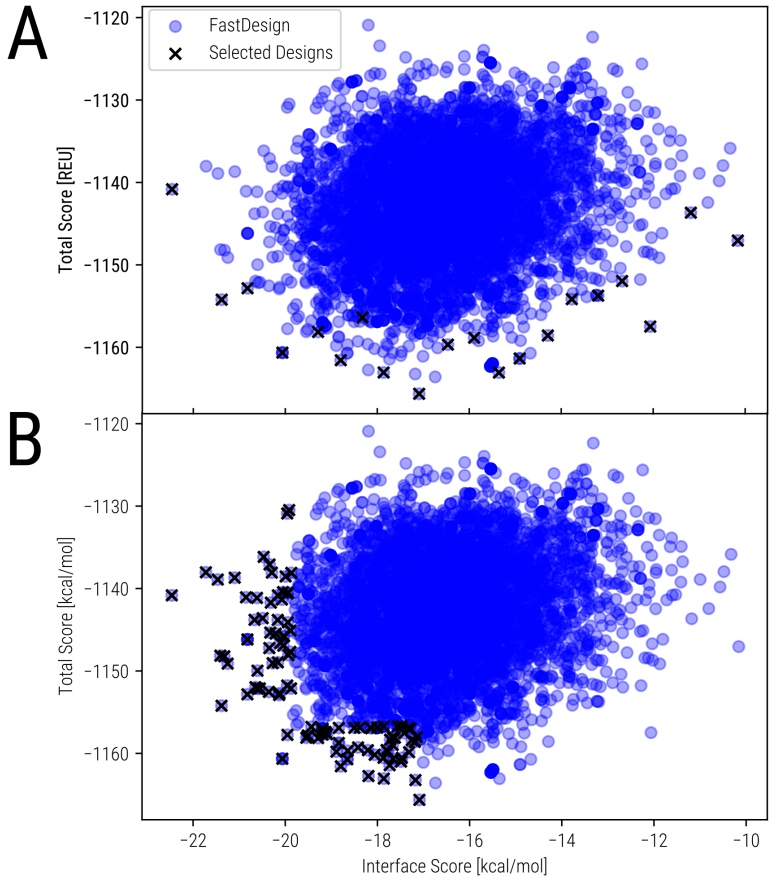

Figure 7. Selected designs for EVB assessment. A) Selected minimum-energy designs encompassing a wide range of interface scores. B) A preselection of a hundred designs with the lowest interface (50) or total score (50) to be ranked according to Rosetta-derived binding free energies.

First, we selected designs covering a wide range of interface score values by taking the minimum energy models that spanned most of the designed interface score range (Figure 7A).

A hundred independent EVB trajectories were run for each selected model, and the estimated activation free energy was taken as the average value of all sampled trajectories. On the other hand, designed models were locally sampled by removing the catalytic constraints and using the Rosetta energy function to generate conformational ensembles of the designed enzymes in complex to the TS ligand. From these ensembles, the expectation value of the interface score was calculated from a Boltzmann distribution based on their total scores (see "Local conformational search of designed enzyme" below for details on the method). From now on, these Boltzmann-averaged interface scores will be referred to as TS binding free energies.

For the 20 models evaluated, the correlation between their EVB-estimated activation free energies and their interface scores obtained from the design protocol is 0.2344. However, when compared with the TS binding free energies, the correlation value increases to 0.3978. These low values could reflect that TS binding free energies are not a direct measure of activation free energies, among other possible artifacts (see Discussion for more details). However, since the correlation was higher when deriving the TS binding free energies from an ensemble of conformations, it seems reasonable to select models using this metric for further screenings with EVB simulations.

Screening a set of designed variants with EVB simulations

Given the cost of obtaining convergent results with MD simulations, it seems infeasible to sample all designed models. Therefore, we focused on selecting designs according to two general metrics (Figure 7B). On the one hand, we selected the 50 models with the lowest interface score, which ensures designs with good interactions between the TS ligand and the protein. In addition, we selected the 50 designs with the lowest total score, increasing the likelihood that designed models are more stable to fold into the designed conformation correctly. The one hundred designs selected this way were subjected to a local conformational sampling to estimate their TS binding free energies. For further EVB assessment, we only selected the 20 models with the best TS binding free energies. 25% of these models came from the lowest total scores selection, while 75% were from the ones selected by the lowest interface scores.

We only obtained good EVB reaction free energy profiles for 18 of the 20 designs (Table 1). Most designs have higher activation free energies than the EVB-derived value of the starting GNCA4-W229D-F290W variant (14.4 ± 1.70 kcal/mol). Only the 5FQK_KET_0487_00523 variant had a significantly better activation free energy (13.9 ± 0.70 kcal/mol) than the original variant, although this value was only 0.5 kcal/mol better.

The low success obtained from this screening stage raises concerns regarding the current strategy's ability to optimise *in silico* low-activity enzymatic variants. Three main reasons could have affected these results. First, there is no certainty that the enzyme design strategy, in its current state, can produce models that have better catalytic properties than the starting variant. Second, the filtering stage is still deficient and has not been validated to deliver correct binding free energies; additionally, because of computational limitations, not all 10000 models produced in the design stage were considered for this filtering, and only a subset of 100 models was. Finally, there was no high-quality validation that the implemented EVB simulations could correctly rank the models regarding their catalytic ability.

We first decided to explore a more thorough validation of the EVB simulations, specifically their ability to rank related enzymatic variants with increased catalytic activity.

Evaluating KE directed-evolution trajectories with EVB simulations

The ability of the EVB set up to correctly rank the free energy of related KE enzymatic variants can be queried using a dataset of available structures and experimentally obtained catalytic constants. To this end, we employed a small dataset of enzymatic structures representing the DE of the HG-3 artificial Kemp eliminase system.[42] This system represents different structure-activity points on the evolutionary trajectory towards improving the catalytic activity for the KE reaction (Table 2).

For the variants in Table 2, we carried out EVB simulations analogous to the ones employed to screen the catalytic activity of the KE enzymatic designs. However, to match more closely the experimental activation free energies of these variants, we changed the reference reaction to match the starting point of the DE optimisation experiment, i.e., the HG3 variant.[30] We also

include the HG2 variant as a control since it has no reported catalytic constant because it was not detectable in the original publication.[30] The HG2 variant is the precursor of the HG3 variant and was designed computationally from knowledge derived from ground state molecular dynamics.[30]

Design Name	ΔGcat (kcal/mol)	S.D. (kcal/mol)
5FQK_KET_0487_03769	13.5	3.0
5FQK_KET_0487_00523	13.9	0.7
5FQK_KET_0487_09862	15.4	2.4
5FQK_KET_0487_03439	15.8	2.4
5FQK_KET_0487_00207	16.2	2.7
5FQK_KET_0487_09645	16.3	2.2
5FQK_KET_0487_04400	16.3	1.6
5FQK_KET_0487_04913	16.5	1.5
5FQK_KET_0487_03912	16.9	2.4
5FQK_KET_0487_09509	17.4	1.9
5FQK_KET_0487_09594	17.5	2.9
5FQK_KET_0487_05456	17.7	1.9
5FQK_KET_0487_09827	17.8	2.0
5FQK_KET_0487_08314	18.0	1.4
5FQK_KET_0487_00549	18.2	4.8
5FQK_KET_0487_00750	18.4	1.4
5FQK_KET_0487_06142	18.5	1.3
5FQK_KET_0487_01128	19.3	2.5

Table 1. EVB-derived activation free energies of selected designs. All values are in kcal/mol. S.D.: Standard deviation.

A comparison between the resulting EVB activation free energies and the experimental values for each specific variant is shown in Figure 8. There are problems when comparing the theoretical (calculated) and experimental (observed) absolute values. The calculated energies show that the reference reaction (i.e., the HG3 variant) has the best (lowest) activation free energy of all the

simulated models; all evolved variants have higher calculated energies than the HG3 variant, even though they represent superior points in the path of directed-evolution optimisation. Nonetheless, the HG2 variant[30] has higher activation energy than the HG3 variant, which denotes a prediction with the correct trend.

Despite the failure in defining absolute free energies, all evolved variants show the correct trend in their calculated activation free energies among them, agreeing with the idea that EVB could serve the purpose of ranking computational designs. This result is relevant when screening for future experimental validation since correct absolute free energy predictions should not severely affect the ranked selection, given the same trends are maintained. Regardless, a more thorough examination of this specific EVB simulation procedure must still be assessed to understand the origin of the incorrect absolute free-energy predictions.

Variant	Kcat (S-1)	ΔG‡cat (kcal/mol)	Mutations
HG3	3.0 ± 0.1	16.8 ± 1.0	-
HG3.3b	14 ± 2	15.9 ± 0.6	V6I, K50H, M84C, S89R, Q90D, A125N
HG3.7	310 ± 130	14.1 ± 0.4	V6I, Q37K, K50Q, M84C, S89R, Q90H, A125N
HG3.14	490 ± 100	13.8 ± 0.5	V6I, Q37K, K50Q, M84C, Q90H, T105I, A125T, T208M, T279S, D300N
HG3.17	700 ± 60	13.6 ± 0.8	V6I, Q37K, N47E, K50Q, G82A M84C, S89N, Q90F, T105I, A125T, T208M, W275A, R276F, T279S, D300N

Table 2. Catalytic parameters of the HG3 evolved variants. Activation free energies are calculated from the Eyring-Polanyi equation at a temperature of 298K. K_{cat} uncertainties are expressed as standard deviations. Free energy errors are determined by calculating the average error of the conversion of the maximum and minimum (i.e., k_{cat} ± SD) values to free energies. Sequence changes over the HG3 sequence background are indicated. Numbering is according to the original publication.[42]

A detail that stands out about the EVB activation free energies obtained for the HG3 variants (Figure 8) is their high standard deviations. This high variability comes from the different activation free energies of individual trajectories, each one exploring a finite portion of phase space representing possible ways the reaction could occur inside the enzymatic active site.

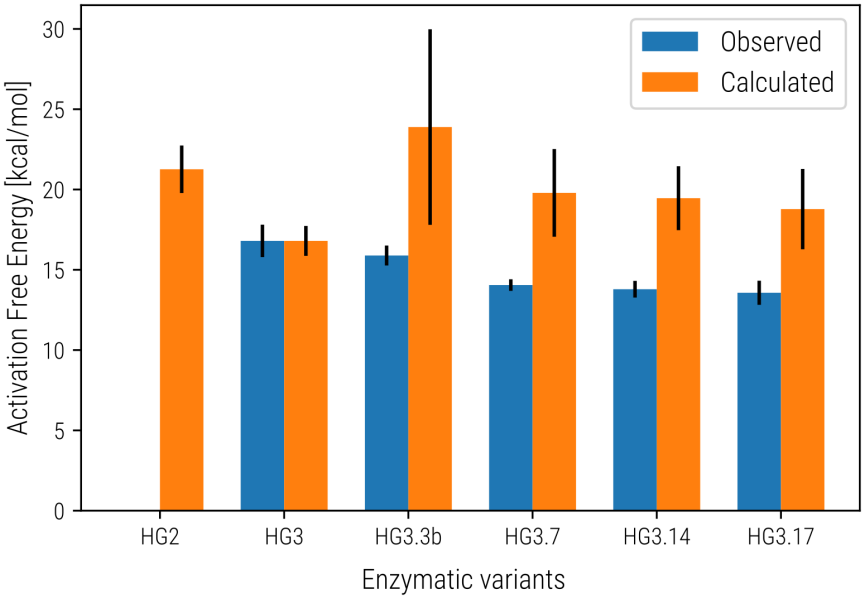

Figure 8. Observed vs. calculated free energies for the HG3 variants in the DE trajectory. Error bars correspond to those reported in Table 2.

Two questions arise about the origin of these high standard deviations. One concerns the simulation time required to obtain convergent results when carrying out EVB simulations. The second one is about the physical origin of the spread in the activation free energies. The first issue can be addressed by exploring more extended simulations (simulation length and number of replicas) to obtain the reaction free-energy profiles.[55] While this can show when a specific EVB setup can start to converge, it does not necessarily mean that simulations have converged over the complete conformational space of the system at the simulated temperature. Indeed, very long simulations show that enzymatic systems can adopt several conformations with different catalytic capabilities.[12] Since the process of showing convergence in MD simulations is computationally too costly, in this work, we have focused primarily on the latter issue, i.e., the physical origin of the large standard deviations in the EVB free energy profiles.

Examining the effect of conformational entropy in EVB simulations over the calculated activation free energies.

We set up additional EVB simulations using positional restraints over the protein coordinates for the GNCA4-W229D-F290W variant. These restraints allow the reaction to occur in a protein environment that stays fixed near the conformation that binds the TSA (i.e., the crystallographic

conformation). We compared the restrained simulations' reactive free energy profiles with the ones derived from the unrestrained simulations (Figure 9).

While the unrestrained simulations have a high standard deviation (1.70 kcal/mol), the restrained simulations have a much lower standard deviation (0.22 kcal/mol). This diminished standard deviation in the restrained simulations is because all reactive trajectories take place similarly, entailing similar energetics among them. More importantly, all the restrained-simulations KE trajectories involve very low activation free energies, as low as the lowest activation free energies sampled by the unrestrained simulations. The average activation free energy difference between these two simulations is 4.45 kcal/mol, implying a change in the reaction's velocity constant of a million fold. This significant energetical difference highlights the potential of pre-organising an enzymatic active site, even without any modification to its interaction potential (i.e., keeping its residue composition constant).

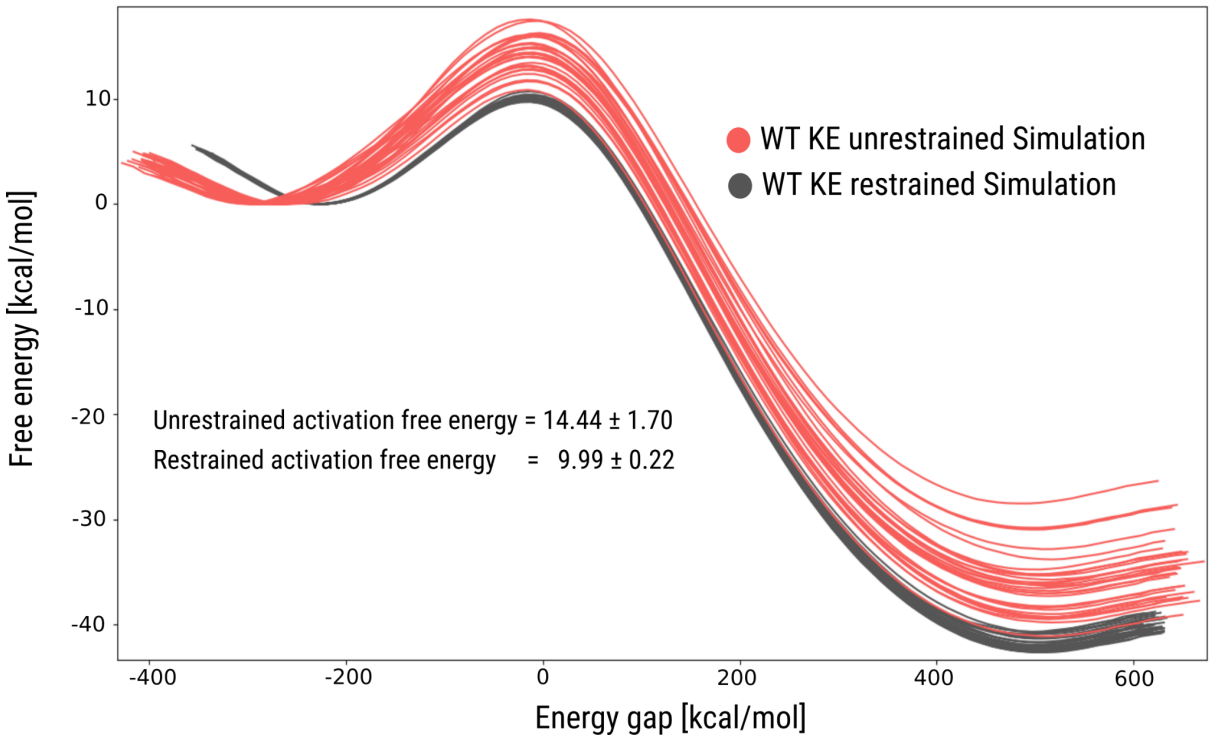

Figure 9. Effect of applying protein constraints over the EVB free energy reaction profiles. Free energy changes for the reactive trajectories of the KE reaction with (gray curves) and without (red curves) positional constraints over the protein atomic positions are shown along the reaction coordinate (energy gap).

Per-residue electrostatic energy contributions

When designing new enzymatic mutations, it is helpful to assert the individual energetic contribution of each residue in the active site. This can aid in identifying possible hotspots for catalytically-improving mutations.

We started by analyzing electrostatic free-energy contributions of individual residues for the (unrestrained) EVB simulations of the GNCA4-W229D-F290W variant, using the linear response approximation (LRA)[56] between the substrate and TS regions (Figure 10).

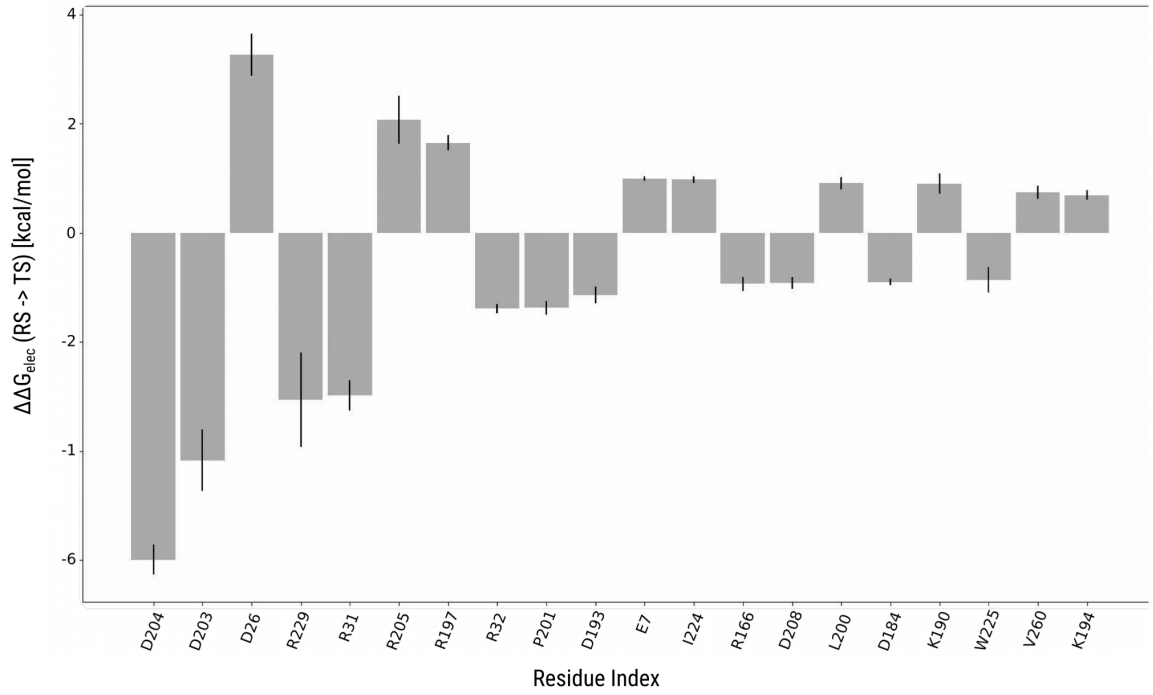

Figure 10. LRA residue-level electrostatic free energy contribution to the activation free energy of the GNCA4-W229D-F290W enzyme over the KE reaction. The error lines represent the standard deviation of applying the LRA analysis to all EVB trajectories.

In Figure 10, contributing residues are ordered by the magnitude of their per-residue electrostatic free energy contribution. The most contributing residue is the catalytic base (D204), followed closely by residue D203, just beside it (see Figure 11). On the other hand, residue D26 is an anti-catalytic residue, and it happens to be at the opposite side of residue D204 relative to the TS orientation along the bond being broken. It appears to be a pattern in the electrostatic free energy contribution for charged residues depending on their location relative to the TS ligand. At the TS's left side (Figure 11), positively charged residues contribute favourably to the catalytic energy,

while negatively charged residues contribute unfavourably. This effect is inversely mirrored at the other side, with positively charged residues at the TS's right side contributing unfavourably, and negatively charged ones do so favourably when at the TS's left side. This catalytic contribution makes sense in the context of the bond-breaking character developed at the KE reaction TS: there is a negative charge accumulation at the TS oxygen (oxyanion formation, pointing left in Figure 11), and a positive charge accumulation at the nitrogen of the bond being broken (pointing right in Figure 11). This charge polarisation suggests an electrostatic field mechanism responsible for the differential stabilisation of the TS regarding the ground state.

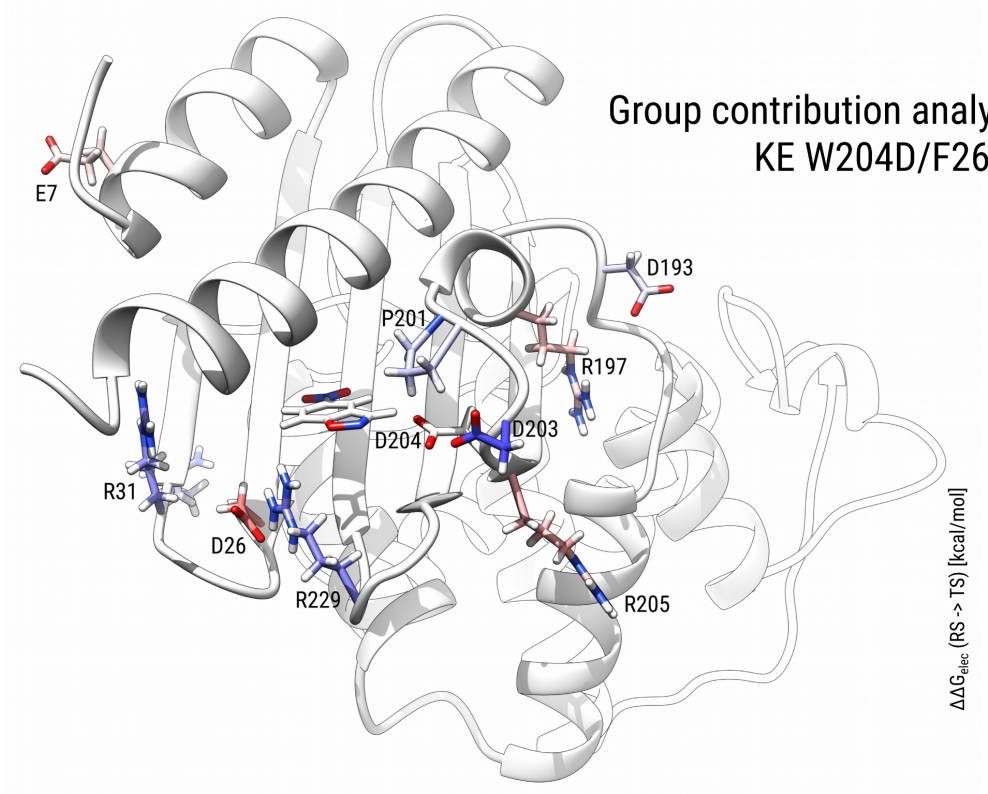

Figure 11. Structural depiction of the electrostatic free energy contribution of individual residues to the catalysis of the GNCA4-W229D-F290W variant.

Another possible practice to evaluate electrostatic contributions of different residues is running EVB simulations using an uncharged version of the residue in question (i.e., all the residue's atomic charges are set to zero). To explore the effect that removing a specific charge would have over the activation free energy of the reaction, we ran separate simulations in which the seven residues with the highest LRA free energy contribution magnitudes were individually modified for their uncharged versions. The activation free energies of the full simulations were plotted together with the LRA free energy contribution of the uncharged residue (Figure 12). If the LRA

free energy contribution is favourable for a particular residue (i.e., it has a negative $\Delta\Delta G_{elec}$ (RS -> TS) value), it would be expected that the predicted activation free energy would increase when uncharging this residue. The opposite should be valid for residues with an anti-catalytic LRA contribution (i.e., having a positive $\Delta\Delta G_{elec}$ (RS -> TS) value).

The previous idea was plotted in Figure 12, where it would be expected that residue contributions should be mapped into the first (for anti-catalytic residues) and fourth quadrant (for catalytic residues). However, it can be seen that either the uncharged simulation did not change the activation free energies significantly or did it so counterintuitively. The case of the R31 residue (red circled in Figure 12) is the most extreme, in which the activation free energy drops in approximately 1 kcal/mol when uncharging it.

We decided to look deeper into the counterintuitive result of the R31 uncharged simulation. We depict in Figure 13 several superimposed snapshots of the trajectories simulated with and without the R31 residue's charges. The conformations between both simulations differ quite significantly; in the charged version, the R31 side chain is solvent-exposed, while, in the uncharged simulation, it makes hydrophobic contacts with the sidechain of residue W263. This result raises concerns on the interpretability of uncharged simulations to study electrostatic free energy contributions since uncharging a large residue, such as R31, comes with the unintended effect of creating a large hydrophobic residue. This behaviour explains the previous counterintuitive effect of discharging the R31 catalytic residue; removing its charges produced an even lower kinetic barrier, probably by stabilizing the contribution of residue W263 or the active site packing.

These results align with the intuitive idea that removing catalytic charges will increase activation free energies and that removing anti-catalytic charges will diminish them. However, it would not be that simple for a real physical system since dynamical behaviours can be unpredictable, and residues with high degrees of freedom (e.g., solvent-exposed) could behave outside the scope of their native structural context to create counterintuitive results upon mutations. These results highlight the importance of running conformational sampling when evaluating (bio)chemical activities.

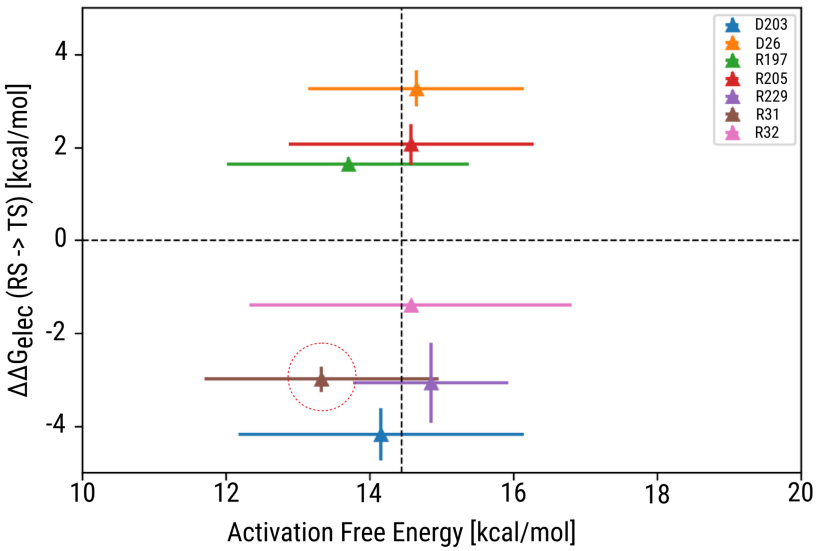

Figure 12. Unrestrained EVB uncharged-residue simulation analysis. Each point represents a simulation run uncharging the specified residue. The activation free energy, derived from this simulation, is plotted against the LRA free energy contribution of the specified uncharged residue. The vertical dashed line represents the average activation free energy of the fully charged simulation (i.e., with normal charges). The error bars represent the standard deviations of all the EVB replicas used in the analysis.

Comparison of residue electrostatic energy contributions for enzymatic designs

To explore the electrostatic catalytic effect developed by a computational design over the reaction coordinate, we repeated the LRA-per-residue electrostatic analysis with the catalytically improved design 5FQK_KET_0487_00523 and compared it with the one for the GNCA4-W229D-F290W original variant (Figure 15).

The catalytic base D204 and other residues: D27, D205, R198, R206, R33, W226, R167, and E8, maintain their catalytic activity in both variants. Residue R230 has a non-obvious impact over the catalysis of 5FQK_KET_0487_00523 by acting as a neutral catalytic position in some trajectories, while in others as a prominent anti catalytic position. This position behaved as a robust catalytic position in the original variant, acting as a stabiliser of the developed oxyanion hole in the TS (Figure 16).

There is a significant shift in the ligand position in the design variant relative to the original one. This shift creates a different contact angle between the ligand and the R230 side chain in the design variant, which can no longer act as an oxyanion hole stabiliser and becomes disorganised.

The shift in the ligand position is provoked by a strong interaction between the ligand's nitro group and a newly designed H237 position (Figure 16). As a positive residue, this interaction creates a catalytic effect unique to the designed variant, which was non-existent in the original design since a hydrophobic residue occupied this position.

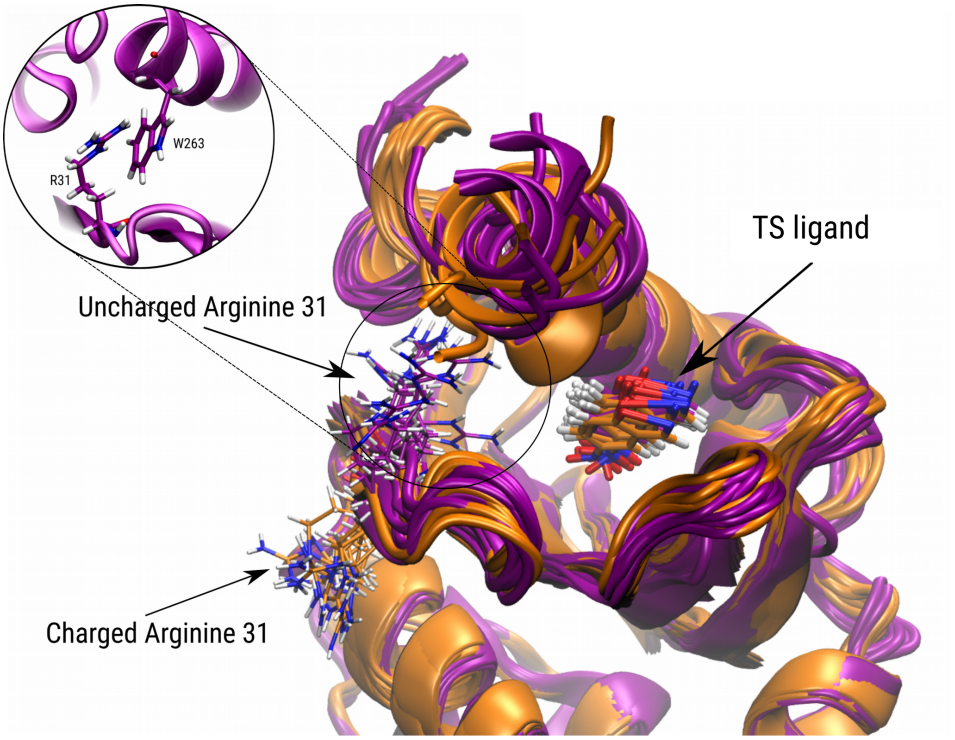

Figure 13. Comparison of conformations for the R31 residue when simulated with and without its atomic charges. EVB simulations' snapshots of R31 residue conformations are compared when the R31 residue contains full charges (orange) or is uncharged (purple). A characteristic snapshot of the uncharged simulation is shown (circled close up) to depict the interaction developed between R31 and W263 residues. Conformations of the KE reaction TS ligand are also shown for both simulations to indicate its relative position to R31.

Many other per-residue contributions are unique either to the designed variant or to the original one. In other cases, like R32, a better catalytic contribution is found in the original variant than in the designed one. This residue has a different rotameric position in both models (Figure 16), explaining the differential catalytic effect.

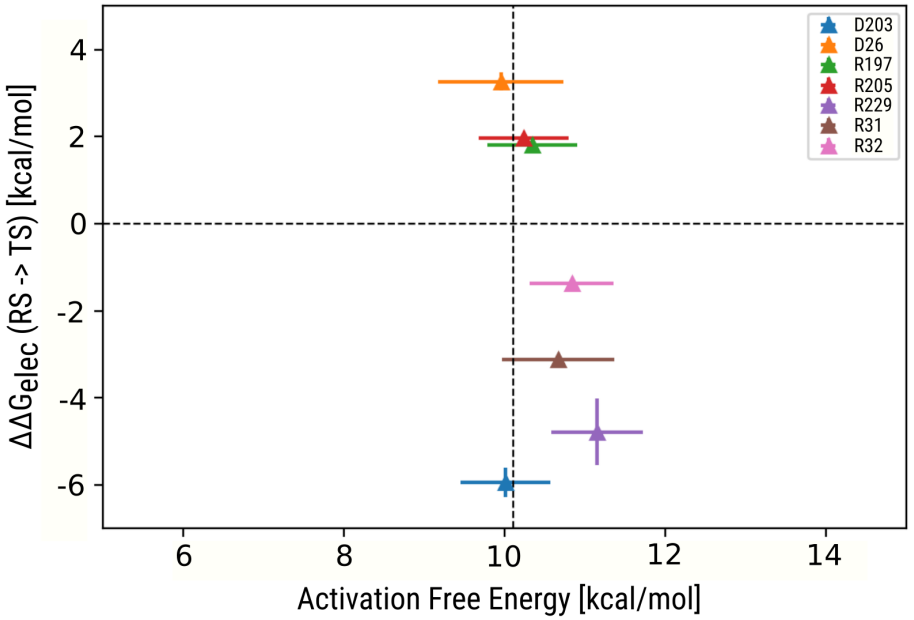

Figure 14. Restrained EVB uncharged-residue simulation analysis. Each point represents a simulation run uncharging the specified residue with positional restraints imposed over all protein atoms. The activation free energy derived from this simulation is plotted against the LRA free energy contribution of the specified uncharged residue. The vertical dashed line represents the average activation free energy of the fully charged simulation (i.e., with normal charges). The error bars represent the standard deviations of all the EVB replicas used in the analysis.

Discussion

Our target was to combine two successful methodologies to create an improved optimisation framework for computational enzyme design. On one side, the Rosetta suite of programs for macromolecular modelling has successfully generated enzyme design proposals that create *de novo* catalytic activity in previously inactive protein scaffolds. On the other hand, the EVB model has established the physical basis of protein catalysis by predicting relevant physicochemical parameters through detailed simulations of enzymatic systems. Combining the two methods seemed an obvious choice to create a robust framework for enzymatic design in which variants from the enzyme design algorithm are later evaluated with EVB simulations.

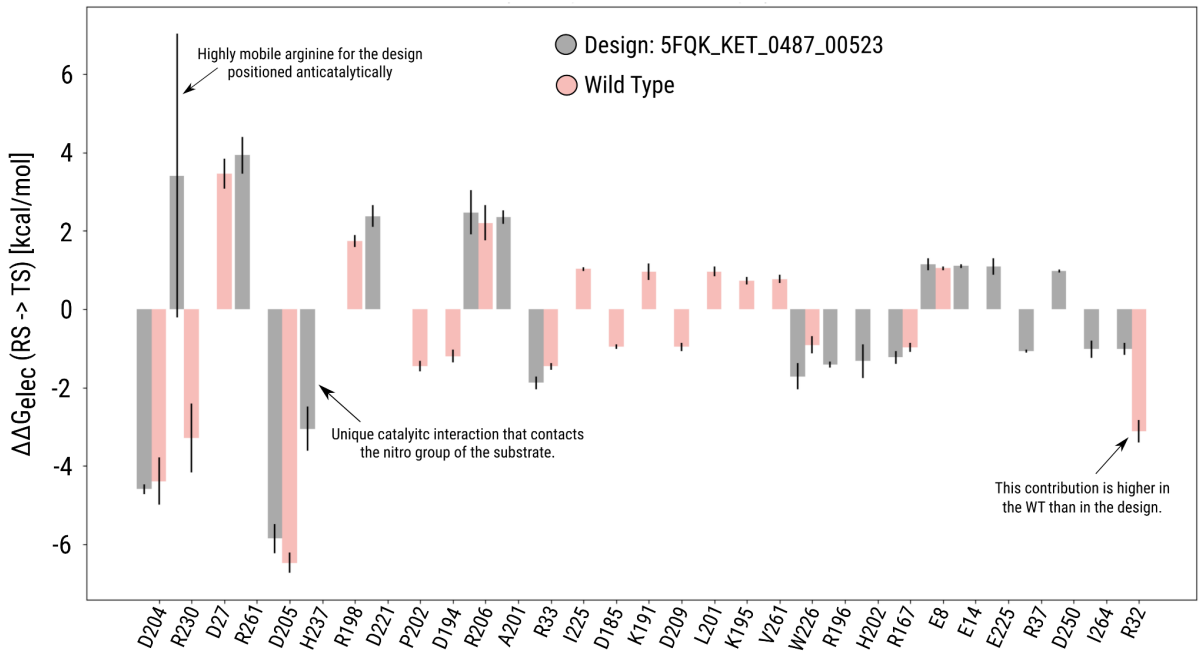

Figure 15. LRA residue-level electrostatic free energy contribution comparison between the starting and a designed enzyme. The LRA electrostatic analysis was done for the GNCA4-W229D-F290W variant (gray) and the 5FQK_KET_0487_00523 designed model (salmon).

We selected the case of redesigning a *de novo* active site of experimentally proven enzymatic activity to test our idea. This system is based on a resurrected ancient beta-lactamase scaffold and has a catalytic activity for the KE reaction comparable to other computational design strategies.[51] We explored two Rosetta design methodologies to generate optimised variants for this system based on optimising the Rosetta total score of the enzymatic complex. This target score allowed, on the one hand, to optimise interactions towards improving protein stability, and on the other, since mutations were only allowed to occur near the ligand position, to improve active site interactions with the TS model.

The EnzDes strategy, based on an almost fixed-backbone optimisation of the rotameric states of the active site, generated very few mutations, indicating that the active site was near its sequence optimum for its native backbone conformation. This result was not surprising since the Rosetta score function has been trained to recapitulate the native sequence of scaffolds based on rotameric optimisation in a fixed backbone protocol.[57] While this strategy could have previously worked for *de novo* design applications,[20] it does not help in redesigning efforts to improve enzymatic variants.

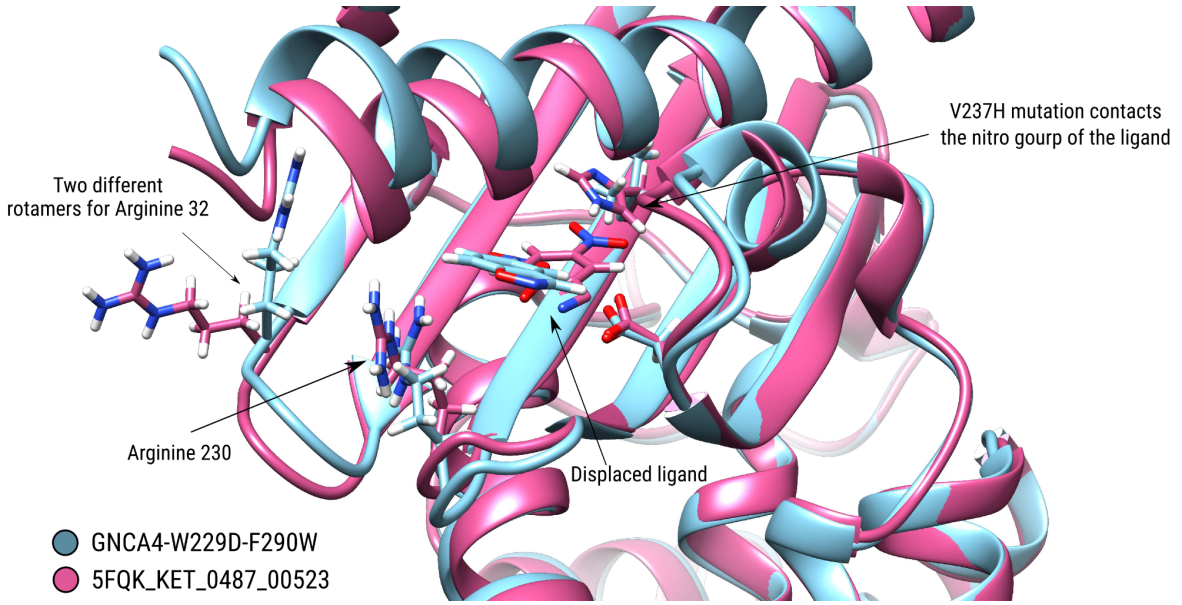

Figure 16. Comparison of residue positions with different catalytic effects between the designed (magenta) and the starting (cyan) enzymatic variants.

The FastDes methodology, on the other hand, allows complete flexibility of the backbone, ligand, and rotameric state. This strategy generated a diverse set of designs that encompassed a wide range of total and interface scores. This strategy is more suitable for enzyme redesign since it can diversify the possible ways the protein scaffold binds the TS state. The need for backbone flexibility is in line with other works in which it was essential for the success of protein design approaches.[58–60]

When we evaluated a ranked selection of Rosetta designs with EVB simulations, very few models were evaluated as improved regarding the original variant, establishing questions and challenges when combining both methodologies:

First, there is no certainty that the design algorithm can generate improved models since the optimisation target employed (i.e., the total score) does not explicitly capture catalytic activity. There is a need for a computationally cheap metric to capture catalytic trends for the target reaction that can be used as the optimisation target or in combination with the current one, either included as a constraint or in a multi-objective optimisation algorithm.

Our results indicate that the preorganisation of key catalytic residues could significantly impact the activation free energy since configurational entropy plays an important role in defining how the reaction can occur in the active site. We observed high variability in the activation free energies of different EVB trajectories, notably decreased upon restricting the protein movements to the conformation that binds the TSA. This effect was accompanied by a significant diminishing in the activation free energies of all individual EVB replicas, pointing to an improved catalytic effect based on residues being preorganised in a catalytically competent conformation throughout the whole enzymatic reaction. In this regard, metrics predicting individual residue dynamics could be helpful to capture preorganisation levels at key enzymatic positions and significantly improve optimisation strategies seeking to enhance active site preorganisation.

Second, due to the limitations of running MD simulations, not all designs can be directly tested by the EVB method. It is, therefore, crucial to have a ranking system of the designed variants to test only the ones most likely to succeed in improving the catalytic activity. As we mentioned before, most of these selections have been carried out by the chemical intuition of the designer,[26] rendering the success of the methodology ambiguously defined. A systematic, practical, and computationally cheap first screening should be applied to ensure that the more accurate and computationally intensive techniques (e.g., EVB simulations) only evaluate models with the most promising catalytic characteristics.

We explored the possibility of uncovering alternative TS binding modes by sampling the enzymatic system while removing the constraints employed during the design methodology. This new sampling was done since there was no guarantee that the structures derived from the design algorithm would be at their minimum energy conformations. Thus, we predicted TS binding free energies using Boltzmann distributions to integrate all the energetical information from the sampled conformations. However, this binding free energy score did not significantly correlate with the activation free energies derived from EVB simulations. This poor correlation could indicate a difference in the sampling coverage between both methods, a faulty score function for evaluating protein-ligand interactions[61] or, possibly, that TS binding does not necessarily correlate well with activation free energies for this system. Nonetheless, this binding free energy score correlated better with the EVB activation free energies than the interface scores derived directly from the designed structures alone and could, in principle, be applied to select

designs for further characterisation. Improving this correlation by refining the current method or developing new alternative ones would be key for optimising the computational enzyme design pipeline.

Third, predicting catalytic activities with EVB is in itself challenging, requiring suitable parameterisation and validation before being applied for computational design screening. A requirement for the validation step is the availability of a dataset large enough to establish statistical significance. These datasets should comprise variants with low to high catalytic activities for a specific reaction, hopefully paired with structural information for the binding of the reactive system (i.e., substrate, TSA, or product). There are relatively few datasets with these characteristics that are large enough to be statistically significant; however, experimental techniques for obtaining structural information from sequence alone,[62] and for the rapid and simultaneous evaluation of multiple variants are starting to catch up with these goals.[63]

Our validation of the EVB method was partially successful in reproducing trends in a small validation dataset containing evolved enzymes for the KE reaction, failing mainly to predict absolute activation free energies. While essential to validate the simulated results, failure in predicting absolute catalytic parameters would not impede the filtering out of designs, especially since designs will ultimately be tested experimentally. Establishing a ranking of the most promising models can still help increase the ratio of success of our evolutionary strategy. Nevertheless, the failure of our EVB implementation in predicting catalytic parameters needs to be addressed more thoroughly. Special attention needs to be paid to the method's convergence, which has been seldom addressed in the EVB literature.[55] Also since proteins can adopt multiple configurations differing in catalytic activity, running simulations only focusing on the minimum energy conformation could seriously disregard the full effect of the protein's configurational entropy over the predicted catalytic barriers.[12] This is particularly true for enzymatic designs whose sampled configurational space can differ significantly from the original scaffold. There is an all-important need to overcome this problem of configurational sampling that, in our specific case, given the large number of variants to be tested, should be as computationally cheap as possible.

Despite not being successful in uncovering improved designs, the application of the EVB method to assess computationally designed enzymes helped us reveal important hypotheses for improving the catalytic activity of the studied reaction:

The electrostatic contribution analysis showed that the location of charged residues followed a trend in agreement with the idea that electrostatic fields[11] can help to stabilise the differential charge developed at the TS due to its bond-breaking character. However, as we observe in many of our EVB simulations, the strategy cannot be simply applied by changing the charge of surface residues since their mobility can be hard to predict. Unexpected dynamics can occur given the alternative positions these residues can adopt relative to the location of the reactive ligand, specially for residues with low preorganisation (i.e., solvent exposed). Residue mobility predictions should be used to ensure that positioned charges maintain the correct polarisation throughout the full enzymatic dynamics.

Studying the system using positional constraints over the protein revealed the importance of configurational entropy over the catalysed reaction. Improving active site preorganisation helps more reactive trajectories to occur optimally, indicating that optimisation should not only focus on improving the interaction potential of the system but also on stabilising the preorganisation of key target residues. This fact is highly important for residues stabilising the bond-breaking polarisation developed at the TS of the KE reaction, at which correctly positioned charged residues are essential.

Finally, comparing the predicted catalytic effects of different residues in enzymatic variants is challenging. Subtleties in the ligand position and additive effects of multiple mutations over the reaction coordinate generate unique idiosyncrasies regarding how the catalytic effect is achieved. Nonetheless, these comparisons can inform and guide the improvement of catalytic activities by highlighting effects profited by specific variants, which could be later combined to improve the computational enzyme design strategy. Specifically, we observed the novel effect of a positive charge over the nitro group in the KE reactive system, whose catalytic effect demands a more in-depth exploration.

Conclusions

Our attempt to combine the Rosetta enzyme design methodology with EVB screening failed to uncover a significant number of improved variants according to the predicted absolute activation free energies. It could still be the case that improved variants exist among the set of designs since our application of the EVB method did not excel for absolute activation energy prediction and was better at ranking them. More thorough validation and parametrization of the EVB simulations can improve these predictions by encompassing convergence studies and assessing the effect of including alternative protein configurations when evaluating activation free energies.

We identified several key aspects to improve the Rosetta and EVB pipeline: the development of a cheap metric to capture residue preorganisation to include it in the optimisation algorithm, a rapid systematic assessment method to select the most promising variants for further and more intensive evaluation, and to improve the robustness of the EVB protocol to predict activation free energies. These aspects inspire the subsequent work carried out in this thesis.

Finally, it is of the utmost importance that the information derived while developing this framework for computational enzyme design be experimentally tested to corroborate and demonstrate the hypotheses derived from it, thus validating the overall value of the methodology.

Methods

Small-molecule Docking

TSA-ligand self-docking into the GNCA4-W229D variant was carried out using Rosetta scripts.[64] The target complex was downloaded from the PDB[65] with code 5FQJ. TSA parameters were obtained directly through scripts inside Rosetta for ligand preparation.[52]

The ligand docking protocol consists first of a random initial placement that searches the rigid body and torsional degrees of freedom of the ligand relative to the protein receptor. Then, small perturbations of the ligand and repacking the receptor side-chains. Finally, the entire system is minimized using gradient-based minimization, including the ligand, receptor side-chain, and backbone atoms. The first two steps (i.e., ligand positioning and side-chains repacking) are

optimised with a Monte Carlo algorithm using the Metropolis criterion. The entire protocol ligand docking protocol using Rosetta Scripts is described in [52].

Quantum chemical calculations

The geometry of the KE reaction TS model was calculated from QC modelling of the reaction at the B3LYP/6-31+G(d) level of theory, optimised with the CPCM implicit solvent model using a water dielectric. The TS was confirmed via an oscillatory mode frequency analysis, characterised by a single imaginary frequency connecting the product and substrate of the reaction when integrated through its intrinsic reaction coordinate. QC calculations were carried out with the Gaussian 09 program.[66]

Charges employed in all calculations for the KE substrate, product, TSA, and TS ligands were calculated with the RESP method using the RED tools program[67] with default options.

Enzyme design protocol

The enzyme design protocols were executed through the Rosetta Scripts platform.[64] The GNCA4-W229D-F290W variant starting model was downloaded from the PDB database[65] with code 5FQK. The TSA bound into this structure was replaced by a TS model of the KE reaction (see "Quantum chemical calculations" section in Methods), and the entire complex was minimised with Rosetta's Relax protocol[54] 10000 times. The lowest energy model produced was selected for applying the enzymatic design protocols.

All enzyme design protocols apply a single catalytic constraint during the optimisation, maintaining the hydrogen bond distance between the KE reaction TS and the catalytic base D229 at a proton transfer distance. Other degrees of freedom associated with this bonded distance, i.e., angles and torsions, were also constrained to maintain the proper TS state geometry.

The enzyme design protocols define which residues will be designed (change sequence), repacked (maintain sequence but optimised), or fixed (not included in the according to 4 cutoffs. The first cutoff is set for residues with their CA atoms within 6.0 angstroms from any ligand atom. All residues within this cutoff are allowed to change their amino acid identity (i.e., design) by repacking their side chains from a library containing all 20 natural amino acids, except for the

catalytic base, which was maintained as an aspartic acid residue. The second cutoff, at 8.0 angstroms, allows residues with their CA atom inside the cutoff but with their beta carbon closer than the CA to any ligand atom to be designable. The third cutoff, at 10.0 angstroms, sets residues with their CA atom inside it to be repackables (side chains are optimised without changing their amino acid identities). Finally, at 12.0 angstroms, the fourth cutoff defines residues containing the CA inside the cutoff, but with their beta carbon closer than the CA to any ligand's atom, to be repackable. Each cutoff is applied in order, and any residue that fulfils a cutoff is excluded from applying the following restrictions.

The EnzDes protocol works in two stages. The first stage is when residues are allowed to change their identity through a Monte Carlo search through the sequence space by rotamer optimisation, followed by a gradient-based minimisation that includes ligand rigid-body, side-chain and backbone degrees of freedom. The second stage follows the same steps as the first stage; however, no design (only repacking) occurs during the rotamers optimisation.

The FastDes protocol consists of the application of the Relax protocol[54] with design capabilities at the repacking stages. The Relax protocol searches the local conformational space of residues included in the optimisation, using cycles of repacking and minimisation that scale the repulsive term ('fa_rep') of the Rosetta score function.[47] The weights used are 0.02, 0.25, 0.55 and 1.00 of the original 'fa_rep' weight. During minimisation, ligand rigid-body, side chain, and backbone degrees of freedom movement are included in the optimisation scheme. The rotameric behaviour of residues during repacking are assigned according to the rules of the four cutoffs defined for designability (see above).

Local conformational search of designed enzyme

A sampling of the conformational space of enzyme design models was applied to characterise their binding modes. The protocol consisted of an unconstrained (i.e., catalytic constraints were removed) search of the local conformational space employing the Rosetta Relax protocol to the whole system. Each model was sampled with 200 trajectories, and the total and interface scores were used to calculate their binding free energy contributions.

The binding free energies of designed models were calculated from a Boltzmann distribution based on the total scores of sampled conformations. The binding energy values were calculated as the expectation value of the interface score between the transition state and the enzyme, using a Boltzmann distribution based on the total complex energy:

$$< E^b >= \sum_i^N p_i E_i^b \tag{1}$$

Here, $< E^b >$ is the expectation value of the interface score, N is the total number of sampled conformations, and E_i^b is the interface score of the *i*th conformation. The interface score for each enzyme design conformation is calculated as:

$$E_i^b = E_i^{complex} - (E_i^{TS} + E_i^{Enz}) \tag{2}$$

With $E_i^{complex}$ is the total energy of the enzyme-TS complex structure, E_i^{TS} the unbound TS ligand energy, and E_i^{Enz} the unbound enzyme energy. Probabilities (p_i) are obtained from a Boltzmann distribution using the N sampled enzyme-TS conformations $E_i^{complex}$ scores as:

$$P_i = \frac{e^{-E_i^{complex}/KT}}{Q} \tag{3}$$

Here, KT is the characteristic energy partition, and *Q* represents the partition function calculated as:

$$Q = \sum_i^N e^{-E_i/KT} \tag{4}$$

EVB analysis

The EVB model, used for studying chemical reactions in enzymes and solution, begins by describing a chemical reaction using a valence bond approach. In the EVB, the system

wave-function is represented by a linear combination of the most essential ionic and covalent resonance forms (diabatic states) of the system.[68]

Reactant and product are treated as basis states that are mixed to describe the reacting system. The potential energies of the diabatic states (H_{ii} and H_{jj}) and the mixing term (H_{ij}) are represented by the Hamiltonian matrix elements in equations:

$$H_{ii} = \epsilon_i = \alpha^i_{gas} + U^i_{intra}(R,Q) + U^i_{inter}(R,Q,r,q) + U^i_{solvent}(r,q) \quad (5)$$

$$H_{ij} = Ae^{(-a|\Delta R'|)} \quad (6)$$

R and Q represent the atomic coordinates and charges of the reactants or products (i.e., the solute) in the diabatic states, and r and q are the coordinates and charges of the surrounding water or protein (i.e., the solvent). α^i_{gas} is the energy of the i^{th} diabatic state in the gas phase, where all the fragments are taken to be infinity. $U^i_{intra}(R, Q)$ is the intramolecular potential of the solute system (relative to its minimum) in this state. $U^i_{inter}(R, Q, r, q)$ represents the interaction between the solute atoms and the surrounding solvent atoms. $U^i_{solvent}$ represents the potential energy of the solvent. The adiabatic ground state energy E_g and the corresponding eigenvector C_g are obtained by solving the secular equation:

$$H_{EVB}C_g = E_gC_g \quad (7)$$

Using the Hellmann-Feynman theorem[69] for obtaining the first analytical derivatives of E_g, the EVB energy surface can be sampled directly by MD simulations. However, in practice, this is done by a combined free energy perturbation and umbrella sampling (FEP/US) procedure[70,71] that provides the free energy function needed to calculate the activation free energy.

The free energy associated with the transformation of a molecular system from state i to another state j, described by the potentials V_i and V_j, respectively, can be calculated using the perturbation formula represented by the equation:

$$\Delta G_{i\rightarrow j} = -RTln\langle e^{\frac{-(V_j-V_i)}{RT}}\rangle_i \quad (8)$$

The angular brackets denote the average ensemble calculated with MD simulations, using the potential V_i. Although the perturbation formula is exact, it is only applicable when the states are so similar that configurational sampling using V_i also samples relevant (i.e. low-energy) configurations on V_j. By introducing a set of intermediate mapping potentials (representing unphysical states) as linear combinations of V_i and V_j, sufficient sampling can be attained, provided that these intermediate potentials are sufficiently closely spaced. This sampling is carried out given that the following equation is satisfied:

$$V_m = \sum_{i=1}^{N} (1 - \lambda_m) V_i + \lambda_m V_j \tag{9}$$

Here, V_m is the effective mapping potential formed as a linear combination over all N states using the consecutive mapping $\lambda_m = (\lambda_1, \lambda_2, \ldots, \lambda_{N-1}, \lambda_N)$. The total free energy change is then calculated as a sum over all steps between the end-point potentials, as shown in equation:

$$\Delta G_{initial \rightarrow final} = \sum_{i=1}^{N-1} \Delta G_{i \rightarrow i+1} \tag{10}$$

Where N is the total number of steps (mapping potentials), and the above equation gives each sum term.

Finally, and considering only the case of two diabatic states, the free energy functional that corresponds to the adiabatic ground state surface, E_g is obtained by:

$$\Delta G(x') = \Delta G_m - \beta^{-1} ln \langle \delta(x - x') e^{-\beta[E_g(x) - \epsilon_m(x)]} \rangle_m \tag{11}$$

Where ε_m is the mapping potential that keeps the reaction coordinate x in the region of x', The angular brackets denote the average ensemble calculated with MD simulations using the ε_m potential, $\beta=(k_B T)^{-1}$, with k_B as the Boltzmann's constant and T the temperature. If the changes in ε_m are sufficiently gradual, the free energy functional $\Delta G(x')$, obtained with several values of m

overlaps over a range of *x'* values, gives the complete free-energy curve for the reaction when the full set of *ΔG(x')* are patched together.

Electrostatic free energy contributions at the residue level were calculated using the LRA approach [56] between substrate and TS state regions. The evaluated configurations are taken from simulations employing the specific mapping potentials (ε_m) that most contribute to these states. The residue level electrostatic energies were calculated with the Qcalc6 routine of the Q MD package.[72]

EVB simulations

We set up EVB simulations using as starting point structures extracted directly from the PDB database,[65] or derived from the different enzyme design protocols, specifically the models for the HG3-related enzymes, with IDs: 3NYD (HG3) and 4BS0 (HG3.17). The remaining models (i.e., HG2, HG3, HG3.3b, HG3.7, HG3.14) were generated through mutation and minimization with the Relax protocol.[54] The lowest energy model among 10000 produced trajectories, with bound KE reaction TS model, was selected for EVB simulations.

For all simulations, the AMBER14SB forcefield[73] was employed to represent protein solvent molecules. Bonded and nonbonded parameters, except for charges, were derived from the GAFF force field.[74] Charges for the solute substrate and TS system were calculated using the RESP fitting method.[75] Product charges were defined to match the TS charges at the TS region (i.e., when $\lambda_m = 0.5$, see "EVB analysis" section above) with the following formula:

$$Q_i^P = 2Q_i^{TS} - Q_i^S \tag{12}$$

Where Q_i^P, Q_i^{TS}, Q_i^S represent the charges of the i^{th} atom in the product, transition, and substrate state, respectively.

The solution reference reaction for the KE reaction was set up in a 25 angstroms radius sphere of water centred at the solute system. Water molecules were represented with the TIP3P water model.[76] The sphere integrity was maintained with the SCAAS method.[77] For each replica, the system was slowly heated up to 298K for a total of 120 ps applying a positional constraint of 100

kcal·mol^{-1} Å^{-2} in all solute atoms. Later, the system was equilibrated during 1.5 ns gradually removing constraints during the first 100 ps. Each replica was further simulated during 50 FEP windows (mapping potentials) of 50 ps each (i.e., 2.5 ns per replica). Energies, coming from 100 replicas (i.e., a total of 250 ns), were collected during this last step of the simulation to obtain the EVB activation free energy profiles of each variant simulated.

Enzymatic simulations for the KE reaction were set up by building a sphere of water molecules with a radius of 25 angstroms centred on the ligand. Any charged protein residue outside the sphere was modelled as its neutrally-charged version to avoid unwanted electrostatic effects over the solute system due to unscreened charges in vacuo. The water sphere integrity was maintained with the SCAAS method,[77] and any protein atom outside the simulated sphere is positionally restrained with a harmonic force constant of 200 kcal mol^{-1} Å^{-2}. For each replica, the system was slowly heated up during 300 ps with soft positional constraints over protein atoms (10 kcal mol^{-1} Å^{-2}) and hard constraints over solute atoms (100 kcal mol^{-1} Å^{-2}). The system is later equilibrated during 1.5 ns, gradually removing the constraints over protein and solute atoms during the first 100 ps but maintaining a half harmonic restraint (0.5 kcal mol^{-1} Å^{-2}) between the geometric centres of the catalytic base and the KE solute. This soft distance restraint is applied to ensure the solute is near the catalytic base side chain to engage it in the reaction. Finally, the system is simulated during 50 FEP windows (mapping potentials) of 50 ps each (i.e., 2.5 ns per replica). Energies, coming from 100 replicas (i.e., a total of 250 ns), were collected during this last step of the simulation to obtain the enzymatic EVB activation free energy profiles.

Chapter 2. Predicting binding free energy in MHC-I-peptide complexes

An essential requirement to explore methods for predicting free energies is using appropriate experimental datasets. For enzyme design, very few datasets relate experimental activity with structural information for the system; while many trajectories for the DE of enzymes pair the sequence of the improved variants with their catalytic activity, few have a significant number of solved structures for the relevant points along the evolutionary trajectories. This lack of a statistically significant dataset hinder efforts that correlate structural modelling data with experimental activity.

While the above issue is specific for enzymatic design, some protein systems contain appropriate datasets to predict free energy changes in biochemical processes. In this regard, protein-protein and protein-small molecule interaction datasets have had an enormous relevance in validating methodologies to predict changes in binding free energies.[78–83] Although predicting activation free energies is in principle different from predicting binding free energies, most methods in enzyme design employ almost the same principles as when approaching binding studies. Many are based on the binding of a fixed-TS model modelled with force fields typically employed to study systems in their ground state. Moreover, there are approaches for computational enzyme design directly employing the binding of substrates or products to create active catalytic sites.[84] Also, many simulation methods that query catalytic activities using simulations focus solely on the Michaelis complex to study the conformational dynamics of enzymatic systems.[12,85] Therefore, validating methods to predict binding free energies in protein-protein or protein-small molecule systems can still be highly relevant for computational design endeavours.

A popular benchmark system for protein-protein interactions is the Major histocompatibility complex class I (MHC-I) receptors. This receptor is a central piece in defining the repertoire of antigenic peptides that can activate the cytotoxic response of CD8+ lymphocytes.[86] A ternary complex between the MHC-I receptor, the antigenic peptide, and the T-cell receptor of the CD8+ lymphocyte must form for proper activation of a T-cell-CD8+ clone.[87] The MHC-I receptor binds peptides between 8 to 11 residues in length with a highly promiscuous specificity. The binding mode of peptides to MHC-I receptors is highly characteristic, with the peptide lying extended in

the MHC-I pocket with their termini frequently anchored at the same MHC-I receptor regions (see Figure 17).

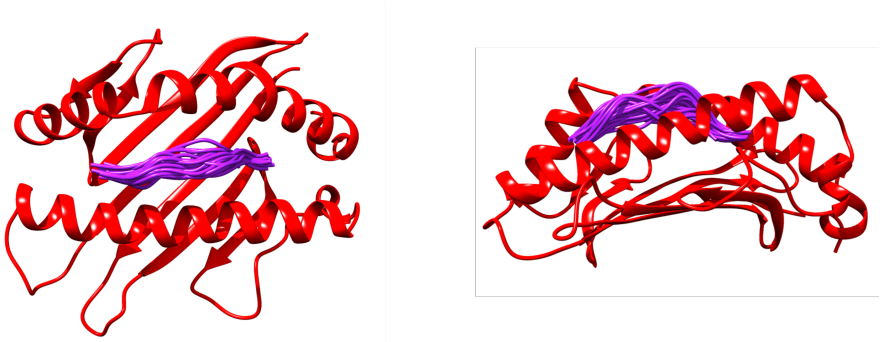

Figure 17. HLA structural set of peptide conformations. Fifty peptide conformations (purple) were derived from clustering a set of 534 PDB structures only containing the three human leukocyte antigens (HLA) MHC-I groups (HLA-A, HLA-B and HLA-C) bound to different peptide antigens. For simplicity, only the binding pocket domain (red) of the HLA-A receptor is shown.

The MHC-I-peptide complexes have several characteristics that make them ideal for exploring methods to predict binding free energies in systems dominated by protein-protein interactions. On the one hand, there is numerous experimental and structural information available for the system that relates the MHC-I receptor-peptide (MHC_p) complex structures and their immunogenic activities, which is directly correlated to the binding activity of the peptides for their MHC receptors.[88] On the other hand, the characteristic peptide binding to MHC receptors makes these systems simpler to explore since conformations tend to be highly similar for all bound peptides (see Figure 17). Finally, any developed method displaying good predicting performance for MHC_p binding can be used in the design of peptide-based vaccines.[89,90]

In this chapter, we have used MHC-I binding data to query the Rosetta Score function to predict peptides' binding activity towards a specific MHC-I allotype. This validation is relevant for a proof-of-concept demonstration that binding free energies can be predicted employing this knowledge-base scorefunction.

Building a compelling MHC-I peptide experimental dataset to predict binding free energies

The dataset was compiled from a single publication employing the same methodology to estimate all immunogenic activities.[91] This way of obtaining experimental data ensures that the

activity values are self-consistent, avoiding excessive experimental noise that could interfere with correlation predictions. The selected experimental dataset pertains to the HLA-A*02:01 allotype and encompasses diverse peptide sequences and binding activities (measured as IC50 values).

We filtered all peptide activity data to contain only nine-mers and group them using a hierarchical clustering strategy based on their pairwise sequence similarities (see "The PyCBBL library" in Appendix 2 - Code for more details). A threshold for the sequence similarity was defined to partition the group of sequences into 50 clusters. Only each cluster's centroid sequence was selected to compile the dataset, thus maximizing the diversity of peptidic sequences considered. The distribution of IC50 values of the final selected peptides is diverse and spans a wide range of experimental binding values (Figure 18).

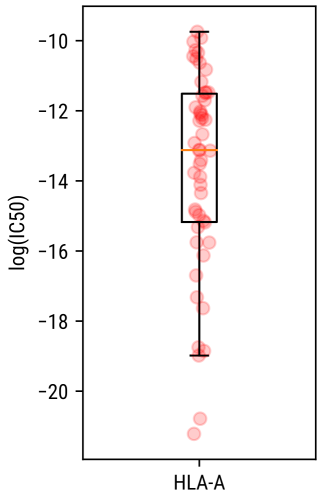

Figure 18. Distribution of log(IC50) values for the HLA-A allotype HLA-A*02:01 contained in the selected dataset.

Modelling MHC-peptide complexes

We aim to predict the binding free energies for MHC_p complexes for a dataset composed of diverse peptide sequences and a wide range of experimental activities (from now on, the binding dataset). The approach consisted of sampling an ensemble of conformations from a discrete set of experimental peptide conformations extracted from the PDB database (from now on, the structural dataset).

Each MHC_p complex in the binding dataset was modelled by setting 50 trajectories, each using a different peptide conformation from the structural dataset. The corresponding peptide

sequences were first threaded into each peptide backbone conformation, and then the complexes were minimized by searching their local conformational space (for more details, see Methods). For each MHC_p complex structure, we ran 30 cycles of the fastrelax protocol[54] by allowing the peptide and the neighbouring receptor residues to sample backbone, side chain, and rigid body degrees of freedom. The progression of the scores for all replicas modelled is shown in Figure 19, where it can be seen that total and interface scores converged early in the search, at less than five fastrelax cycles.

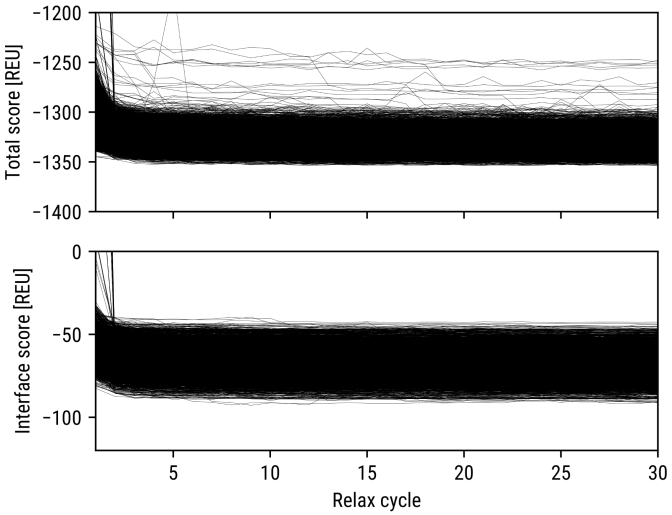

Figure 19. Convergence of the Interface and total score along the fastrelax optimisation. Each modelled trajectory is plotted individually. For clarity, the y-axis was truncated to depict only values close to the final converged scores.

Binding free energies correlations with the experimental data

The total and interface scores distributions for all sampled conformations of each peptide in the dataset are plotted in Figure 20. Peptides were ordered by their log(IC50) values to aid in the visualization of trends between the scores and the experimental activity values. The minimum total score values sampled by each peptide (Figure 20 upper plot) correlate poorly with the log(IC50) experimental values, with a Pearson Correlation Coefficient (PCC) of 0.4412. On the other hand, when correlating the lowest sampled interface scores (Figure 20 lower plot), the PCC increases to 0.7586. From the boxplot of interface score values distributions (Figure 20 lower plot), it is apparent that the average values of the distributions have a higher correlation with the experimental data, with an actual PCC of 0.8053.

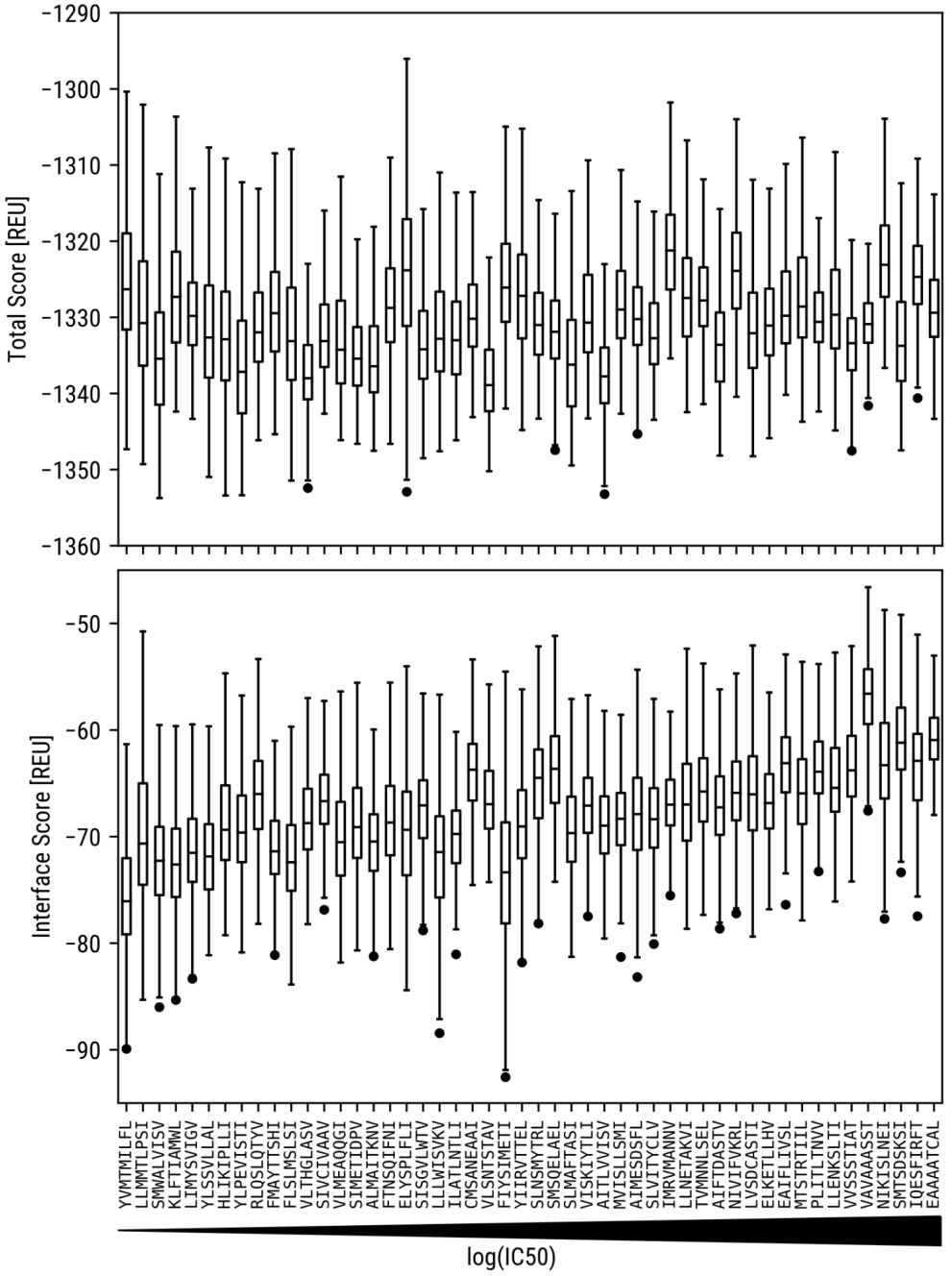

Figure 20. Distribution of total (up) and interface (down) scores for each peptide in the binding database. Each distribution summarises all values for each cycle step and all trajectories. Their log(IC50) values order peptide sequences displayed in the x-axis. If outliers exist in a distribution, only the lowest-score outlier is shown.

The fact that the interface score averages correlate better than the minimum sampled interface score values indicates that the ensemble of conformations can be more informative of the experimental activities than the lowest interface scores alone. This result can be rationalized by looking at the bidimensional distributions of total and interface scores (i.e., the conformational

energy landscapes, Figure 22). We observe that the lowest interface scores rarely coincide with the system's lowest-energy conformations. Because of their higher energy, these lowest-interface score conformations should be less populated than the system's lowest energy microstates. Therefore, a more proper way of counting the contribution of each conformation to the binding energy would be through applying a proper statistical-mechanics distribution.

We calculated the binding free energies of each MHC_p complex by setting a Boltzmann distribution for all sampled conformations based on their total score values. The calculated probabilities allowed us to estimate the binding free energy as the expectation value of the entire distribution of sampled conformations (see Methods section for details on these equations). However, the energy-partition parameter of this distribution (i.e., the KT parameter) has not been adequately characterised for the knowledge-based potential employed here.[47] Thus, we decided to explore the PCC as a function of this parameter (Figure 21).

The PCC increases readily, reaching a maximum of 0.8185 at 11.0 Rosetta Energy Units (REU), after which it drops very slowly at larger KT values. To explore the origin of this behaviour, we decided to repeat the analysis by dividing the peptides into two equally-sized groups based on their hydrophobicities. We estimated theoretical LogP values for each peptide, and the groups were defined based on the LogP median value. According to the estimated LogP values, most peptides have a hydrophobic nature (Figure 21, right plot). Despite this, we designated the 50% less hydrophobic peptides as the Hydrophilic group and the other more hydrophobic half as the Hydrophobic group. The Hydrophilic group presents a behaviour similar to the one when all the peptides were analyzed. The PCC peaks notoriously at a KT value of 6.6 REU to a PCC of 0.8185 and then decreases appreciably at larger KT values. On the other hand, the Hydrophilic group has significantly lower PCC values than the Hydrophilic group and presents a monotonous, although very slight, increase of the PCC at increased KT values, converging to a PCC of 0.6841.

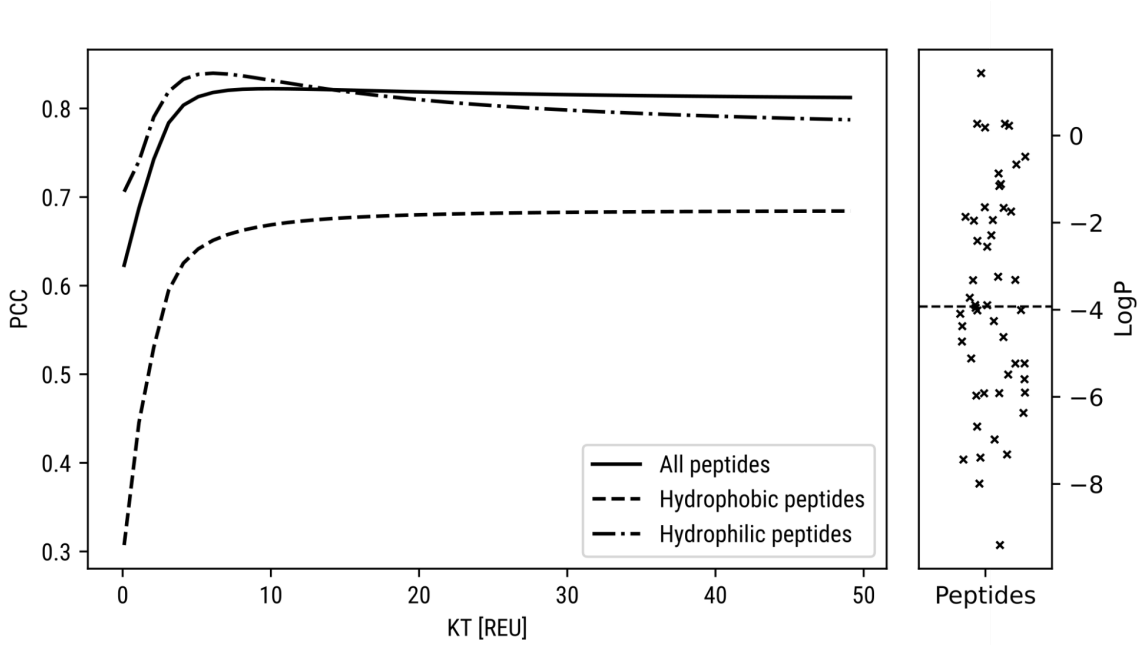

Figure 21. PCC at different KT values. (left) PCC variation is shown for three sub-groups of peptides in the dataset. The full subset of peptides (All peptides) was subdivided into two groups if they were above (hydrophilic) or below (hydrophobic) the median value of their predicted LogP values. (Right) Predicted LogP distributions for peptides in the binding dataset. The median LogP value is indicated with a dashed line.

To explore more in-depth the previous result, we performed a leave-one-out analysis to observe the effect of each energy term over the PCC (the Rosetta score function is defined as a linear combination of 19 terms).[47] In Figure 23, we plot the four terms with the highest effect over the calculated PCC values. Leaving out the 'fa_sol" or 'fa_atr' term significantly decreases the correlation with the experimental data; leaving the 'fa_atr' term generates a curve with a pronounced maximum PCC of 0.2255 at 0.8 KT; similarly, when the 'fa_sol' is left out, the correlation PCC maximum is less notorious and has a value of 0.1638 at 4.2 KT. The other two terms, 'fa_elec' and 'hbond_bb_sc", barely affect the PCC, behaving similarly to the full score function with PCC maxima of 0.7922 and 0.7933, respectively.

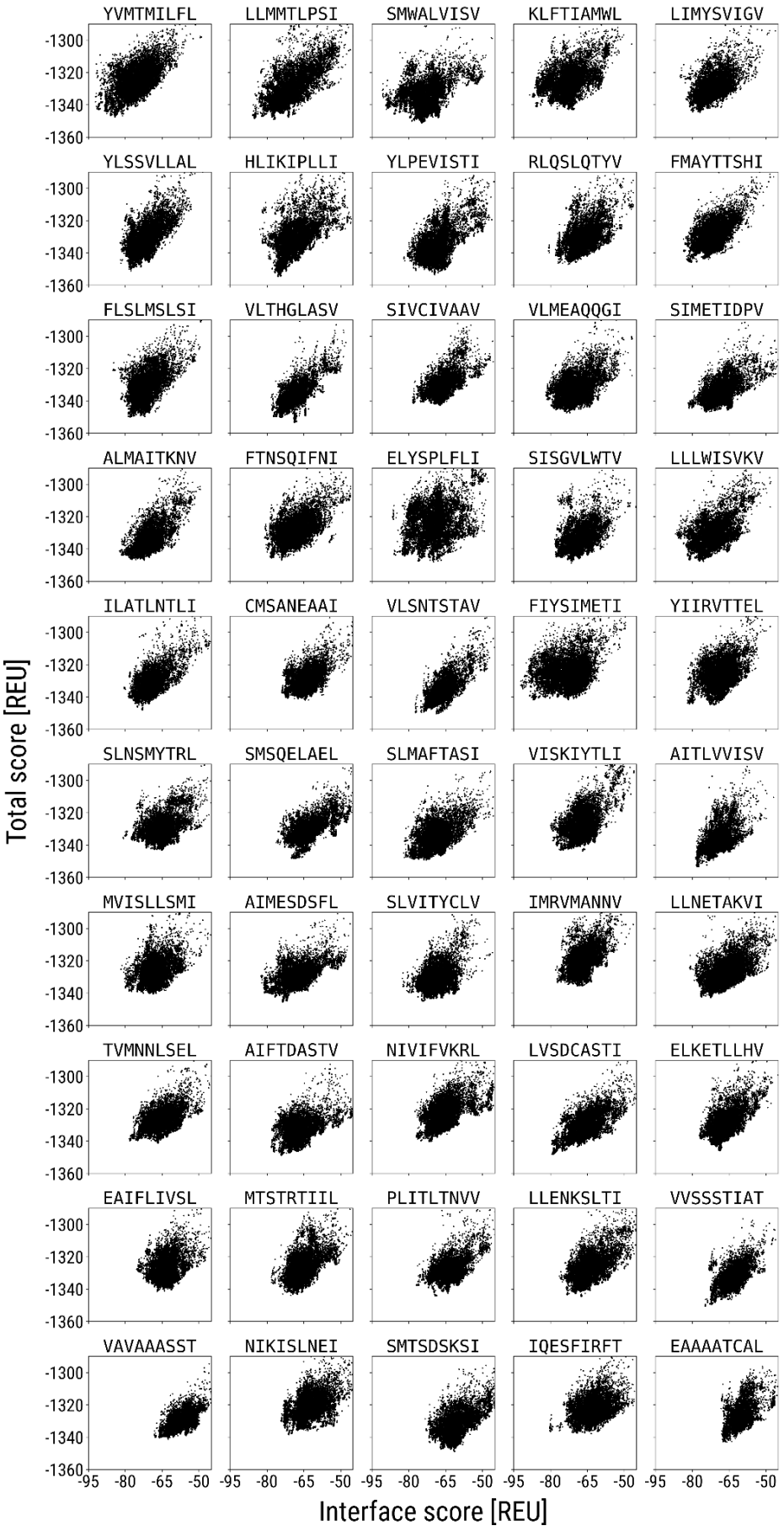
YVMTMILFL
LLMMTLPSI
SMWALVISV
KLFTIAMWL
LIMYSVIGV
YLSSVLLAL
HLIKIPLLI
YLPEVISTI
RLQSLQTYV
FMAYTTSHI
FLSLMSLSI
VLTHGLASV
SIVCIVAAV
VLMEAQQGI
SIMETIDPV
ALMAITKNV
FTNSQIFNI
ELYSPLFLI
SISGVLWTV
LLLWISVKV
ILATLNTLI
CMSANEAAI
VLSNTSTAV
FIYSIMETI
YIIRVTTEL
SLNSMYTRL
SMSQELAEL
SLMAFTASI
VISKIYTLI
AITLVVISV
MVISLLSMI
AIMESDSFL
SLVITYCLV
IMRVMANNV
LLNETAKVI
TVMNNLSEL
AIFTDASTV
NIVIFVKRL
LVSDCASTI
ELKETLLHV
EAIFLIVSL
MTSTRTIIL
PLITLTNVV
LLENKSLTI
VVSSSTIAT
VAVAAASST
NIKISLNEI
SMTSDSKSI
IQESFIRFT
EAAAATCAL
-1300
-1320
-1340
-1360
-95
-80
-65
-50
Total score [REU]
Interface score [REU]

Figure 22. (previous page) Total and interface energy landscapes for all peptides in the binding set. The total and interface score values for all sampled conformations are plotted as black dots. Plots are sorted by the peptides' log(IC50) values.

We repeated the PCC analysis only including individual energy terms to observe their behaviour (Figure 24). As expected, the 'fa_atr' and 'fa_sol' terms show good correlation curves when contrasted against the experimental log(IC50) values. The 'fa_atr' has a slight maximum PCC of 0.5596 at 4.1 KT, while the 'fa_sol' has a more notorious maximum PCC of 0.6078 at 1.1 KT. This behaviour agrees with what happens in the leave-one-out analysis of energy terms (see Figure 23). When both terms are combined, the PCC curve is close to the "Full score" curve (Figure 24), with a maximum PCC of 0.7783 at 1.1 KT. On the other hand, the 'fa_elec' component behaves oppositely to the experimental binding values, with a negative asymptotic PCC behaviour reaching a minimum PCC of -0.4411.

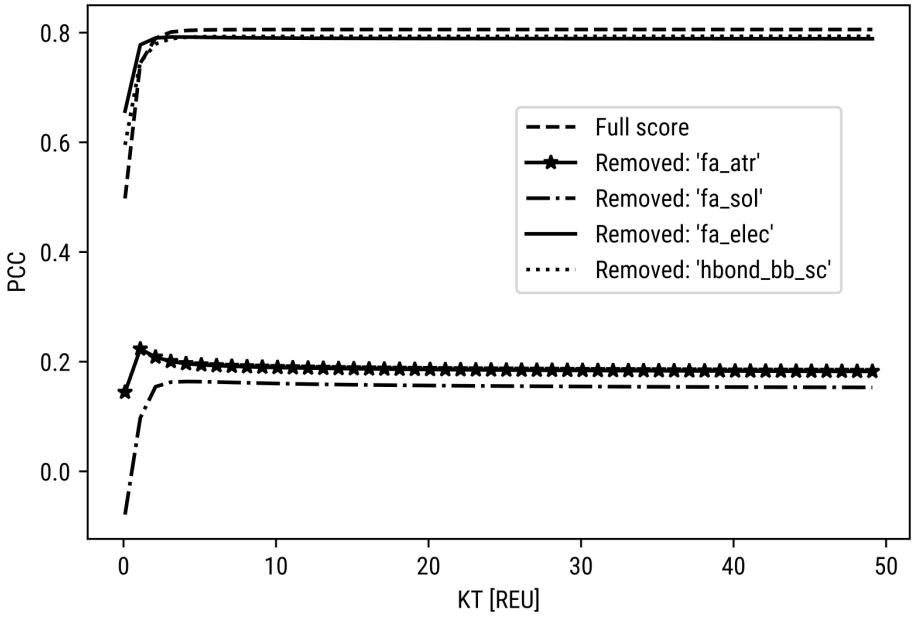

Figure 23. Leaving-one-out analysis of the score function energy terms. The curves show the PCC dependency on KT when using all energy terms (Full score curve) or leaving individual terms outside the interface score calculation (Removed curves).

Simulation convergence

We next consider the convergence of our sampling method according to specific sampling parameters. As we showed above, already at five or fewer relax cycles, low values were reached

for both interface and total scores (see Figure 19) while still retaining shallow variability at a higher number of relaxing cycles. Inspection of the PCC convergence concerning the same parameter shows that the curve has almost converged at around four relax cycles (Figure 25). Jointly, this shows that the convergence of the minimization of the HLA-A*02:01-peptide complexes using our strategy was fully achieved early regarding the number of minimization relax cycles applied.

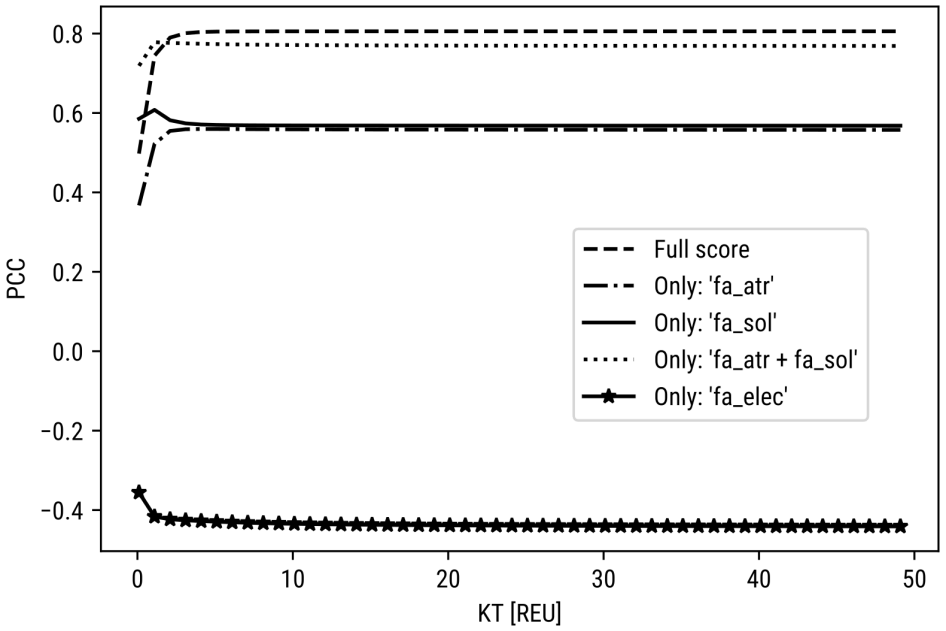

Figure 24. Effect of Individual Rosetta score function energy terms over the PCC. The curves show the PCC dependency on KT when using all energy terms (Full score curve) or only including individual terms (Only curves) to calculate the interface scores. Also, the effect of combining the two most important terms (Only: 'fa_sol + fa_atr' curve) is shown.

Our sampling method starts from a predefined and representative set of 50 peptide backbone structures. Since our protocol only generates a local search around these backbone conformations, we evaluated the convergence of the PCC as a function of the number of starting backbone conformations. This analysis was carried out by bootstrapping trajectories coming from specific starting backbone conformations. The PCC was calculated by repeating the bootstrapping at a progressively increased number of starting structures (Figure 26). The PCC converges when around half of the representative starting backbone conformations are being used, with good correlations even when only 5 to 10 conformations are employed.

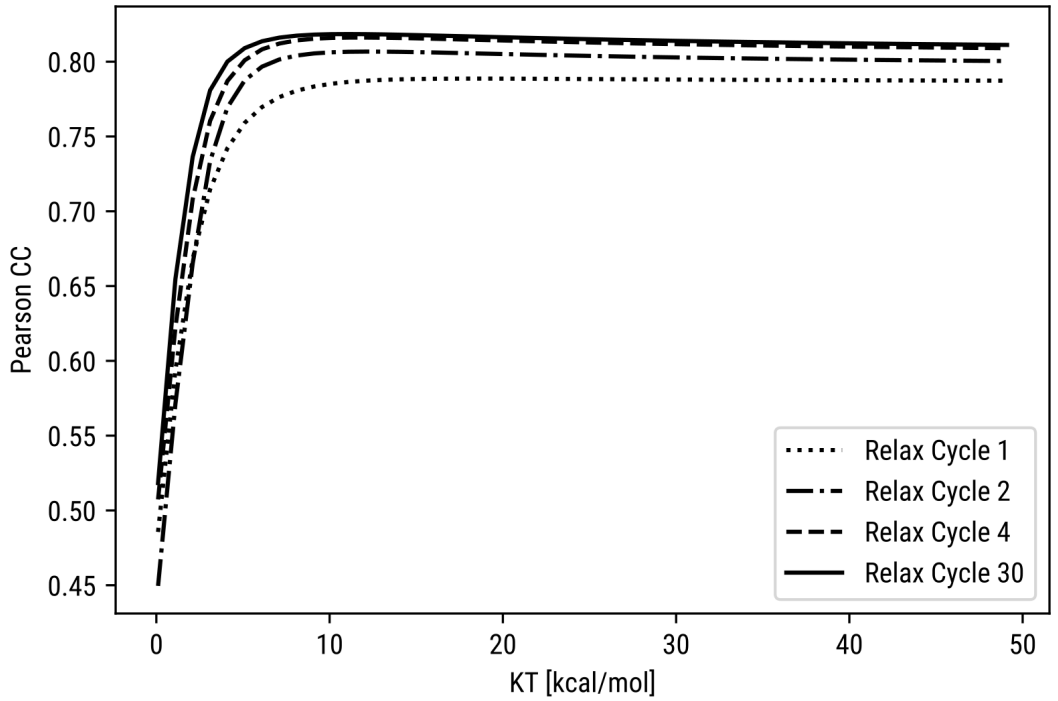

Figure 25. PCC convergence as a function of the relaxed cycles applied. Each curve considers the PCC at different KT values, only considering energy values from conformations generated until the specified relaxation cycles.

To look further into which starting conformations were more beneficial in terms of energy optimisation, we plotted the Boltzmann probability contribution of each starting conformation to each of the peptides in the binding dataset (Figure 27). This probability was calculated by summing the Boltzmann probabilities for all conformations whose trajectories originated from a particular peptide conformation in the structure dataset.

All peptide conformations have a non-zero probability contribution to the Boltzmann distribution of each peptide in the binding dataset. However, a cluster of related peptide conformations has higher Boltzmann weights for all peptides in the dataset (Figure 27). This result suggests that most peptides bound to specific MHC receptors seem to adopt prefered backbone conformations independently of their sequence.

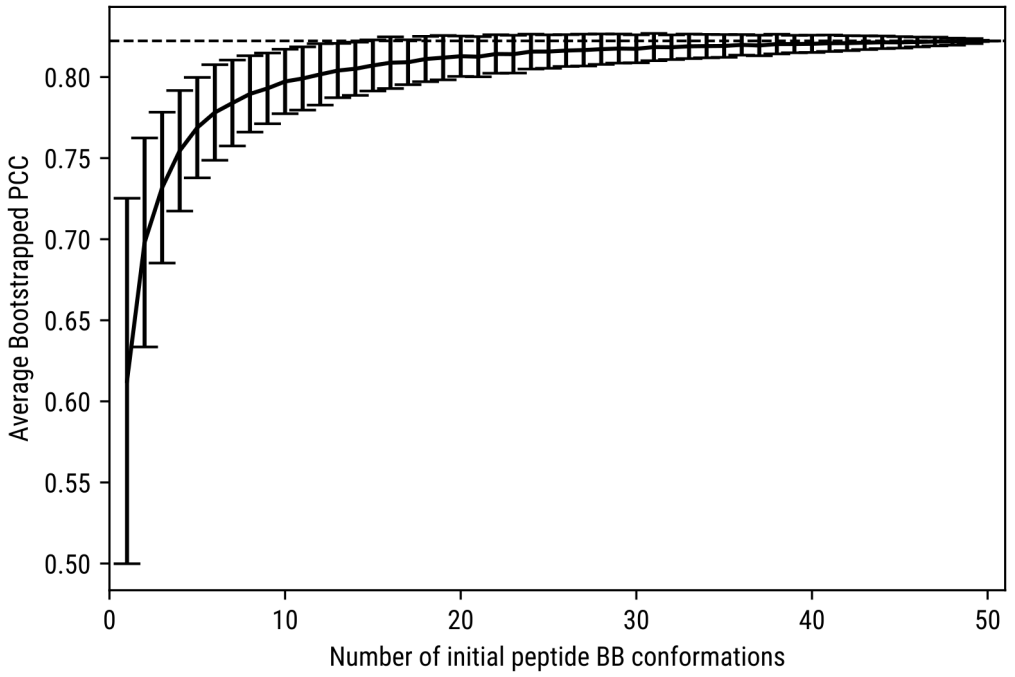

Figure 26. PCC convergence as a function of the number of initial backbone conformations. The average of the bootstrapped PCC is shown, and error bars represent its bootstrapped standard deviation. Bootstrapping was carried out by resampling trajectories starting from different numbers of peptide backbone conformations.

Structural binding analysis

A set of predicted free energy values that correlate with experimental data allows us to look at the physical interactions behind MHC-specificity more confidently. Since the rosetta score function is residue-decomposable, we mapped the Boltzmann-averaged interface energies into the surface of the minimum-energy sampled complex structures (Figure 28). The MHC receptor seems to have considerable plasticity to bind each peptide. Residues on top of the peptide-binding pocket, which are fairly solvent-exposed, are very dynamic and help accommodate diverse peptide side chains, sometimes contacting peptides of different sequence positions. Still, peptides make strong interactions at their N- and C-termini with the MHC pocket's anchorage points, while very little at middle-bottom regions. These regions with absent or loose interactions are expected to have higher conformational dynamics when unbound, thus possibly changing when binding with the T-cell receptor.

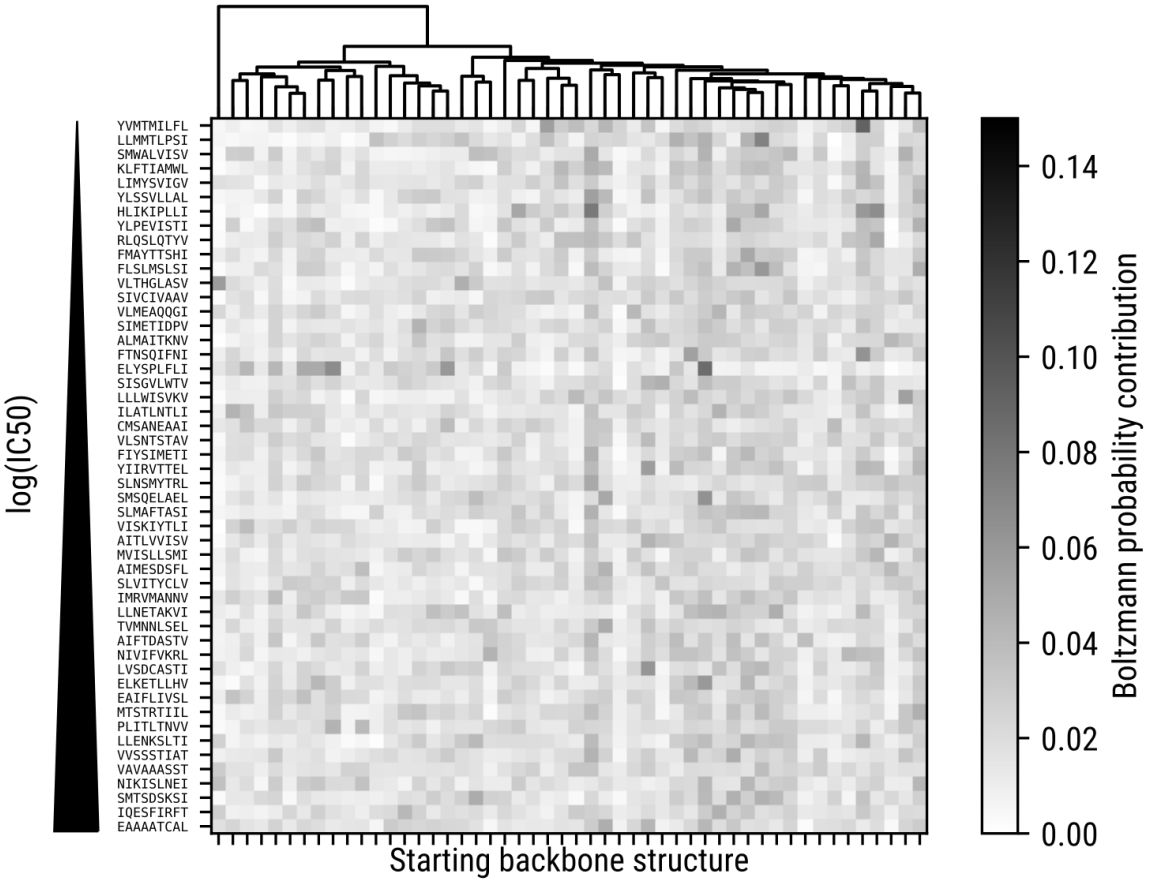

Figure 27. Boltzmann probability contribution of each starting structure to the modeling of peptides in the validation dataset. Peptides are ordered by log(IC50) values and starting peptide structures by a dendrogram based on their pairwise RMSD values.

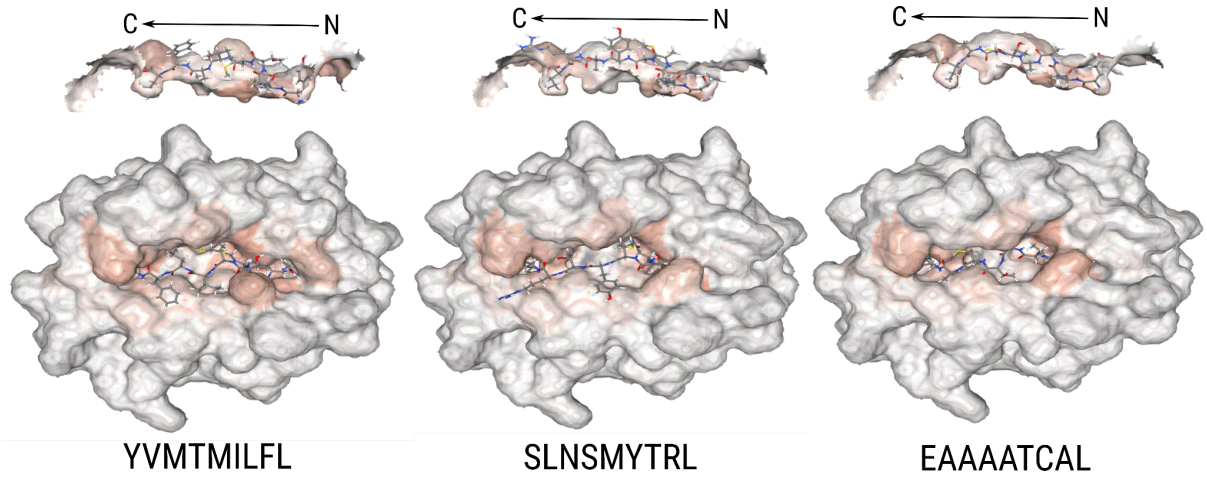

Figure 28. (Previous page) Surface mapping of interaction energies in the MHC receptor. Three MHC of high (YVMTMILFL), medium (SLNSMYTRL) and low (EAAAATCAL) affinity for the MHC receptor are shown. The minimum energy structures sampled are used to depict each residue's interaction free energy values in the MHC pocket. (Up) Side view of the MHC peptide-pocket showing the lowest energy peptide conformation over it. The peptide polarity is indicated from the N-terminus to the C-terminus. (Down) Top view of the MHC_p complex. Each peptide sequence is indicated at the bottom for each complex.

Discussion

We have selected a MHCp system to benchmark our method for predicting binding free energies. Despite not being directly related to activation free energies, predicting binding free energies can be a close and relevant proof-of-concept scenario in which substrate, TS, and product binding could later be related to enzymatic catalytic parameters.[92] More importantly, the Rosetta energy function has not usually being employed in predicting free energies, and since it has been mainly trained and validated to model protein-protein interactions, benchmarking it first in the MHCp system at which these interactions dominate could alleviate artefacts coming from small-molecule modelling and parameterization. In this regard, a recent benchmark concluded that the latest Rosetta energy functions performed worse than previous versions in predicting protein-ligand interactions.[61]

The advantages of the MHC system for studying binding rely on their comprehensive experimental characterisation, but also in that these systems have characteristic binding modes while retaining high promiscuity of the peptidic sequences bound to each MHC receptor, making the predictions challenging, especially at the enthalpic level and less so at the entropic level.

Our simulations correlated well with experimental data for a diverse dataset of peptide sequences and experimental values. The modelling, however, required experimental knowledge of characteristic binding modes to sample peptide conformations. Although this strategy severely limits the application of our method for other MHC allotypes, our primary goal was to benchmark the Rosetta energy function for generating trustable free energy values. In this sense, we reached high correlations (PCC 0.8185) with the experimental dataset, indicating that the binding free energies, derived with the Rosetta energy function (i.e., the Boltzmann-averaged interface scores), applied to a set of sampled conformations, capture experimental trends for

protein-peptide binding. Further benchmarks, however, would be necessary to demonstrate the generalizability of our method to predict thermodynamic parameters in other systems.

When looking in detail at the origin ohe correlation with experimental data, we noted the need for a high KT value when calculating the Boltzmann partition function of the conformational ensemble. The Rosetta energy function combines physical- and knowledge-based energy terms.[47] These knowledge-based terms make the energy function bear arbitrary units (REU); however, this function was recently parameterized using thermodynamic data to convey values in kcal/mol units.[93] The KT constant for a statistical ensemble at room temperature is 0.593 kcal/mol. However, our PCC peaked at a very different value of 11.0 kcal/mol. Although at this point, we cannot discard overfitting to our data or problems with the Rosetta energy function parameterization to kcal/mol, the significant difference remains puzzling.

We decided to look deeper at the composition of our dataset since there are many hydrophobic residues in the peptide sequences. As expected, most peptides were predicted to be hydrophobic; however, the more hydrophobic half had a very different behaviour when predicting correlations than the less hydrophobic half. The KT peaked at lower values (6.6 kcal/mol) for the former group than for the latter, whose PCC did not peak and continued increasing at larger KT values. This result indicates that a biased dataset composition could have affected the fact that the maximal correlation was obtained at a high KT value. Other datasets with different peptidic compositions should also be tested to shed light on this matter.

Two Rosetta energy terms were responsible for most of the conveyed correlation. The 'fa_atr' term represents the attractive part of the Van der Waals interactions, and the 'fa_sol' term represents a penalty for desolvating atoms (i.e., the implicit solvent term). The combination of both terms was almost wholly responsible for the obtained correlation, leaving almost negligible contributions for other energy terms. On the contrary, when removed, the 'fa_elec' term had little influence on the correlation and, when compared alone, negatively correlated with the experimental data. These results make sense if we consider the mainly hydrophobic composition of the dataset: for hydrophobic residues, electrostatic interactions are minimal, and the Van der Waals term is primarily responsible for their non-bonded interactions. Besides, a significant

driving force for the binding of hydrophobic molecules in solution is the hydrophobic effect, which is indirectly captured by the implicit solvation term, 'fa_sol', of the Rosetta energy function.

Since many MHC binding modes have already been characterised, the sampling was started from a comprehensive set of relevant experimental peptide structures. Peptides were subjected to a local exploration from these starting conformations, thus generating structures close to them. We showed that the method generated good correlations with the experimental data with just a few of these starting structures. Moreover, modelled peptides, almost independently of their sequence, preferred a structurally-related subset of these conformations to generate their lowest energy conformations. This result points to a mechanism in which most peptides are bound to a particular MHC using distinctly related backbone conformations. This fact could have significant effects on modelling MHC interactions since discovering these preferred binding conformations could allow for faster structural modelling of peptide interactions for specific MHC allotypes.

Since the Rosetta energy function is also residue decomposable, we could map the interaction energy contributions into the MHC pocket residues. This analysis confirmed the canonical view that peptides make the most substantial interactions at anchorage points in the MHC pocket. The pocket also presented high plasticity, with many solvent-exposed residues located at the top of the helices, helping accommodate peptide side chains of different sequence positions in different peptides. On the other hand, interactions were absent in the middle regions of the pocket. This lack of interactions makes sense in the mechanism of T-cell receptor recognition, in which many peptides seem to bulge out at their middle section, helping them reach out of the pocket for interacting with the T-cell receptor.[94] These loose interactions will also allow for larger dynamics of the peptide at this middle regions, allowing the peptide to adopt alternative conformations for T-cell interaction, as has been observed in specific cases of T-cell recognition.[95]

Conclusions

We have validated the use of the Rosetta energy function to predict experimental activities by using conformational sampling in combination with a statistical-mechanics ensemble analysis. The architecture of the energy function allowed us to understand the physical origin of the

predictions and track these contributions at the residue level, which is highly important for experimental design since mutations target changes residue wise.

The results obtained are relevant for developing fast techniques to predict thermodynamic activities since the simplifications made over the energy function seem to have maintained a good balance between physics- and knowledge-based energy terms that allows for fast sampling of protein conformations. Although further benchmarks are necessary, this method could also be used to predict protein-small molecule experimental data and, if properly validated, adequately help filter model proposals in protein design algorithms.

Methods

Peptide conformational dataset

A structural dataset containing MHC_p complexes was compiled to assess how different peptide conformations affect binding affinity predictions. A set containing only HLA-A, HLA-B, and HLA-C MHC-I receptors was built from a list of 534 PDBs. MHCp complexes (i.e., MHC alpha, β2-microglobulin, and peptide chain) were built and filtered to leave only complexes with nonameric peptides and no missing coordinates at residues near the peptide binding pocket. The selected MHCp complexes were further clustered, based on peptide pairwise RMSD, to select a set of 50 representative peptide conformations.

Modelling MHC-I and peptide bound conformations

All PDB structures of the HLA-A*02:01 MHC receptor were used to build MHC complexes consisting only of the MHC alpha chain and its corresponding β2-microglobulin chain. Using the Rosetta all-atom score function, these models were subjected to an all-atom minimisation,[47] and the best scoring pose was chosen for modelling the peptide conformations (PDB code 3UTQ). Each peptide in the binding dataset was modelled into the MHC receptor using all peptide conformations in the previously built peptide conformational dataset. The peptide sequence was threaded into each conformation and then subjected to 30 cycles of local backbone, side-chain, and rigid body optimisation using Rosetta's[50] fastrelax algorithm[54] with default options. For each cycle we stored the total system energy (total score) and the binding energy between the peptide and the MHC-I receptor (interface score). Flexible residues at the backbone and side-chain levels

included peptide residues and receptor residues near 8 Å of the peptide coordinates. This list of residues was updated at the beginning of every cycle. The fastrelax method minimises the complex by searching its local conformational space by decreasing the repulsive term and increasing the attractive term of the score function in the early inner cycles, scaling back to the default weights as the minimisation progresses. Ten replicas of this complete procedure were carried out to estimate the method's convergence, totalling 750000 conformations for the full set of peptides (15000 for each MHC_p complex).

Binding energy calculations

Binding energy values were calculated as the expectation value of the interaction (interface) score between the peptide and the MHC receptor, using a Boltzmann distribution based on the total complex energy:

$$< E^b >= \sum_{i}^{N} p_i E_i^b \tag{13}$$

Here, $< E^b >$ is the expectation value of the interaction score, N is the total number of sampled MHC_p conformations, and E_i^b is the interaction score of each specific MHC_p conformation. The interaction score for each MHC_p conformation is calculated as:

$$E_i^b = E_i^{MHC_p} - (E_i^p + E_i^{MHC}) \tag{14}$$

With $E_i^{MHC_p}$ is the total energy of the MHC_p complex conformation, E_i^p the unbound peptide conformation energy, and E_i^{MHC} is the MHC receptor conformation energy without the bound peptide. The probabilities p_i are obtained from a Boltzmann distribution using the N sampled MHC_p conformations $E_i^{MHC_p}$ scores:

$$P_i = \frac{e^{-E_i^{MHC_p}/KT}}{Q} \tag{15}$$

Here, KT is the characteristic energy partition, and Q represents the partition function calculated as:

$$Q = \sum_{i}^{N} e^{-E_i/KT} \tag{16}$$

Bootstrapping analysis

Peptide backbone conformations were bootstrapped by taking resampling with replacement subsets of conformations increasing in number. The PCC analysis was repeated by only using conformations that started from the selected conformations. After all the resamplings were carried out, the average and standard deviation of the PCC distributions were calculated and reported.

Chapter 3. Dynamical and binding predictions using the WCN metric

The WCN is a metric developed to estimate the packing density around a specific residue or atom in a protein structure. It has been described to correlate well with dynamic[96] and evolutionary[97,98] profiles in proteins and to be helpful to characterise protein-protein interfaces[99] and conserved catalytic residues in enzymes.[100,101] On the one hand, the possibility of predicting dynamic profiles from the structure alone and with little computation, makes the WCN an attractive metric that capture the preorganisation of residues for being used directly in an enzyme design algorithm. On the other hand, the degree of agreement between the WCN and evolutionary data could aid in determining functional patterns in proteins with uncharacterised functions.

To demonstrate the practicality of the WCN in modelling protein dynamics and functional characterisation, we sought to validate its use two-folds:

We first decided to explore the ability of the WCN to predict dynamic profiles of diverse protein structures, using as a an experimental descriptor the Debye-Waller or temperature factor (B-factor), derived from X-ray crystallography experiments.[102,103] This parameter indirectly measures the squared atomic displacement of the atoms in the crystal system, which is associated with the uncertainty in the x-ray scattering patterns. The origin of these variabilities can be related to the different conformational substates of the protein in the crystal and their corresponding vibrational modes, but also to other factors like lattice disorders and translational and rotational diffusions. Despite these possible non-dynamical artefacts, B-factors describe structural fluctuations occurring at very different timescales, from femtoseconds to seconds, provided that relevant conformations are highly populated in the crystal lattice.[104] This description gives rich information about protein motions, which is fundamental to understanding the relationship between the dynamic structure-activity that underlies their biochemical functions.

Secondly, we sought to define a general method to map possible regions of protein-protein binding. For this goal, we used the problem of characterising the unknown binding mode between the IκBα protein and its binding partner, the histone H4.[105] Canonically, the IκBα receptor binds to

the NF-κB transcription factor in the cytosol, preventing its migration to the nucleus, and thus inhibiting its action as a transcriptional regulator.[106] It has been proposed that the IκBα protein can also bind the N-terminal tail of histones H2 and H4, which adds a new role to the IκBα protein as a histone-code regulator.[105] While the interaction of IκBα with NF-κB has already been characterised at the structural level, the interaction of IκBα with the histone's N-terminal tails is yet to be characterised.

We start this chapter by analysing the ability of different WCN variants to predict dynamic profiles, and we continue with the derivation of a general score helpful to map residues with essential functions beyond their structural role.

Results

Validation of the WCN to predict dynamical profiles

We first decided to test the ability of WCN to predict the dynamic profiles of a diverse dataset of protein structures. We compared three different WCN profiles against experimental B-factor values considering separate atoms subsets. The dataset comprises 153 single-chain proteins with 1.5 Å or better resolutions, diffracted between 95K and 105K (for more details, see the "Protein dataset collection for dynamical predictions" section in methods). Two coarse-grained models of WCN, considering only backbone CA atoms ($WCN_{C\text{-}alpha}$) and backbone CA atoms plus sidechain centroid values ($WCN_{Centroid}$), were contrasted to the per-residue-averaged B-factor profile of each protein. An average Pearson(Spearman) correlation coefficient (CC) of 0.676(0.697) and 0.684(0.715) was obtained for each model, respectively. The best and worst predictions are shown in Figure 29 to gain insight into the performance of these WCN models.

Analogously, we calculated a higher resolution WCN profile, considering all the atoms in the protein structures ($WCN_{All\text{-}atom}$), and compared it to the all-atom B-factor profile of each protein. An improved average Pearson(Spearman) correlation coefficient of 0.691(0.733) was estimated for the entire data set. The worst and best predictions for this model are shown in Figure 30.

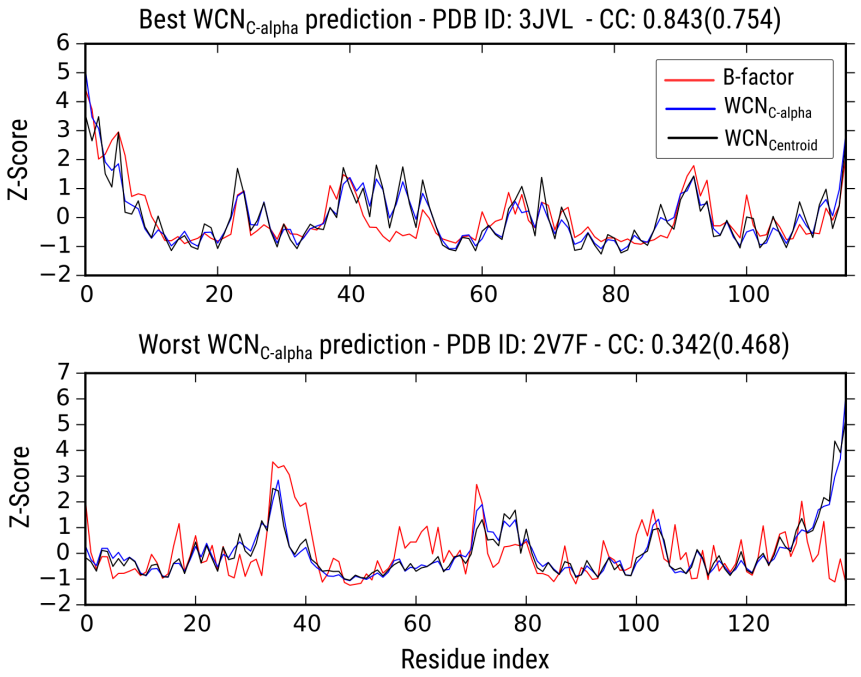

Figure 29. Best and worst predictions for the WCNC-alpha and WCNCentriod profiles. The Pearson(Spearman) CC between the WCNC-alpha profile and the B-factor profile is shown in the title together with the PDB code of the structural entry being compared.

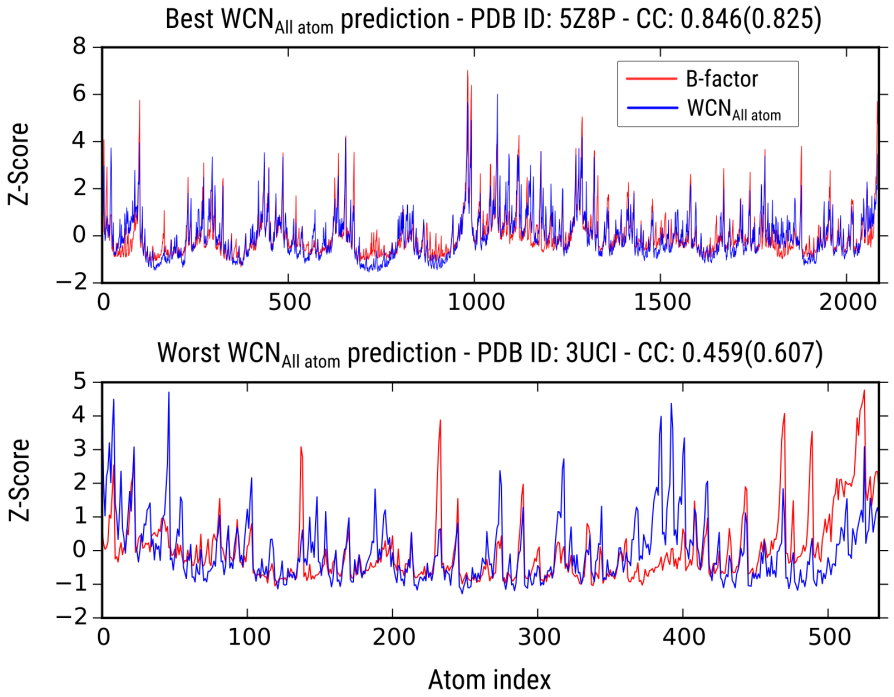

Figure 30. Best and worst predictions for the WCNAll atom profiles. The Pearson(Spearman) CC between the WCNAll atom profile and the B-factor profile is shown in the title together with the PDB code of the structural entry being compared.

These results validate that predictions made with alternative versions of the WCN, differing in the granularity, significantly agree with the crystallographically-derived dynamic profiles of different protein folds.

We now proceed to explore the utility of WCN in the prediction of binding regions based on evolutionary information.

The IκBα and NF-κB complex

IκBα has an ankyrin repeat fold that exposes a large interaction surface. It binds the two subunits of the NF-κB dimer (i.e., the p65 and the P50 subunits) at different sides of its ankyrin repeat fold (Figure 31). The crystal structure shows six ankyrin repeats forming the central module of the IκBα protein; however, there is missing structural information at the N- and C-terminus. The C-terminal domain of the Rel Homology Regions (RHR) of the p65 subunit forms ample interactions with IκBα, while the N-terminal RHR domain makes scarce interactions. Only the C-terminal RHR domain of the p50 subunit interacts with IκBα using a different region than the p65 domains with scarce interface overlap. The nuclear localisation signal (NLS) of the NF-κB p65 subunit extends away from the p65 C-terminal RHR domain and adopts an alpha-helix conformation that contacts the first and second ankyrin repeats of the IκBα domain.

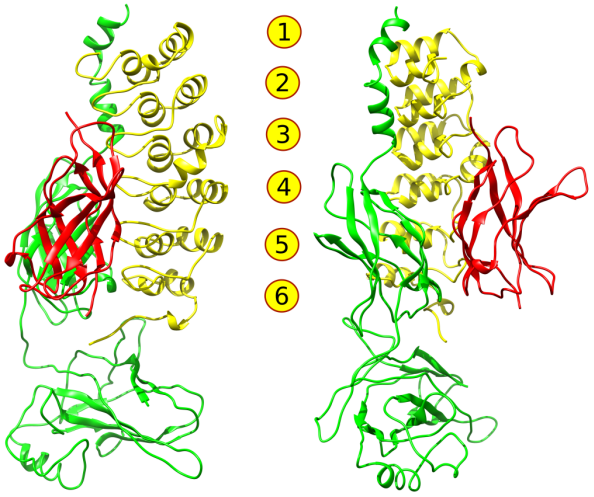

Figure 31. The crystallographic structure of IκBα in contact with two subunits of NF-κB. IKBA is shown in yellow with 6 stacked ankyrin repeats. The RHR of the p65 (green) and the P50 (red) subunits contact different interfaces of the IκBα protein. The complex is rotated by 90º in the right relative to the left model — PDB structure 1NFI.

Correlations between WCN and evolutionary information

WCN profiles for the IκBα structure were built using only the contacts present in the IκBα single-chain structure ($WCN_{single\text{-}chain}$) or using the contacts present in the structure in complex with NF-κB ($WCN_{complex}$). These two profiles were compared to the sequence conservation (SC) data derived from the Consurf[107] web server (Fig. 40). The $WCN_{single\text{-}chain}$ profile, derived using the single-chain structure of the IκBα domain, follows the SC profile with many differences in several regions (PCC: 0.3515). However, when contact information from the IκBα structure in complex with NF-κB is included in the WCN calculation (i.e., $WNC_{complex}$), the correlation augments significantly (PCC: 0.5080), evidencing that NF-κB positions participating in binding are conserved among the set of homologous proteins considered to produce the SC profile.

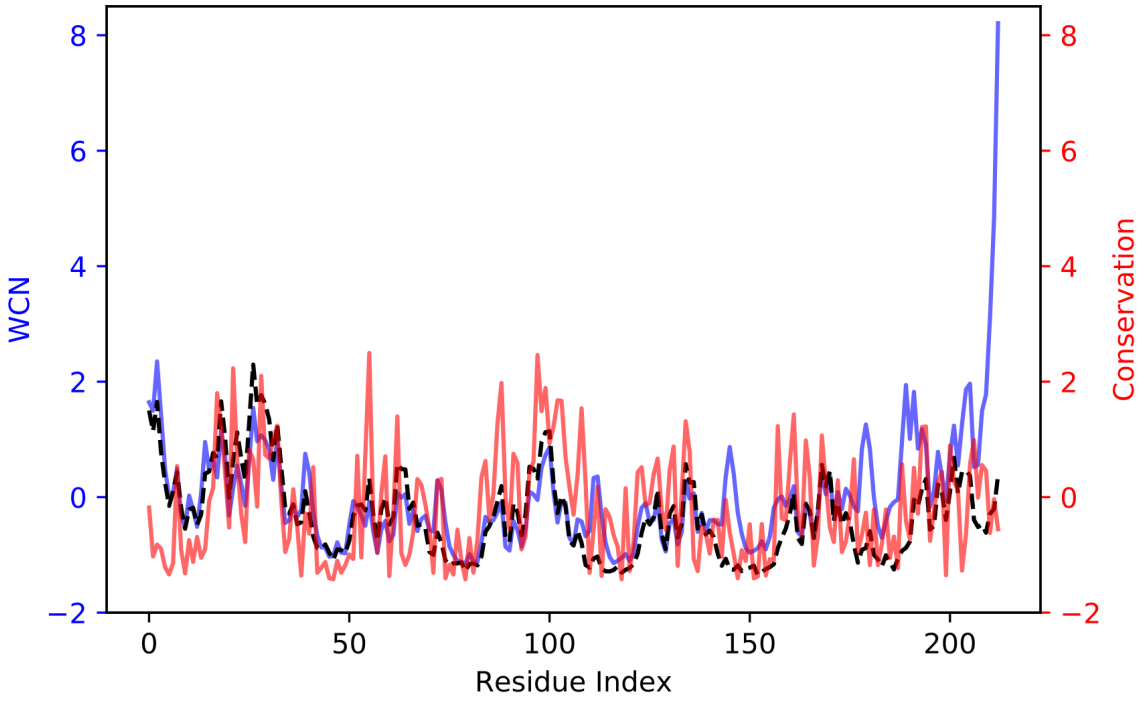

Figure 32. WCN and SC profiles for IκBα. The WCN profiles for the single-chain (blue line) and complex (dashed black line) structures are shown together with the SC data (red line) derived from the comparison of homologous proteins. All profiles represent their normalized Z-score values.

With this result and to validate the prediction of interface residues, we next wondered which IκBα residues of its interface with NF-κB could be predicted by only considering the $WCN_{single\text{-}chain}$ and the SC profile for the IκBα protein.

Predicting binding interface residues using WCN and evolutionary information

The SC profile represents the variation in amino acid identity at each residue position of the IκBα protein, calculated by comparing a set of homologous proteins. As measured above, the $WCN_{complex}$ profile is closer to the SC profile than the $WCN_{single\text{-}chain}$ profile. This more significant similarity stems from the fact that the conserved biological unit is the IκBα protein in complex with the NF-κB subunits, and not the monomeric IκBα structure. Residues conserved on the surface of IκBα are better explained by the $WCN_{complex}$ profile since residues in contact between the two proteins must have certainly co-evolved.[108] Therefore, it is reasonable to think that differences between the SC profile and the $WCN_{single\text{-}chain}$ could reflect residues with further evolutionary constraints than only maintaining the IκBα domain folded structure.

We derived a simple score to estimate IκBα residues with additional roles besides maintaining its ankyrin repeat fold. The score is deemed FEEC (Fold-Excluded Evolutionary Conservation) and is defined as the difference between the $WCN_{single\text{-}chain}$ minus the SC values. Based on this definition, residues with a positive FEEC score could have additional conservation constraints than those imposed by the protein's tertiary structure alone. To validate this score, we classify which interface residues between IκBα and NF-κB could be predicted using the FEEC score alone. The interface residues were defined as having more than 20% of their solvent-accessible surface area hidden ($SASA_h$) upon complexation.

From the 45 IκBα residues in the IκBα/NF-κB complex, 36 belong to the interface with the NF-κB p65 subunit and 11 to the interface with NF-κB p50 subunit (two residues are shared between these interfaces). From the 213 IκBα residues in the crystallographic structure, 121 residues have a positive FEEC score. For the interaction with NF-κB p65 subunit, 83% (30/36) of the interface residues with IκBα are correctly included in this set of residues, being also true for 73% (8/11) of the interface residues interacting with the NF-κB p50 subunit.

To visualize the pattern of surface residues with positive FEEC scores, we mapped their FEEC values onto the surface of the IκBα protein (Figure 33 top). Positive FEEC values are clustered in specific regions of the protein, and they seem to be higher at the terminal residues of the IκBα structure. It is relevant to point out that residues with extreme FEEC values appear to be an artefact of missing structural information of the available IκBα structure. Terminal segments of

the protein are not represented in the structure, which significantly affects the WCN calculation of nearby residues (see Figure 32), resulting in very high FEEC values. Because of this artefact, the selected high-end colour cutoff, representing the mapping of the positive FEEC score values into the protein surface, was kept low to depict better the FEEC-delineated regions (Figure 33 top).

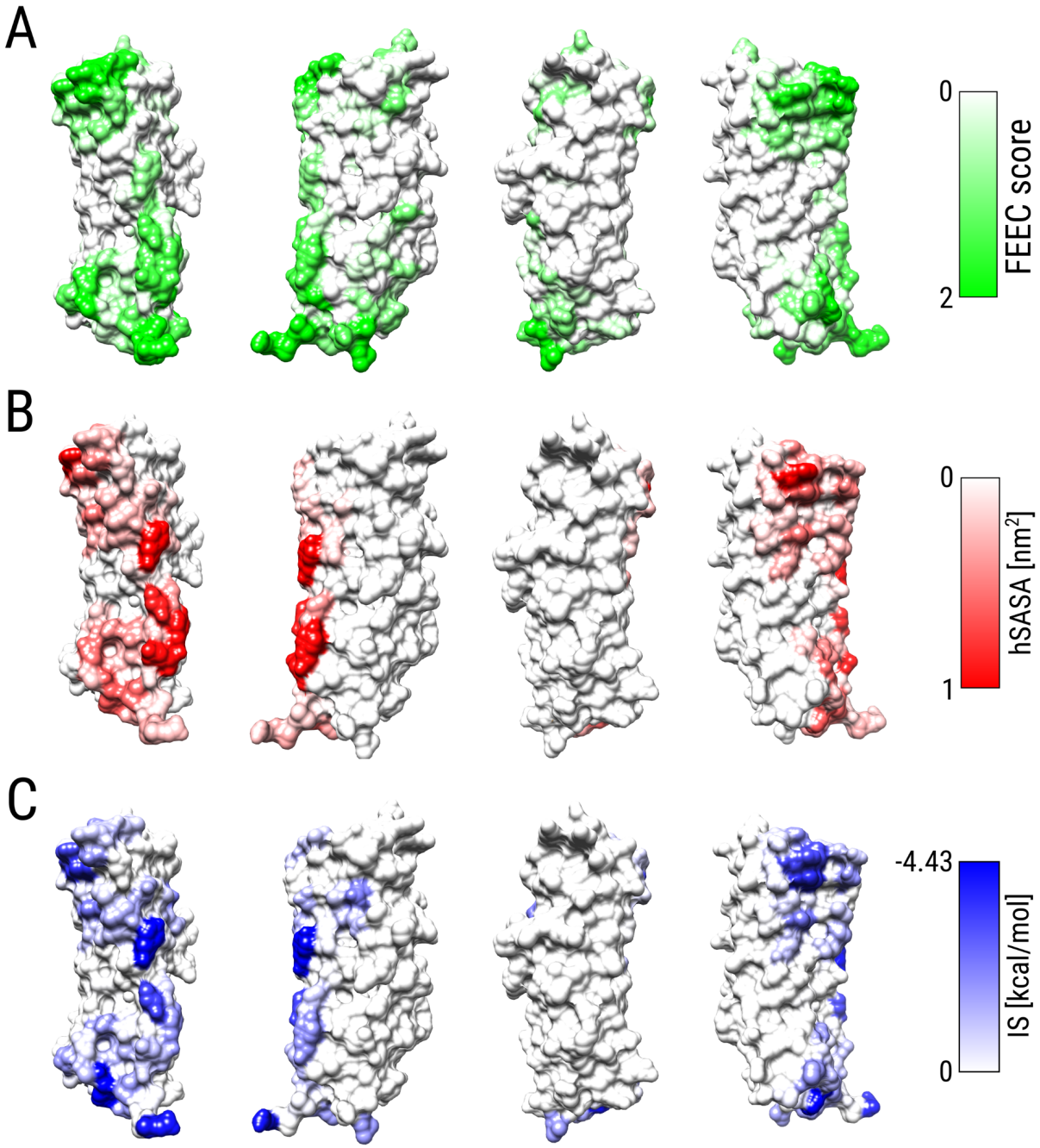

Figure 33. Different metrics characterising surface residues on the IκBα structure. (Top) Surface mapping of FEEC values of residues with positive FEEC scores. (Middle) hidden SASA value upon complexation calculated from the IκBα and NF-Kb complex structure. (Bottom) Binding scores calculated from sampling local conformations of the IκBα and NF-Kb complex structure. Each structure to the right represents a 90º turn of the visualization of the structure to its left – PDB structure 1NFI.

To compare the FEEC scores distribution with the distribution of residues participating in the interface of IκBα with the NF-κB subunits, the per-residue $SASA_h$ values were also mapped into the IκBα surface (Figure 33 middle). Interestingly, there is a similar distribution pattern between residues with positive FEEC scores and residues hidden upon complex formation. While $SASA_h$ is derived from a geometric definition of residues belonging to the interface, it does not account for chemical interactions nor their strength. Thus, it is expected that residues not participating in significant interactions are still included in the interface because they are occluded by nearby interface residues. Therefore, we mapped per-residue binding scores to define more quantitatively the residues' role in the IκBα and NF-κB interface. When mapped into the IκBα surface, the binding scores values also show a similar distribution to FEEC values (Figure 33 bottom).

When classifying residues belonging to the interface, it is essential to study the effect of changing the threshold that defines which residues pertaining to the interface. Thus, the percentage of residues in the interface correctly classified by the FEEC score at different SASAh threshold percentages was plotted in Figure 34. The number of residues classified as being at the interface drops almost linearly when the SASAh threshold percentage is increased. However, at least 80% of residues had a positive FEEC score at all threshold values. A 100% of interface residues have positive FEEC scores when the SASAh threshold is 74% or higher, although, at this threshold value, the interface is only composed of the innermost 16 residues.

We repeated the threshold analysis using now per-residue binding energies (Figure 35). The number of residues in the interface with favourable binding energies (i.e., negative values) slowly increases when the binding energy threshold increases, with most residues having lower binding energy contributions to the interface. At least 79% of residues with negative binding energy values have positive FEEC values. 100% of interface residues have positive FEEC scores when the binding energy threshold is -2.1 kcal/mol or lower, with the interface containing nine or fewer interface residues.

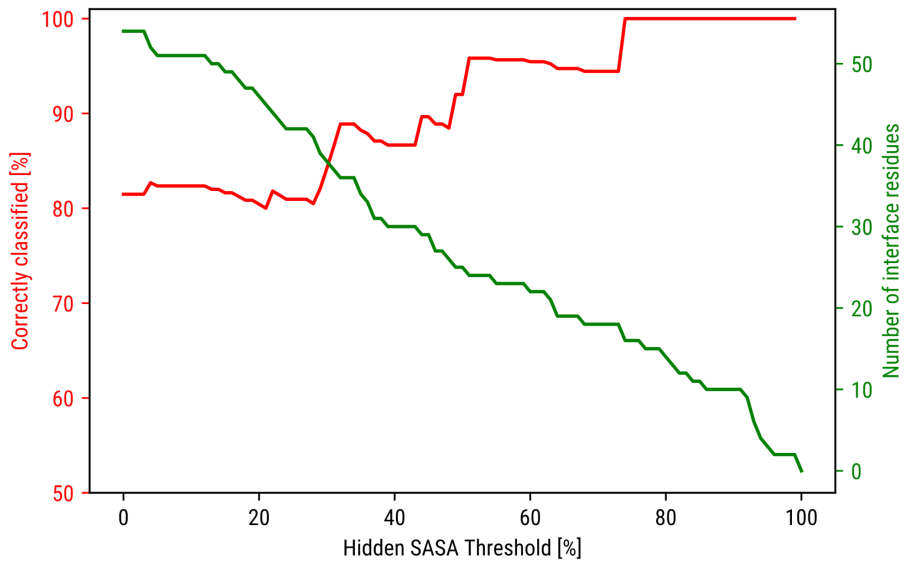

Figure 34. FEEC-classification of interface residues as defined by their $SASA_h$ values. The percentage of correctly classified interface residues by the FEEC score (red line) is shown as a function of the $SASA_h$ threshold percentage used to define which residues belong to the interface. The number of residues included in the interface for each specific $SASA_h$ threshold (green line) is also shown.

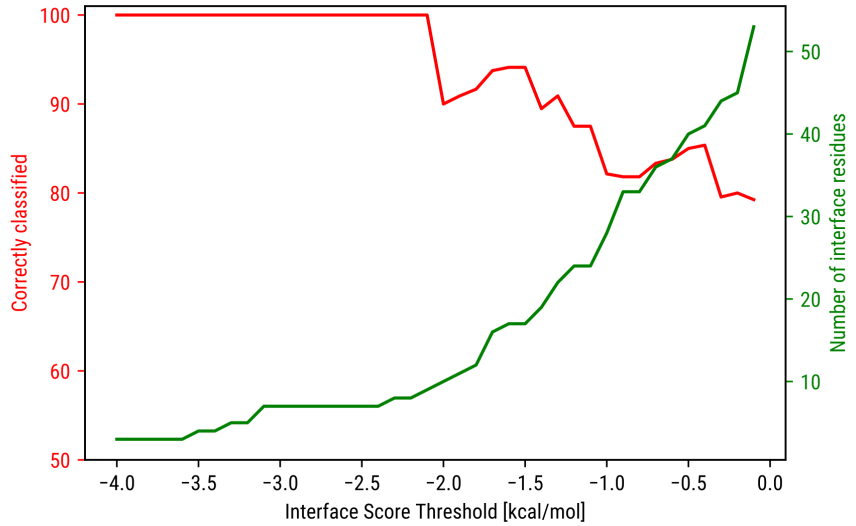

Figure 35. FEEC-classification of interface residues as defined by their interface scores values. The percentage of correctly classified interface residues by the FEEC score (red line) is shown as a function of the interface score threshold value used to define which residues belong to the interface. The number of residues included in the interface having a score less or equal to each interface scores threshold (green line) is also shown.

A putative binding site for H4 N-terminus to IκBα protein

Having investigated the predicting ability of the FEEC score to map regions where the NF-κB protein can bind to IκBα, a definition of a putative region for the binding of the H4 histone N-terminal region is addressed. The sequence of the H4 histone N-terminal region that binds IκBα is: SGRGKGGKGLGKGGAKRHRKVLR.[105] The sequence contains many arginine and lysine residues (red letters) which makes the overall region highly positively charged. Therefore, it is reasonable to propose that any region binding this protein segment should have a complementary electrostatic surface (i.e., a region with many negatively charged residues). To narrow down possible binding regions, we searched for all negatively-charged surface residues with positive FEEC values (Figure 36).

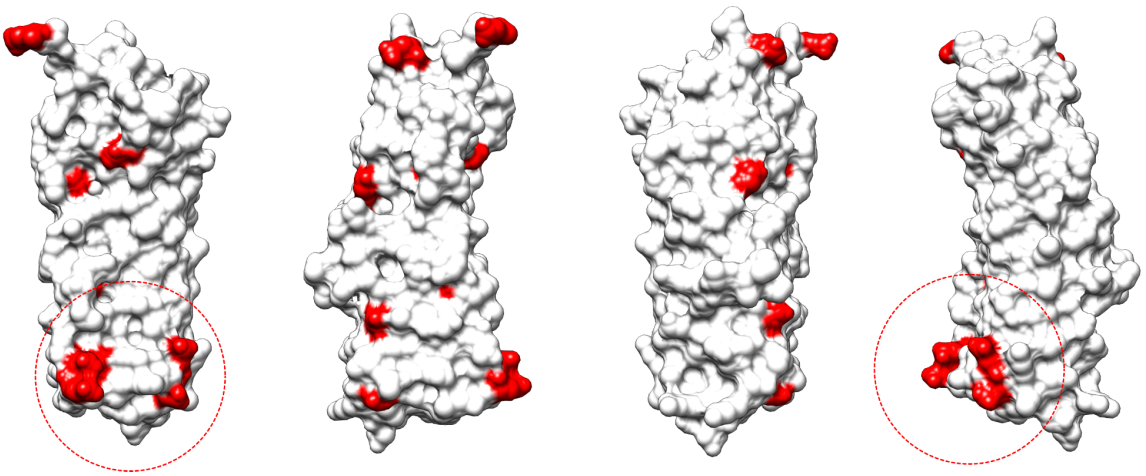

Figure 36. Mapping of negatively charged residues with positive FEEC score on the IκBα surface. The dashed red circle marks the largest negatively-charged region, with four negative residues — PDB structure 1NFI.

The N-terminal side of the IκBα protein comprises the more significant number of negatively charged residues with positive FEEC values (red circle in Figure 36). In this region, five negatively charged residues are found, E72, D73, D75, E85, and E86, coincidentally also part of the IκBα and NF-κB p65 subunit interface (Figure 37). Notably, this region binds the NLS of NF-κB, also containing a high number of positively charged residues.

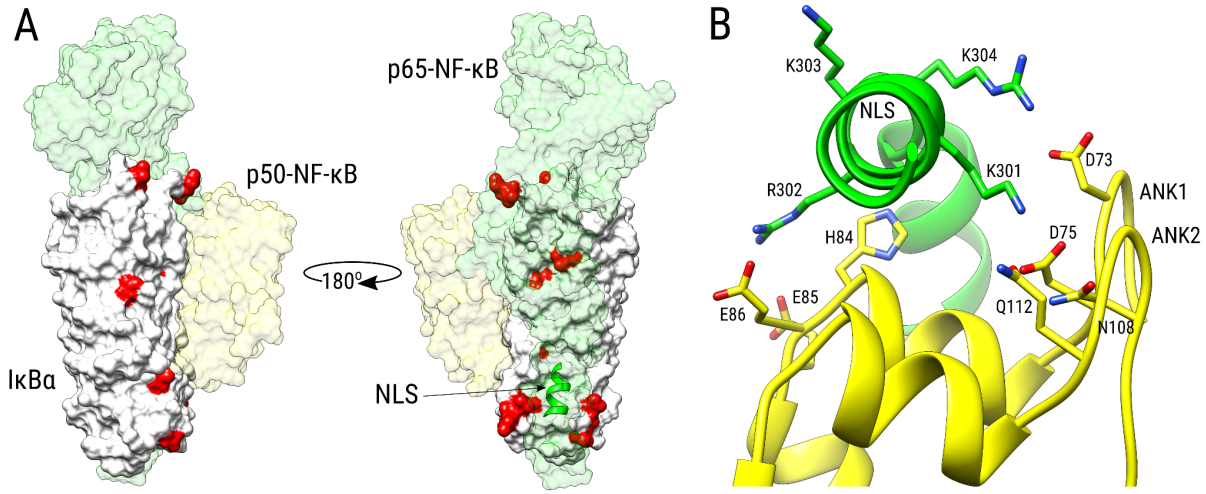

Figure 37. A) The IκBα (white) and NF-κB/p65 (translucent green) and p50 (translucent yellow) complex is shown to indicate the relative location of negative surface residues with positive FEEC values (red surfaces). The NF-κB/p65 nuclear localisation signal (NLS) region is drawn as a green cartoon. B) Interaction of NF-κB/p65 NLS motif with IκBα ANK1 and ANK2 repeats. Polar IκBα residues interacting with the NLS are depicted yellow, the NLS motif, KRKR, is shown in licorice. They are numbered according to Uniprot entries Q04206 (NF-κB/p65) and P25963 (IκBα) — PDB structure 1NFI.

When comparing the sequences of the NF-κB NLS and the H4 N-terminus regions, they have a shared motif of positively charged residues (Figure 38), suggesting this five-residue-containing region as the putative binding region of H4 histone N-terminus to bind the IκBα protein at the defined region (see Fig. 36).

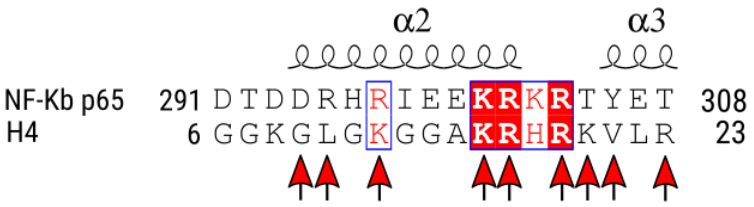

Figure 38. Alignment of the NF-κB/p65 NLS region to the H4 N-terminal tail. Similar positions are indicated in blue rectangles; similar and identical residues, inside a position, are in red and white letters, respectively. Red arrows indicate NF-κB/p65 residues participating in IκBα binding. The secondarySecondary structure shown above was calculated from PDB structure 1NFI.

Discussion

Residues conserved in a particular set of homologous proteins need to fulfil a relevant thermodynamic or kinetic role for the given biomolecular system. This role can be the folding

structure's maintenance, or it can regulate other phenomena that, without being exhaustive, could mediate protein-protein interactions, maintain protein solubility, speed up folding kinetics, or affect the preorganisation of interface or catalytic residues. A three-dimensional protein structure contains the chemical contact information necessary to adopt its folded conformation. The greater the number of chemical contacts a residue participates in, the more restricted it is to change its identity without compromising essential interactions that support the folded structure. Likewise, if the residue is not involved in many chemical interactions, it can vary more freely along the structurally-permitted sequence landscape.

The atomic contact density, measured here continually as the WCN metric, has been shown to correlate well with the dynamic and evolutionary profiles of proteins.[96,98] The latter correlation is higher when the WCN values are derived using the full biological complex structure.[98] Since the WCN values only consider the contact information derived from a particular protein structure, residue positions with higher SC than the one dictated by the WCN metric could have different or additional roles than maintaining the folded configuration. Based on this logic, we defined a FEEC score as the value of the WCN minus the SC scores. Based on this definition, if a residue's FEEC score is positive, additional conservation constraints than those imposed by the protein's tertiary structure alone can apply to them.

When positive FEEC scores were mapped on the surface of the IκBα structure, it delineated regions highly similar to those employed by the protein to interact with its partner protein, NF-κB. All IκBα residues, either deeply located in the interface core or having considerable interaction energies at the interface, were correctly classified by the FEEC score metric. This result is remarkable since the FEEC score derivation is carried out only with the unbound IκBα structure and the sequence information of orthologous proteins. Thus, this metric seems promising to help experimentalists narrow down binding interface location for partner proteins.

We employed the FEEC metric to define the binding region between the IκBα protein and the H4 histone protein. Experimentally, the N-terminus region, specifically the first 23 residues, was shown to bind the IκBα protein,[105] however the exact mechanism of the interaction is unknown. Employing the FEEC metric, we were able to narrow down the H4 histone binding to a specific region, which, according to the crystallographic structure of the IκBα protein in complex with the

NF-κB subunits, was also bound by the NLS of the NF-κB p65 subunit. Interestingly, the NLS and the H4 histone N-terminus sequences share a significant motif of charged residues, indicating that these two segments could have a similar binding mechanism and, therefore, compete for binding IκBα. Ongoing studies have confirmed that this region is indeed the region used by the H4 histone to bind to IκBα and that NF-κB and H4 compete for their binding to IκBα (unpublished results).

Finally, as a validation test, we also explored the utility of the WCN metric to correlate with dynamic profiles for a diverse set of protein structures. This test showed good correlations when employing WCN versions at the residue and atomic levels. The ability of WCN to inexpensively predict dynamic profiles from the protein structure alone opens up the possibility to apply this metric as a fast predictor of atomic- or residue-level conformational dynamics. Importantly, this could be especially useful in tracking the degree of residue preorganisation in protein design algorithms.

Conclusion

The WCN is a helpful metric for quantitatively studying the relationship between protein structure, dynamics, and evolution. We have employed it in conjunction with evolutionary information to derive a score that can identify residues with other roles beside maintaining the folded protein structure. As a particular case, we predicted the protein-protein interfaces of two proteins with an unknown binding mechanism. This same information could easily help classify biological from non-biological interfaces in crystallographic complexes.[109]

The validation of the WCN contact number to predict protein dynamic profiles and suggest hypotheses for residues participating in protein binding makes it an ideal metric to estimate the preorganisation level of protein residues. Such is the case for enzyme design, in which the optimisation of the preorganisation of residue partaking in fundamental catalytic interactions is a promising proposition to create new active sites with increased catalytic activities.

Methods

Protein dataset collection for dynamical predictions

The dataset for fitting the WCN models was selected by searching the whole PDB protein database[65] for structures solved by X-ray crystallography. We considered only crystals composed of a single polypeptide protein in the asymmetric unit, with resolutions better or equal to 1.5 Å. Because B-factors are affected by the temperature at which X-ray dispersion data is collected, for consistency, we only selected structures obtained in the range of 95 and 105 K. Peptide models were filtered out by putting a lower limit of 30 residues to the length of protein sequences. Structures with ligands present that had more than 5 atoms or membrane proteins were discarded. A 35% sequence identity filter was used to generate a non-redundant set of structures, obtaining a maximum value of 32% in the final set. To remove possible crystal contacts bias in the set, we first built all symmetry-related chains at least 5 Å in proximity to the target chain. Then, WCN profiles were derived using the single-chain structure only and, also, using the rebuilt crystal contact neighbouring chains. Both WCN profiles were correlated with the normalized B-factor data available, and structures with PCC not differing more than 5% were regarded as unbiased and included in the final dataset of 153 structures.

Weighted contact number

The WCN is a metric that quantifies the contact density for a particular atom or residue in a protein structure. It is defined as the sum of all inverse squared distances of all other particles in the system to that specific atom:

$$WCN_i = \sum_{j \neq i}^{N} \frac{1}{r_{ij}^2} \tag{17}$$

Here, N is the number of atoms in the structure, and r_{ij} is the distance between atoms i and j. The profiles are normalized by calculating their standard score (z-score). The calculation only considers a subset of atoms in the protein structure depending on the comparison (i.e., at the residue or atomic level). For the $WCN_{\text{alpha-carbon}}$, only the CA atoms are considered; for the WCN_{Centroid}, only the CA atoms and a centroid atom representing the side chain geometric centre

are employed for each residue; finally, for the $WCN_{All\text{-}atom}$, all the protein atoms in the structure are considered. Any WCN metric, as described here, has an inverse relationship with the evolutionary SC;[98] therefore, we use the inverse of the WCN (WCN^{-1}) to make comparisons to the evolutionary information.

Sequence Conservation Score

The SC scores were derived from the CONSURF server.[107] First, the method builds phylogenetic trees from multiple sequence alignment of homologous sequences. Then, considering the stochasticity underlying the evolutionary process, conservation profiles are derived using the Empirical Bayesian Method and smoothed over a window of five residues.[110] Finally, the scores are normalized to their corresponding standard scores (Z-scores).

Fold-Excluded Evolutionary Conservation Score

We defined a FEEC score as the value of the WCN^{-1} minus the SC values:

$$FEEC_i = WCN_i^{-1} - SC_i \tag{18}$$

Analysis of residue interface energy contributions

Considering a protein complex based on two proteins (A and B), we assign a probability to each complex conformation from a Boltzmann distribution based on the full energy landscape:

$$P_i = \frac{e^{-E_i/K_BT}}{Q} \tag{19}$$

$E_i$is the energy of the i[th] structure, K_BT is the energy partition constant, and Q is the partition function of the respective ensemble of N structures:

$$Q = \sum_{i}^{N} e^{-E_i/K_BT} \tag{20}$$

The interface binding energy E_i^B for each complex structure is calculated as

the difference between the energies of the interacting complex structure E_i and the individual chain energies E_i^A and E_I^B as:

$$E_i^B = E_i - (E_i^A + E_i^B) \tag{21}$$

All interface energies were then integrated using the complete set of probabilities to calculate the binding energy expectation value:

$$\langle E^B \rangle = \sum_i^N P_i E_i^B \tag{22}$$

To obtain a set of interacting conformations between IκBα and NF-κB and define an interacting energy landscape, we generated 6500 minimisation trajectories from the crystallographic complex (PDB code: 1NFI) using the fastrelax[54] protocol of the Rosetta software[50] with default options.

We employed the Rosetta energy function[47] in all the modelling steps. The function is residue decomposable, and therefore it is straightforward to do the energy analyses only considering the contribution of individual residues. Accordingly, per-residue interface energy contributions are obtained as the expectation values of individual residues' binding energies.

Chapter 4. Exploring protein conformational landscapes with Structure-Based Model simulations

An important drawback of MD simulations is their slow convergence over the sampled protein conformational space; extremely long simulations are required, if possible, to explore the free-energy landscape of the folded configuration in a convergent manner.[111] This limitation of MD hinders the study of conformational dynamics in protein systems and curtails our capacity to obtain thermodynamically convergent results. This limitation is also valid in the case of enzymatic systems, in which alternative protein configurations can have dissimilar catalytic capabilities, relevant for a proper estimation of their activation free energies.[12]

There are many approaches to deal with the MD sampling problem of protein systems. On the one hand, specialised hardware, such as graphics processing units (GPU), can achieve faster calculations at reduced hardware cost, speeding up MD simulations to sample additional conformational space.[112] Also, many algorithms seek to optimise the conformational search by employing different strategies to enhance the efficiency of the sampled phase-space by a limited number of MD trajectories.[113] Finally, several coarse-grained force fields aim to simplify even more the representation of molecular systems to diminish the number of terms being evaluated by the MD engine, accelerating the conformational sampling convergence.[114]

Structure-based models (SBMs)[115,116] are a special kind of simplifying-force field MD methodology that focuses only on the protein interactions characteristic of its native structure. These models are based on protein energy landscape theory[115] and focus on modelling the protein's native structure as the unique potential-energy-minimum conformation. Since natural proteins have evolved to avoid kinetic traps and maintain a single minimum-energy conformation, the simplifications made by SBMs seem sensible and allow to speed up the MD conformational search to obtain reasonable kinetical and equilibria characterisation of the protein system, to evaluate them in the framework of statistical mechanics.

SBMs can be applied to study native-like conformations and other configurations related to the formers by partial or complete unfolding. Also, by employing more than one protein configuration when building an SBM force field, such as in multi-basin SBMs, the sampling can be expanded to

explore each configuration simultaneously. These single- or multi-basin models have been applied to study protein folding, binding, and conformational changes. For a review on applications of SBMs to study protein biophysics, please see reference [117].

Current implementations of SBMs on popular simulation packages[118] either do not allow setting up custom SBM force fields or do not have hardware acceleration capabilities to run all custom SBMs with GPU acceleration. We consider these characteristics essential to maximize the efficiency of sampling protein conformations and, also, to foster SBM forcefield experimentation among the MD community.

Since the primary goal of this work is to create a framework to develop computational enzyme design methods, and since for enzymatic systems, studying protein conformations is highly relevant when considering how conformational entropy affects their catalytic parameters, we considered it not only important but necessary to expand the implementation of SBMs to include programmatic versatility coupled with GPU acceleration.

Results

Implementation of SBMOpenMM

We developed the SBMOpenMM Python library for setting up and running SBM simulations using the OpenMM toolkit for MD simulations. The OpenMM Python API offers tools to develop custom force fields that use the OpenMM GPU platforms to accelerate MD simulations. The SBMOpenMM library can set up custom SBM force fields and pre-implemented versions of popular SBM force fields, such as the SMOG force fields.[119,120] It can work with different coarse-grained protein models and implement their respective multi-basin force field versions.

The SBMOpenMM library is divided into three main classes (Figure 39). The *geometry* class allows the calculation of the equilibrium values of all bonded (i.e., bonds, angles, and dihedrals) and non-bonded (i.e., protein native contacts) terms from the protein input structure. The *system* class is the main class, where all the force field information is stored. This class contains methods to coarse-grain the protein system, set up force field parameters, and create the force and system objects to simulate them with OpenMM. Finally, the *models* class is where default

SBM models reside. It is a straightforward class to deploy SBMs ready to be simulated by the OpenMM engine, and also can serve as a starting point to create variants of these default implementations.

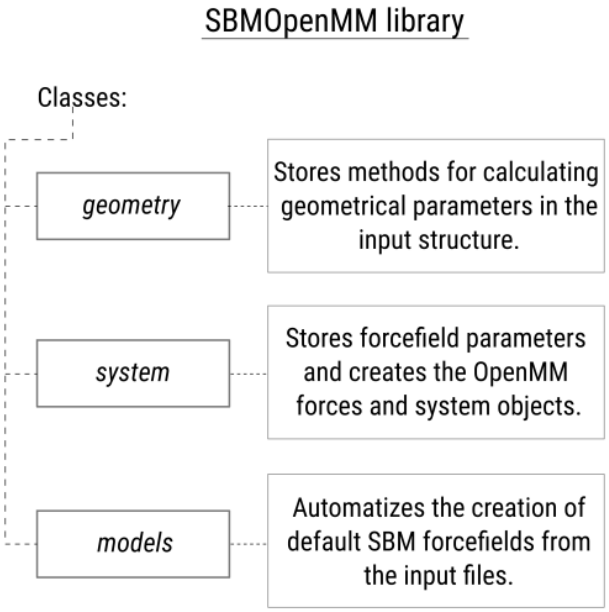

Figure 39. Structure of the SBMOpenMM classes. The program is divided into three main classes that automates the set up of SBM force fields. The system class is the main class which contains all information to create the OpenMM system object for MD simulation.

The workflow to set up SBM models (Figure 40) with SBMOpenMM starts with an input PDB (or CIF) file containing the protein native structure and a contact file containing the set of non-bonded native contacts. The library helps set up the protein topology, which will be defined according to the coarse-graining level required; currently, it can be all-atom (not including hydrogens) or CA atoms only (CA model). The library also has methods to remove any non-protein atom so that it can be employed directly using structures coming directly from the PDB database.

After loading the structure and contacts files to create an instance of the SBMOpenMM *system* class, the library allows to set up the force field parameters for each degree of freedom in the forcefield, i.e., for each bond, angle, torsion, and contact considered. Once forcefield parameters have been given, different force types can be selected, accordingly, to model the interactions that maintain the input structure as the minimum energy configuration. The final step in setting up the SBM is creating the OpenMM system class (different from the SBMOpenMM system class),

which will serve to run the MD simulation. The OpenMM system class creation method inside SBMOpenMM allows checking the magnitude of the forces acting on the system and minimizing the starting coordinates if they happen to be different from the native structure used to set up the SBM force field.

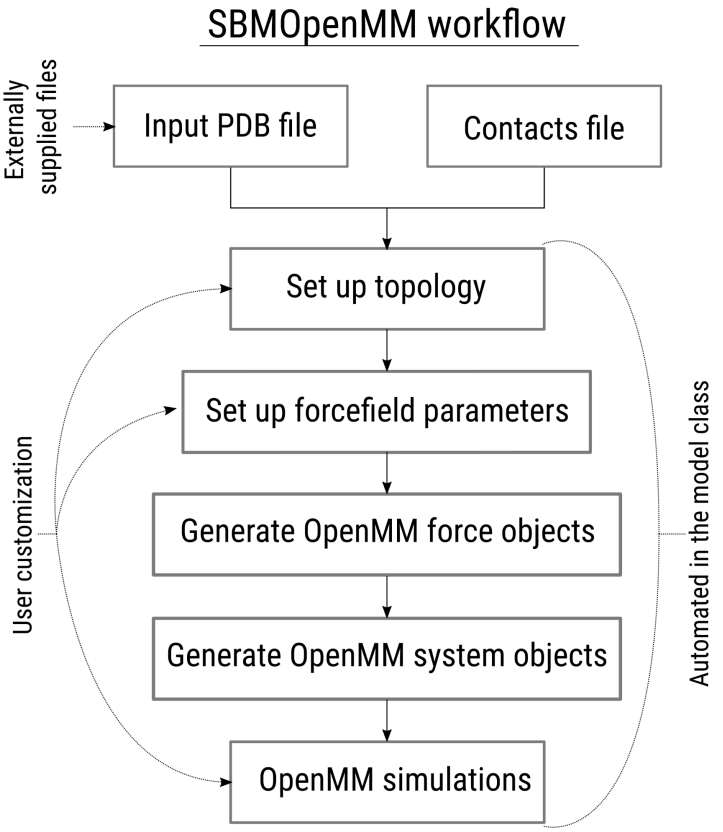

Figure 40. Workflow for setting up custom SBM force fields with SBMOpenMM. An input file is required to calculate the equilibrium values of the degrees of freedom and set up the system's topology. A set of parameters are generated to give to the different available force classes. Finally, an OpenMM system object is created and directly deployed to run the simulation. These steps are automated in the models class to set up predefined SBMs.

Please review the library's documentation for more details on the SBMOpenMM library usage: https://compbiochbiophlab.github.io/sbm-openmm/build/html/index.html, or review the related publication.[121]

Validation tests of SBMOpenMM

To validate our SBMOpenMM library implementation, we run simulations for a small protein system using files derived from a standard SBM implementation: the SMOG server.[122] The

simulations were run with the Gromacs program[123] using the all-atom SMOG force field.[119] The derived trajectories were then re-evaluated using our implemented force objects in the SBMOpenMM library employing precisely the same parameters. A step-by-step comparison between the energies derived from each program is shown in Figure 41. The test shows reliable energy calculations using the implementation of our library in OpenMM that have a very high correlation with the Gromacs derived energies (0.9960, Pearson correlation coefficient (PCC)). Slight differences between both implementations arise because of numerical and additive errors due to the employment of dissimilar numerical libraries by both MD programs. These errors accumulate across many different interactions at each energy term in the force field and result in average differences of a few kilocalories between them for the entire simulated system.

A second validation test was carried out using a set of diverse protein structures from the PDB database (Figure 41). Using the SMOG all-atom force field[119] implemented in the SBMOpenMM library, we determined the folding temperature for each protein model and ran several simulations at different relative temperatures. For each protein model and at each temperature, we calculated correlation values between their experimental crystallographic temperature factors (B-factor) and derived RMSD from the generated SBM all-atom trajectories (Figure 42).

Before reaching the folding temperature, correlations remain high (0.6605 PCC) until around 0.85Tf. After this, correlations drop abruptly at temperatures near, at, or higher than the folding temperature. They also have more significant standard deviations than temperatures equal to or lower than 0.85Tf. This result shows that the derived SBM trajectories explore configurations in agreement with the experimental dynamic profiles of the simulated proteins. Despite the good correlations found, it is expected that dynamic profiles based on protein motions, and crystallographic temperature factors, do not fully agree since B-factors are not a direct measure of protein motions, and can also have significant noise originating from the crystallization environment (for a discussion, see reference [124]).

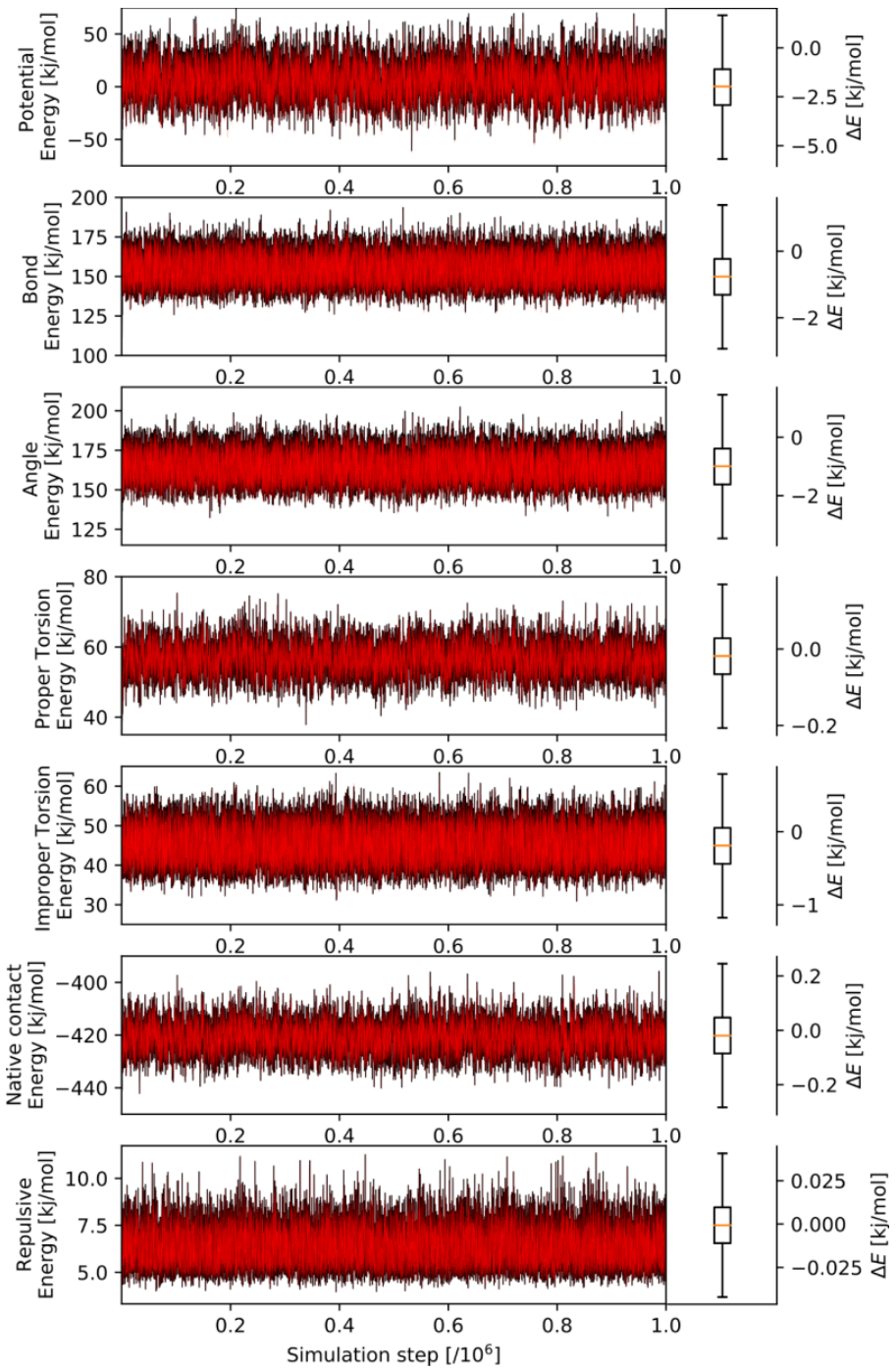

Figure 41. Energy reproducibility of the SBMOpenMM implementation. A 1 ns (SBM timescale) simulation using SMOG's AA force field was run to generate a set of probe conformations. The energies of each conformation were recalculated using the implementation of the same force objects in the SBMOpenMM library. Left-side plots show the full potential energy separated by its composing energy terms, from Gromacs (black lines) and OpenMM (red lines). The right-side boxplots show the energy differences distributions between both programs for all steps.

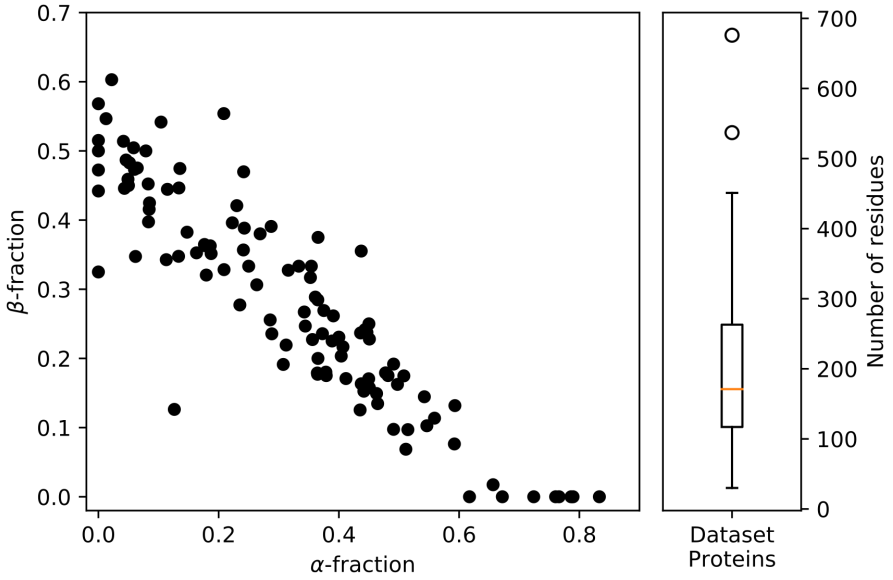

Figure 41. Composition of the structural dataset for validation of SBMOpenMM. (Left plot) Fraction of helical (α-fraction) and strand (β-fraction) character of each PDB structure (Right plot) Distribution of protein lengths in the dataset.

We decided to explore the evolution of the trajectories' RMSD to their corresponding crystallographic structures (Figure 43). This control is done to check that the previously found correlations (Figure 42) arose from the native structural basin exploration. We found that RMSD values remain low up until 0.85Tf and then start to increase with a big transition to larger values at the Tf. RMSD values remain high at temperatures of the Tf, which corresponds to the exploration of the unfolded basins. In Figure 43, an outlier explores lower RMSD values even at temperatures higher than the Tf. This structure is a thirty-residue protein containing three disulfide bonds (PDB code 3E8Y), which helps the protein remain closer to the native basin despite being at its unfolding temperature or above.

Protein folding simulations with SBMOpenMM

To test our SBM implementation for gaining insight into biophysical phenomena, we studied the folding process of a small 88-residue polypeptide. We ran several trajectories at the folding temperature of the protein model, and we analyzed the resulting trajectories employing a Markov State Model (MSM) framework.

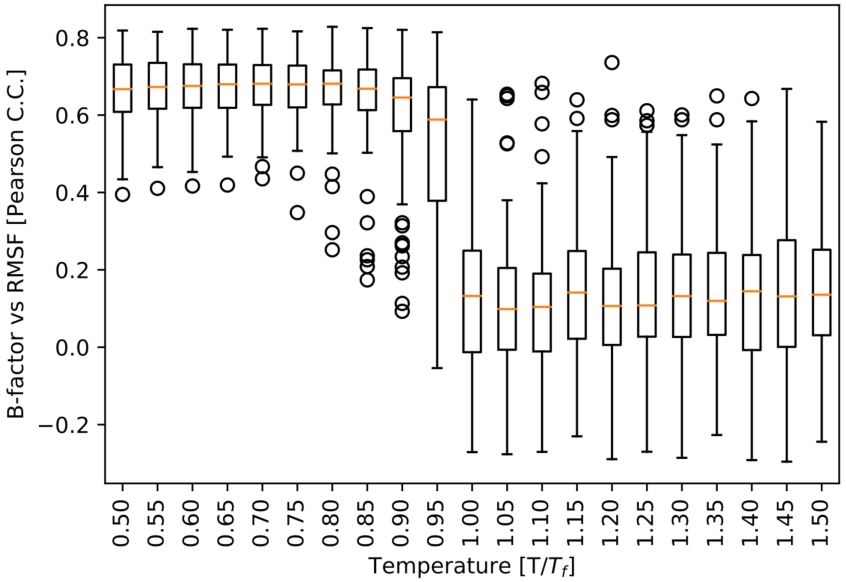

Figure 42. Dynamic and B-factor correlations for AA SBM simulations. Simulations from the 110 structures in the validation dataset were run at different temperatures. The boxplot showS the distributions of PCC between the root-mean-square fluctuations (RMSF) values and crystallographic temperature factors (B-factors).

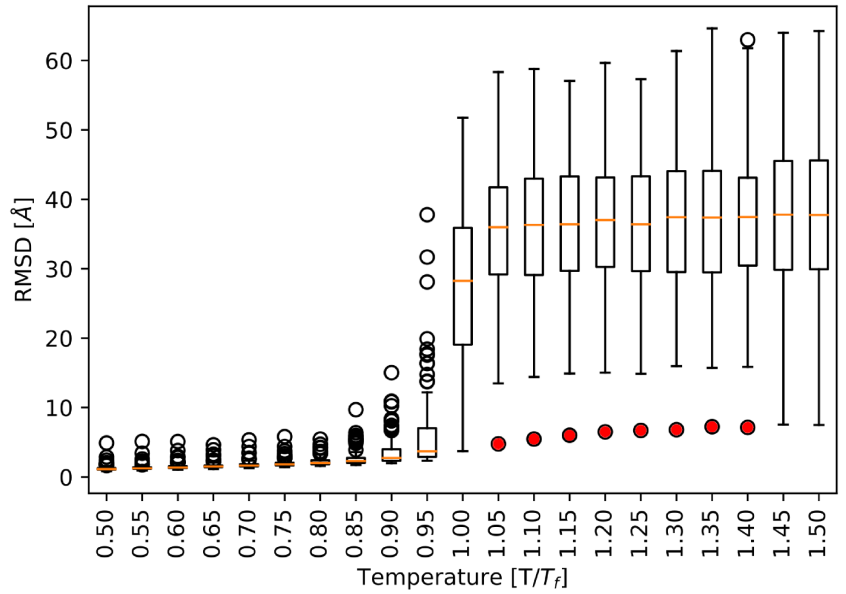

Figure 43. RMSD distributions for AA SBM simulations. The boxplot shows the distribution of RMSD values regarding the native structure at different relative temperatures for the 110 structures in the validation dataset

The protein structure is of FoxP1 (PDB code 2KIU), a DNA transcription factor, and was simulated with the default all-atom model included in the *models* class of the SBMOpenMM library.[121] The Python code to simulate this system can be consulted in Appendix 2. The code allows us to quickly set up a simulation object and use the OpenMM engine to propagate the system's dynamics.

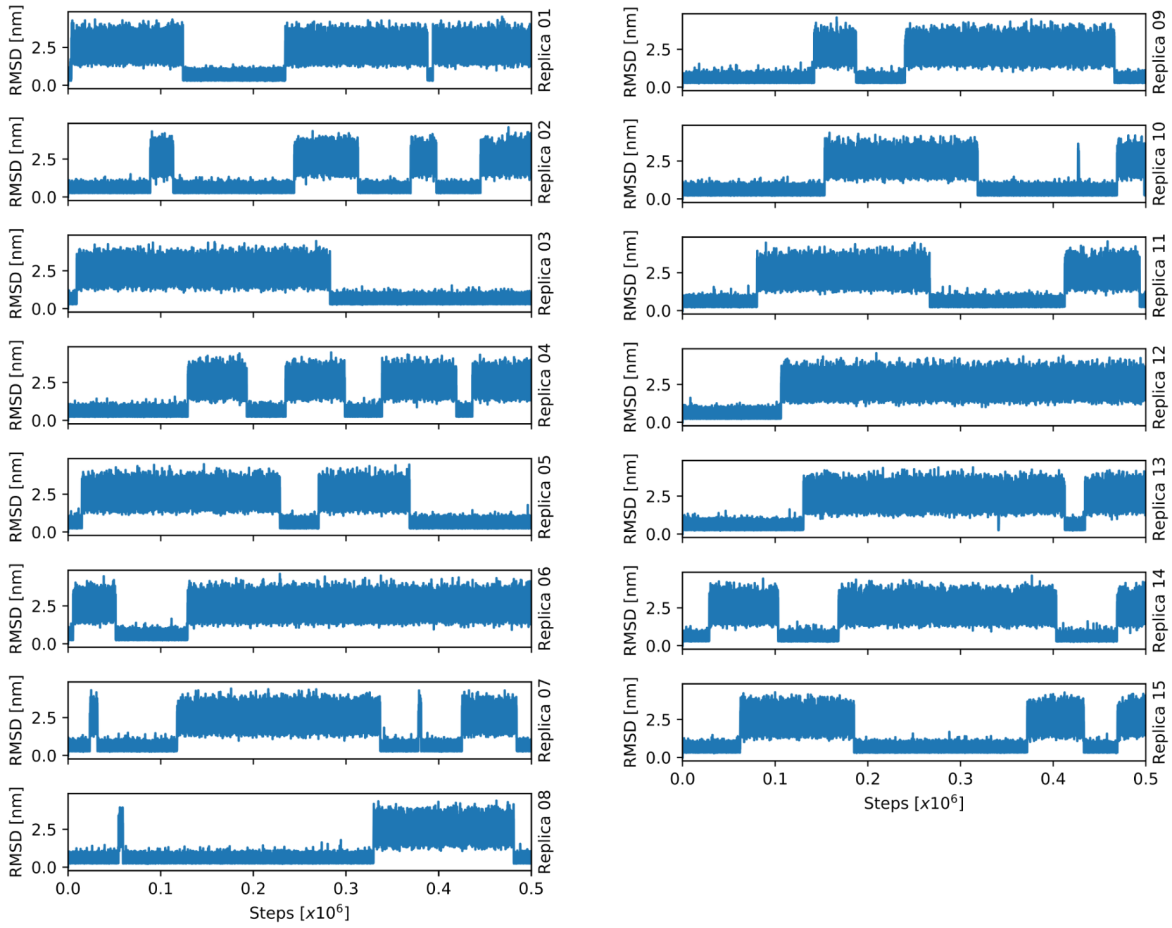

Figure 44. AA SBM folding simulation replicas. Each plot shows the trajectory's RMSD to the native structure for the 15 replicas that were run for the FoxP system. The folded and unfolded configurations are represented by low and high RMSD values, respectively.

For the FoxP system, we ran 15 simulations at the folding temperature, which allowed us to sample its folded and unfolded basins, and transitions between them (Figure 44). The different simulated replicas show several transitions between the folded (low RMSD to the native structure) and unfolded basins (high RMSD values to the native structure), and vice versa, ranging from two to eight. This variability among replicas indicates independence in the phase-space sampled between them.

Exploring SBMOpenMM folding simulations with a Markov State Model framework

To gain insights from the simulated data, we performed a MSM analysis to uncover characteristics of the folding events. The complete set of trajectories was processed using the PyEMMA library[125] (for details see the "MSM validation and construction" section in Methods).

The simulation data was featurized using only the CA distances, since we found this set described better the slow-order kinetics of the full simulation. We diminished the dimensionality of the data by projecting it into a Time-structure independent components (TICA) space[126] in which the folding process was clearly characterised as two configurations (i.e., folded and unfolded). The free energy of the process is shown in Figure 45, in which the left basin pertains to the folded configuration and the right to the unfolded one. Increasingly extended conformations are found inside the folded basin from left to right, and, in the unfolded one, from up to down (see Fig. 34). A fairly sampled TS region connects these two basins.

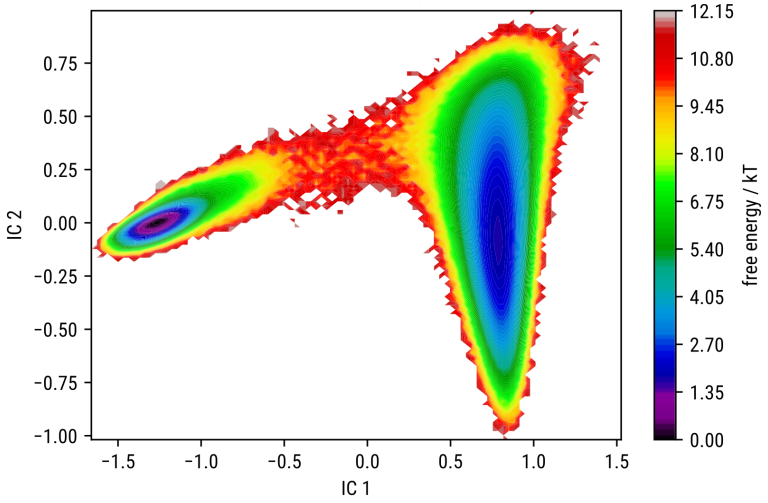

Figure 45. Free energy surface for the FoxP simulation for the two slowest TICA dimensions. The left basin corresponds to the folded configuration and the right to the unfolded one.

To better define the extent of each configuration, we constructed a MSM model[127] by clustering conformations over the sampled TICA space using a k-means algorithm with a thousand discrete sub-states. Then, the PCCA++ algorithm[128] was employed to coarse grain these sub-states into two kinetically-relevant metastable states, establishing clear boundaries for the folded and unfolded configurations (see Fig. 53). Transition Path Theory,[129] applied to study the transitions between the defined states, allowed us to describe a clear boundary to unambiguously define the

folding/unfolding process TS region (see Fig. 56). We estimated an activation free energy of 15.42 KT by counting conformations at this region, which translates to a mean first-passage time of 0.50 μs, in agreement with a folding process characteristic of fast-folding proteins.[130]

After defining conformations belonging to the TS region of the folding simulation of FoxP, we plotted the probability of contact formation for the folded, TS, and unfolded configurations (Figure 46). In the folded configuration at the folding temperature, most native contacts can be formed with reasonable probability (although not necessarily simultaneously), except for N-terminus interactions with the C-terminus. At the other extreme, in the unfolded configuration, native contacts are seldomly formed, and, if formed, they are very close in primary structure. At the TS configuration, we observe structures characterised by the formation of native contacts between the protein's beta-sheets, pointing to a folding mechanism in which this secondary structure is the last to unfold.

Discussion

We have written a new library to construct SBM force fields that build upon the versatility of the OpenMM library to set up customized force objects that can run on hardware accelerated platforms. Given the lower number of computed interactions and the absence of explicit solvent to represent the system, SBM force fields converge notably faster than MD simulations, allowing to explore conformational dynamics with the computational resources available in a present-day personal computer. These simplifications expand the simulation possibilities to studying biomolecular processes with significant energy barriers, such as large conformational rearrangements or protein folding.

The library was successfully benchmarked to reproduce the same energies as standard programs in the field.[118] Additionally, a dataset of structurally-different proteins showed that simulations carried out with the SBMOpenMM library describe the folded basin in a way that correlates with the crystallographic dynamic profiles of the tested proteins. This result is significant since it opens the possibility to study the interactions responsible for the conformational dynamics at the native configuration. However, it is important to note that no alternative near-native configuration can be discovered straightforwardly from simulation using a single-basin SBM potential.

Nonetheless, the sampling carried out with this type of potential can seed other simulations, using more apodictic force fields that take into account non-native contacts or explicit solvent, in a strategy known as adaptive sampling.[131]

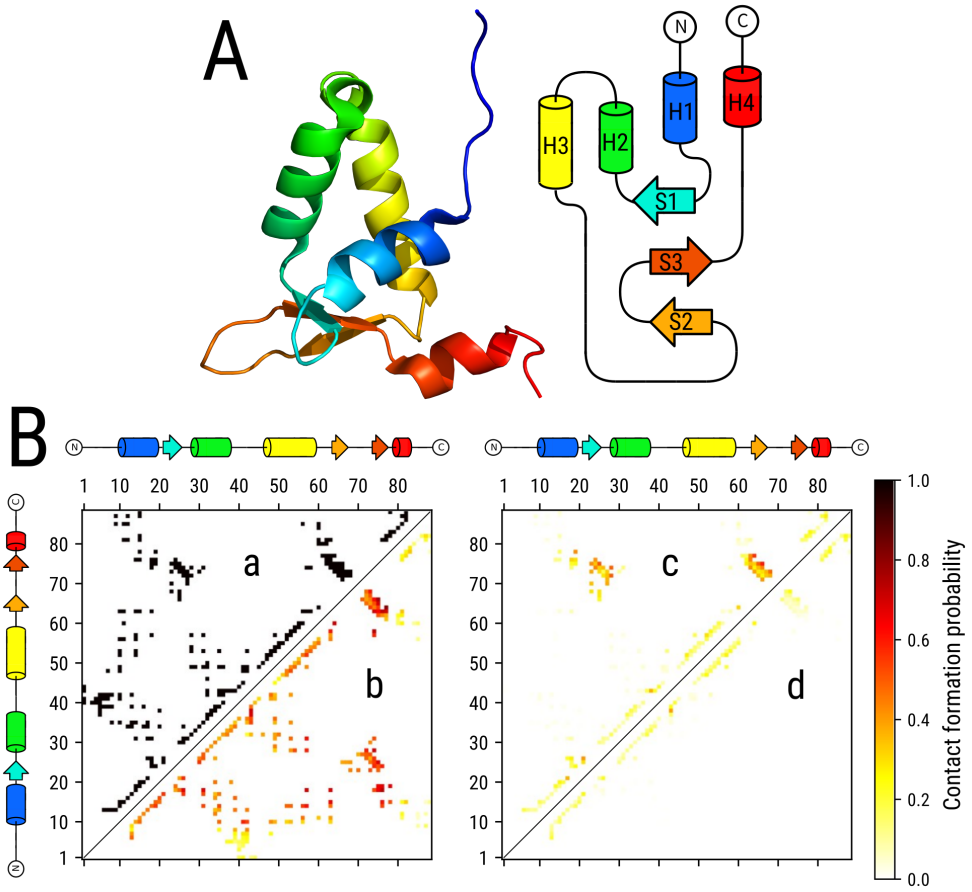

Figure 46. Contact formation probability for different configurations of the FoxP1 protein. (A) Tertiary structure (left) and topology connectivity (right) of the FoxP1 system. (B) The native contact map of FoxP is shown as a reference (a). Probability of contact formation for the folded (b), TS (c), and unfolded (d) configurations of the FoxP protein. The secondary structure of residues is indicated in the plots' axes.

We exemplified the use of our library by studying the two-step folding process of the small FoxP protein. The kinetic characterisation of this simulation was carried out with an MSM analysis of the folding process. Combining the two methodologies clearly defines the folded and unfolded configurations and the TS region connecting them. Once defined, many system observables can be calculated to help make direct comparisons with experimental results, enriching the practicality and relevance of employing SBM simulations to study complex biophysical phenomena.

SBMs are specially used to understand the restrictions that the structure's topology plays on the dynamic behaviour of the protein. Typically, additional force field terms are added to the SBM to assert their effect by studying the differences from applying a purely topological behaviour. The SBMOpenMM library can help this process by facilitating the programmatic deployment of alternative SBM force fields within an efficient and open-source MD simulation package.

In the problem of enzyme design, we previously observed a need to explore the full effect that conformational entropy has over the catalysed reaction (see Chapter 1 - Enzyme optimisation: the Kemp Elimination case). MD simulations are very costly in this regard since they converge very slowly over the sampled phase-space, and alternative methods are needed to sample protein configurations more efficiently. This cost is substantial when several design proposals need to be evaluated to assess their catalytic capabilities. Due to their fast convergence character, SBM simulations can be a first approximation towards this goal; however, if the native fold is ill-defined, conformations extracted from this exploration could be severely biased. Therefore, conformations explored directly from the designed models' structures need to be evaluated with an alternative force field that considers the full extent of the system's interactions. The Rosetta force field[47] could be applied here to re-evaluate the sampled SBM energy landscape and discover conformations closer to the true-native and other relevant near-native configurations, thus, improving the overall protocol of catalytic prediction.

Conclusion

SBM simulations are a relevant methodology for the study of biophysical phenomena. The SBMOpenMM library, here developed, can help deploy tailor-made force fields that can be run efficiently using the OpenMM platform. These SBM can help in the enzyme design framework by providing a fast-converging tool that, in conjunction with others, can aid in assessing the effect that conformational entropy plays over the activation free energies of designed variants.

Methods

SBM AA force field

The SBM forcefield employed corresponds to the SMOG all-atom SBM forcefield[119] as implemented in the SBMOpenMM program.[121]. The potential energy function of this force field is defined as follows:

$$H_{AA} = \sum_{bonds} V_{bond} + \sum_{angles} V_{angle} + \sum_{torsions} V_{torsion} + \sum_{impropers} V_{improper} + \sum_{planars} V_{planar} + \sum_{contacts} V_{LJ_{12-6}} + \sum_{non-contacts} V_{LJ_{12}} \tag{23}$$

Bonded terms of potential energy are defined as:

$$V_{bond} = \frac{k_b}{2}(r - r_0)^2 \tag{24}$$

$$V_{angle} = \frac{k_a}{2}(a - a_0)^2 \tag{25}$$

$$V_{torsion} = k_t(1 - cos(\phi - \phi_0) + \frac{1}{2}(1 - cos(3(\phi - \phi_0))) \tag{26}$$

$$V_{improper} = \frac{k_i}{2}(\phi - \phi_0)^2 \tag{27}$$

$$V_{planar} = \frac{k_p}{2}(\phi - \phi_0)^2 \tag{28}$$

Here, r_0, a_0, and ϕ_0 are the equilibrium distance values for the bond, angle, and torsion (improper) terms, respectively. k_b, k_a, k_t, k_i, and k_p are the force constants of the bond, angle, torsion, improper, and planar terms, respectively. Finally, r, a, and ϕ represents the current value of the bond, angle, and torsion (improper) degrees of freedom, respectively.
Non-bonded terms are defined as:

$$V_{LJ_{12-6}} = \epsilon_c((\frac{\sigma_{ij}}{r})^{12} - 2(\frac{\sigma_{ij}}{r})^6) \tag{29}$$

$$V_{LJ_{12}} = \epsilon_{nc}(\frac{\sigma_{ij}}{r_{ex}})^{12} \tag{30}$$

The V_{LJ12-6} term describes a Lennard-Jones potential that represents the interaction for the native contacts in the system. On the other hand, the V_{LJ12} term defines the volume at which atoms cannot cross each other, and is defined for all non-native contact interactions. The σ_0 values are the equilibrium distances of the non-bonded interactions, r_{ex} is the excluded radius of each particle, and ϵ_c and ϵ_{nc} are the force constants of the native and non-native interactions, respectively.

AA SBM simulation parameters

SBM simulations were run with the SBMOpenMM program[121] using the SBM all-atom force field. The force constant values employed for the harmonic terms were:

$$k_b = 10000\frac{kj}{mol \cdot nm^2} \tag{31}$$

$$k_a = 80\frac{kj}{mol \cdot rad^2} \tag{32}$$

$$k_i = 10\frac{kj}{mol \cdot rad^2} \tag{33}$$

$$k_p = 20\frac{kj}{mol \cdot rad^2} \tag{34}$$

The energy constant ϵ_c for native contacts depends on the number of atoms (N_a) and native contacts (N_c) in the system as:

$$\epsilon_c = \frac{2N_a}{3N_c}(\frac{kj}{mol}) \tag{35}$$

The torsional constant is defined according to the number of atoms (N_a) and the number of proper torsions (N_t) in the system as:

$$k_t = \frac{N_a}{N_t}(\frac{kj}{mol}) \tag{36}$$

However, the torsional energy constant is used unequally for torsions with a backbone (k_t^{BB}), or side-chain-only (k_t^{SC})component, according to the following rule:

$$k_t^{BB} = \frac{2}{3}k_t \tag{37}$$

$$k_t^{SC} = \frac{1}{3}k_t \tag{38}$$

The excluded volume force constant and excluded radius were set to:

$$\epsilon_{nc} = 0.1\frac{kj}{mol} \tag{39}$$

$$r_{ex} = 2.5\mathring{A} \tag{40}$$

In a SBM, by definition, all equilibrium values were defined from the native structure.

All units employed here are compatible with the OpenMM framework, however they do not translate to real physical units, since the SBM is not calibrated to model the interactions of a solvated system.

Folding temperature determination

Folding temperatures (T_f) were determined from short 10ns (SBM timescale) simulations run at different replicas. The T_f is defined as a maximum in the heat capacity (C_v) temperature profile, which was calculated from the whole ensemble of simulations as follow:

$$C_v = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T^2} \tag{41}$$

Here, $\langle E \rangle$ is the ensemble averaged potential energy, k_B is the Boltzmann constant , and T is the temperature at which the conformations were obtained. Calculations of the C_v profiles were carried out with the PyWham program.[132]

Validation dataset

Single chain protein structures were searched in the PDB database. Proteins that contained small molecules, ions, or a missing structure for internal residues (i.e., having chain breaks) were filtered out. Proteins with significant bias in the B-factor profiles due to crystal packing were also removed. To calculate this bias, we used the WCN metric, calculated as the sum of the inverse distances (r_{ij}) from atom i to all other atoms (j) in the structure:

$$WCN_i = \sum_{j \neq i}^{N} \frac{1}{r_{ij}^2} \tag{42}$$

The inverse of the WCN profile (WNC^{-1}) correlates well with the B-factor profile of crystallographic protein structure.[96] Therefore, we used the WNC profiles derived from the single-chain structure alone (WNC_{SC}) and the one derived using the structure of the protein plus all symmetry-related copies of the protein chain that were closer than 5 Å to it (WCN_{SR}) to estimate the crystal packing bias. Both profiles were correlated to the B-factor profile, and we kept protein structures if the change in the correlation value did not change more than 10%. A final set of 110 proteins were thus selected, whose codes are:

1ES5, 1K9Z, 1LC5, 1O4Y, 1OA4, 1OOT, 1P3C, 1QAU, 1QRE, 1S2O, 1TJE, 1ULR, 1UQ5, 1VF8, 1VKK, 1X6X, 1XQO, 1XUB, 1Y0M, 1Z6N, 1ZBF, 1ZHV, 1ZZK, 2CAL, 2E0Q, 2FJ8, 2H1V, 2I49, 2J8B, 2NUH, 2P51, 2QT4, 2VC8, 2WNX, 2X5Y, 3A2Z, 3ALF, 3DFJ, 3E8Y, 3H79, 3H7I, 3K01, 3K6I, 3KBF, 3MBR, 3N79, 3O5O, 3PBC, 3PO8, 3PZ9, 3RJP, 3RKG, 3RT2, 3RVM, 3VQF, 3VZ6, 3W43, 4B89, 4CG0, 4DMV, 4DW8, 4ETL, 4FK9, 4IC4, 4ICV, 4IGV, 4J5Q, 4JF8, 4K0G, 4NPD, 4OOX, 4OUS, 4R6H, 4RWU, 4TO7, 4U94, 4W65, 4W7U, 4WDC, 4XQ1, 4YAP, 4ZC3, 4ZMK, 4ZOT, 5DXW, 5ECA, 5EPF, 5ESR, 5F68, 5FJL, 5H0Q, 5H9K, 5HPJ, 5HQH, 5I4I, 5IDV, 5IWH, 5JW8, 5LQ5, 5M1M, 5MPV, 5OJZ, 5OUO, 5XMO, 5Y4M, 5YDE, 5Z8P, 5ZU6, 6AIB, 6AR0, and 6F47.

SBM simulations

Simulations for the validation dataset were run with the all-atom SBM forcefield. To calculate the systems' folding temperatures (T_f), short constant-temperature simulations of 10 ns (SBM timescale) were run spanning a wide range of temperatures. Then, using the T_f, longer 100 ns simulations were run at different relative temperatures, from $0.5T_f$ to $1.5T_f$, with a temperature step of $0.05T_f$.

For the folding simulations with the FoxP system, the FoxP structure was downloaded from the PDB database[65] with code 2KIU. From all NMR structures, the RMSD-centroid structure was selected as the native conformation. A total of 15 replicas of 10 μs (SBM timescale) each were run at the folding temperature of the system employing the all-atom SBM forcefield.

All simulations were calculated using the OpenMM program.[133]

Free energy calculations

Thermodynamics free energy values were calculated directly from probability values, using the equation:

$$A_i = -k_B T ln \sum_{j \in S_i} \pi_j \tag{43}$$

Here, A_i is the free energy of the thermodynamic state S_i, k_B is the Boltzmann constant, T is the temperature, and π_j is the probability of the system of being in the sub-state j that belongs to state S_i.

The definition of the S_i states, representing configurations, and the j sub-states, representing clusters of conformations, were assigned according to an MSM analysis of the simulation trajectories.

Native contact formation probability

Contact formation probability was calculated for a particular state by counting the number of times a particular native contact was formed at each conformation sampled belonging to that

state. We defined a contact being formed if the distance between the atoms was less or equal than the equilibrium native contact distance times a factor of 1.05.

MSM validation and construction

The trajectory data was featured using different structural descriptors to query which best represents the kinetic information contained in the simulation. We used the VAMP-2 score[126] as a heuristic, and calculated its value for each feature at a lag time of 660 ps (Figure 47). The better the set of features approximate the dynamic process, the higher the VAMP-2 score. Thus, the set of CA native contacts distances was the best performing feature and, therefore, was selected for further analysis.

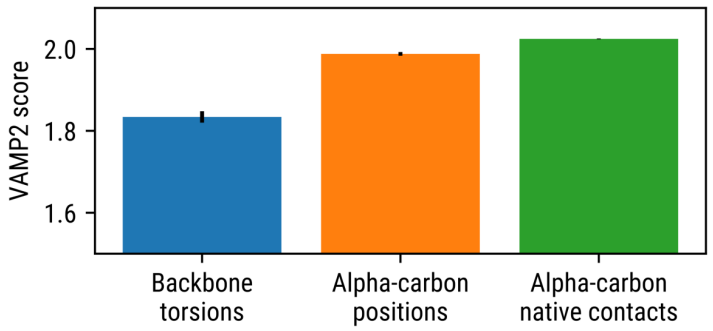

Figure 47. FoxP1 folding simulation featurization. A variational approach to the Markov process (VAMP) dimensionality reduction method was applied to discern different featurization schemes at a lag time of 660ps. Error bars represent the standard deviation of 10 cross-validated estimations.

For better kinetic interpretability of the featured simulation data, we diminished its dimensionality by using a time-structure independent component analysis (TICA).[134] To observe the dependence of our TICA analysis with the simulation lag time, we plotted the number of dimensions needed to explain at least 95% of the kinetic variance (Figure 48). When the lag time increases, we observe a diminishing number of dimensions, approaching only one representative TICA dimension at considerably large lag times. This decrement is indicative of a simulation in which one process dominates the kinetic behaviour when long time scales are considered. A lag time of 660 ps was finally selected, since at this point, most of the kinetic data were contained in only two TICA dimensions, and it was low enough to resolve other, although slower dynamical processes.

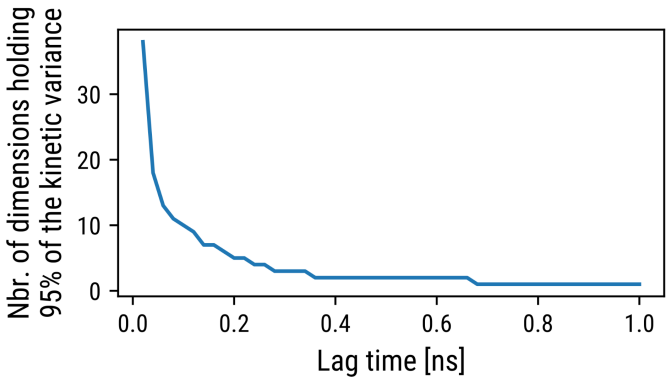

Figure 48. Number of TICA dimensions necessary to explain 95% of the FoxP1 folding simulations' kinetic variance as a function of the lag time.

When the simulation is projected into the two slowest TICA dimensions (IC1 slower than IC2), we observe two minima populated (Figure 49). The IC1 dimension correlates with the folded to unfolded transitions (and vice versa) seen in the independent replicas. On the other hand, the IC2 dimension describes processes of internal variability in both minima, although with more significant deviations inside the unfolded configuration.

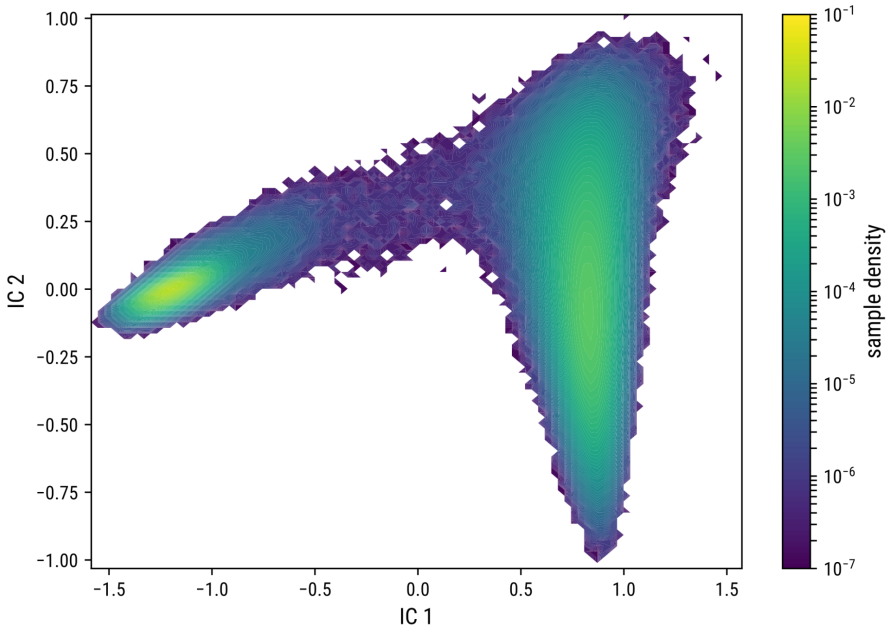

Figure 49. Joint distribution of the two slowest TICA dimensions for the FoxP simulation data.

To build an MSM, we first partitioned the two-dimensional TICA space into discrete clusters using the k-means algorithm together with the TICA-projected simulation data. A thousand clusters were selected (Figure 50) to describe the MSM discrete states by a Bayesian MSM estimator.[135]

The final MSM was validated by analyzing the implied timescales (ITS) at different lag times coming from the eigenvalue decomposition of the MSM transition matrix (Figure 51). Timescales below the black line (shaded grey area) occur faster than the lag time selected to describe the MSM transition matrix and, therefore, cannot be correctly described by the analysis. At a lag time of 660 ps, the only time scale being properly resolved is the folding and unfolding event (blue curve), which is already converging at this lag time. We confirmed the lag time selection by this analysis and focused further analysis in describing the two step folding process of FoxP1.

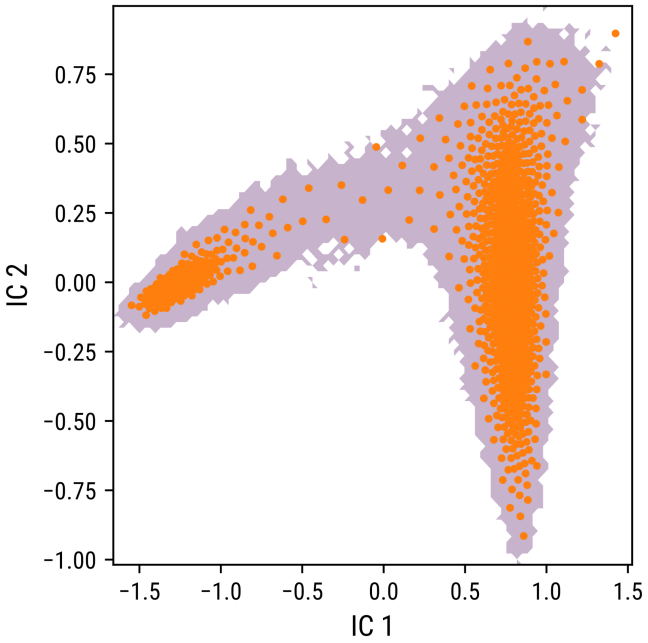

Figure 50. Clustering of FoxP simulation TICA space. The sampled data was clustered in a thousand k-means clusters (orange dots) using all TICA coordinates.The TICA surface sampled by the simulation is shown in the background in gray.

Since our interest is in the folding to unfolding reaction, we plotted the values of the second right eigenvector, which corresponds to the slowest ITS in the MSM transition matrix (Figure 52). We observe that the values of this eigenvector indicate transitions that occur between the left- and right-located minima (according to the IC1 dimension). We partitioned the sampled phase-space into two configurations (i.e., the folded and unfolded configurations) by coarse-graining the MSM clustered space into two metastable clusters using the PCCA++ algorithm.[128] A crisp partitioning of the two states can be observed in Figure 53.

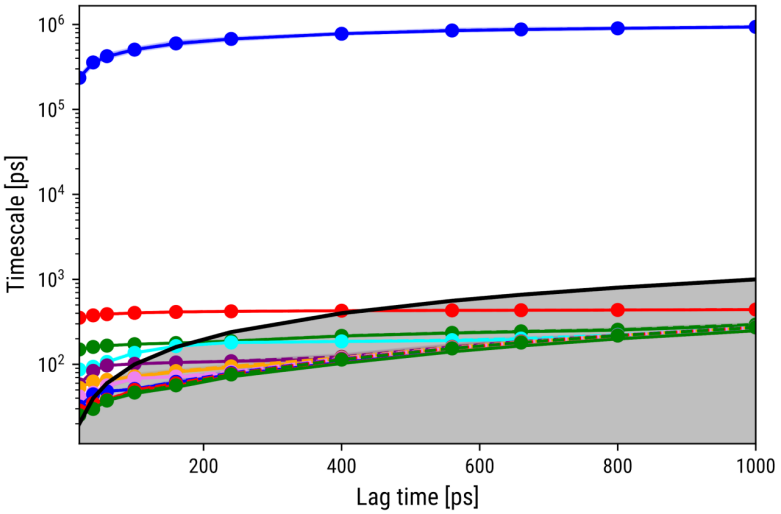

Figure 51. Implied time scales analysis for the FoxP folding simulation. The 10 slowest implied time scales are plotted as a function of the lag time selected to build the MSM transition matrix. Only processes with decorrelation times above the lag time scale (black line) can be correctly resolved by the MSM.

To characterise the metastable partitioning, we plotted the distribution of radius of gyration for the corresponding conformations contained in each state (Figure 54). The two metastable states correspond to a folded and an unfolded configuration. The free energy of the process is indicated in Figure 45 and shows the folded minima at the left of the IC1 TICA dimension and the unfolded one at the right. Significant dispersion of conformations can be observed in the unfolded configuration, which is separated by the IC2 dimension. We plotted the value of the radius of gyration into the two TICA dimensions to gain insight into the compactness of the conformations inside each basin (see Figure 55). The radius of gyration increases slightly in the folded configuration towards the TS region of the two minima. On the other hand, at the unfolded basin, the structures are less compact, with a notorious trend to be highly extended when moving down the IC2 dimension. Despite the unfolded configuration's minima having a large variability in compactness, the fully extended conformations are seldomly sampled because of the large entropy barrier associated with fully extending a polypeptide chain.

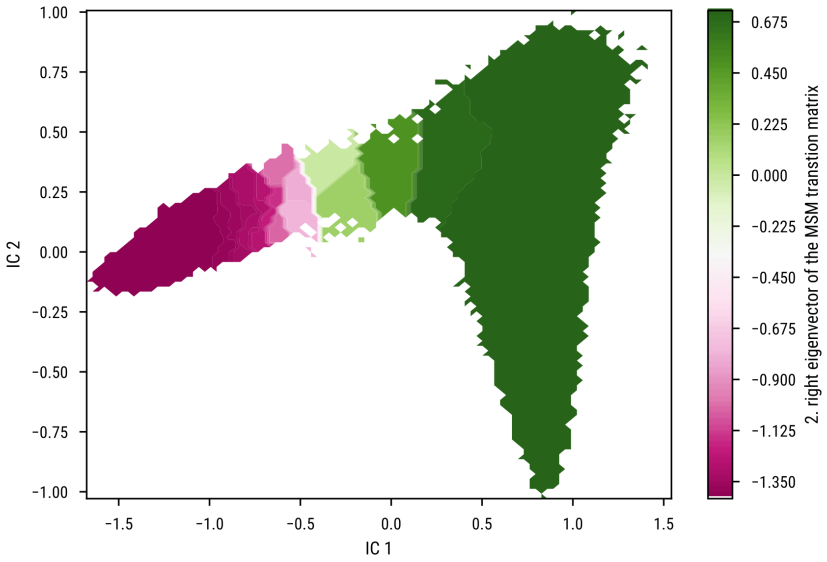

Figure 52. Second right eigenvector of the MSM transition matrix diagonalization. The eigenvector's change in sign indicates the shift between MSM states-regions.

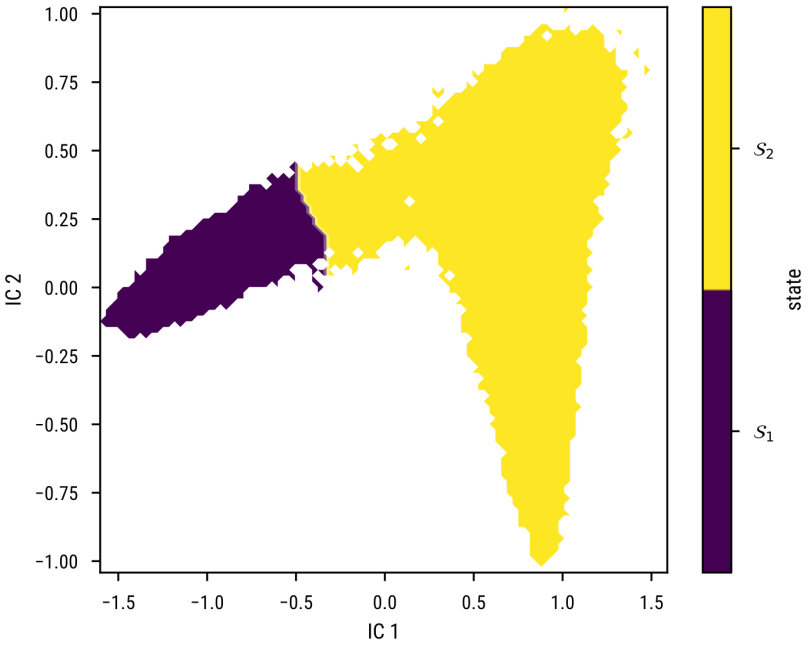

Figure 53. Coarse grained partitioning of the MSM state-space by the PCCA++ algorithm. These two states reflect the partitioning according to the second right eigenvalue, therefore, characterising the slowest process in the simulation.

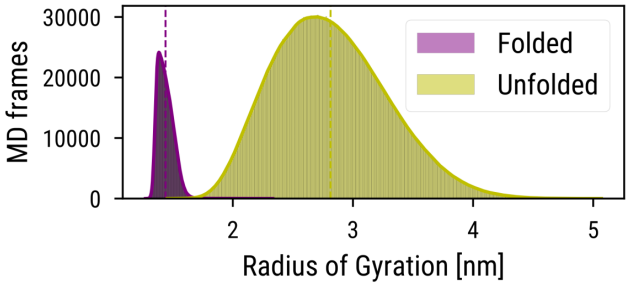

Figure 54. Distributions of radius of gyration values for the two metastable states into which the FoxP simulation-MSM was coarse-grained.

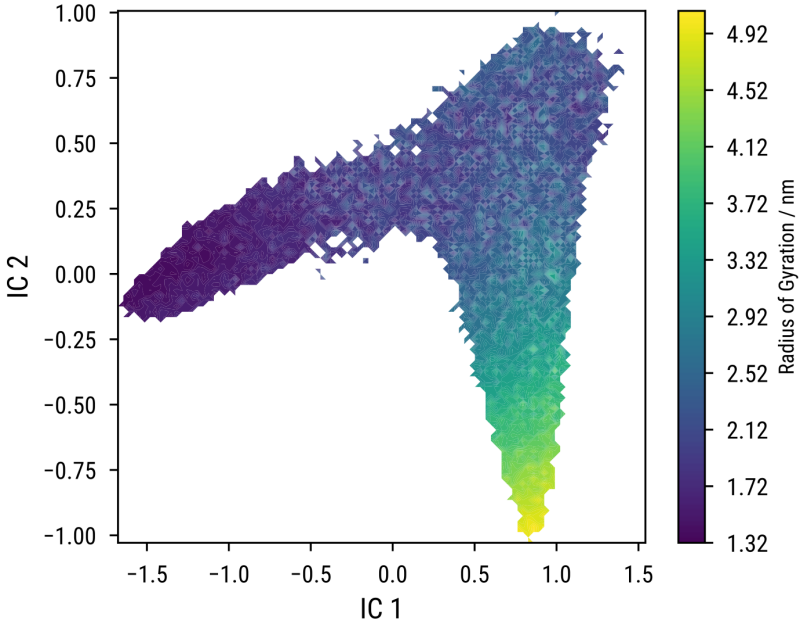

Figure 55. Distribution of radius of gyration for all conformations in the FoxP folding simulation projected into the slowest TICA dimensions.

We wanted to map precisely the configurations belonging to the TS of the folding reaction. To achieve this, we applied Transition Path Theory to define the committor function[129] for the folding process (in the unfolded to folded direction, Figure 56). Defining the TS hypersurface as the region where the commitor function has a value of 0.5, we estimated the convergence of our simulation by estimating the activation free energy of the process at different simulation lengths and counting configurations at different distances from the TS dividing surface (Figure 57). The analysis shows that the simulation has converged regarding the length of the simulated time for

each replica. By repeating the analysis taking the limit of the distance to the TS hypersurface, we estimated an activation free energy of 15.42 KT.

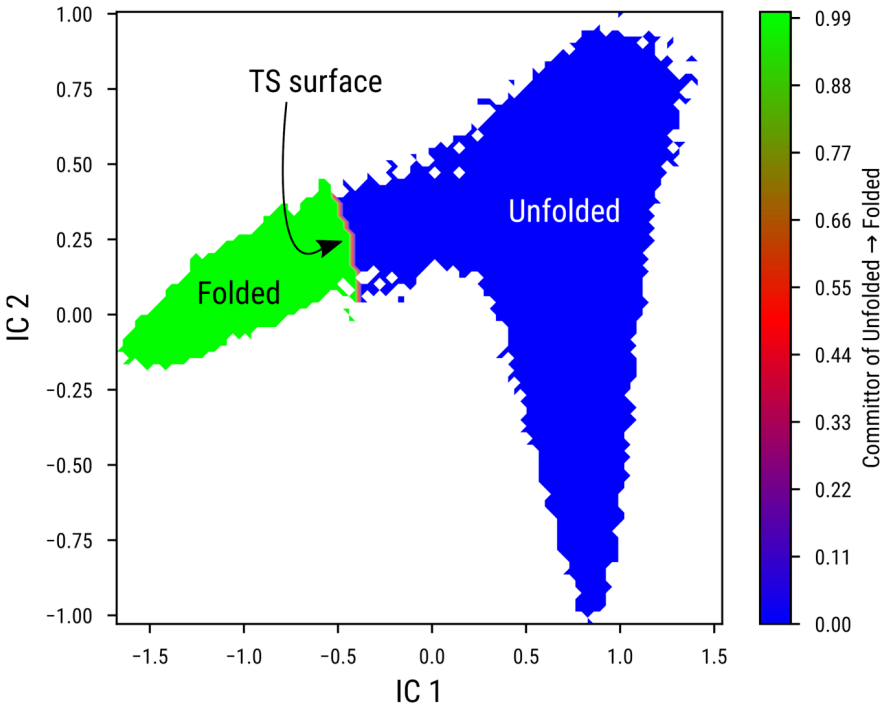

Figure 56. Committor function for the folding transition of FoxP1. The unfolded configuration (source state) has a value of 0, the folded configuration (sink state) has a value of 1, and the TS state region (between both states) is defined at a value of 0.5 for the commitor function.

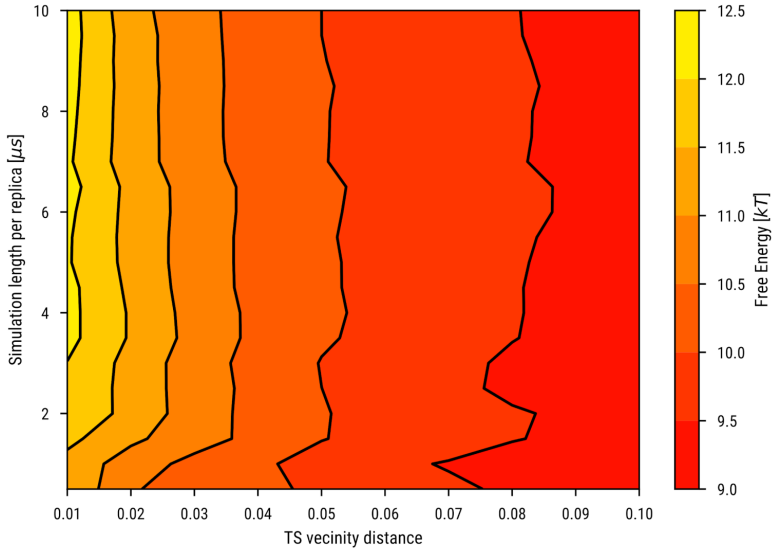

Figure 57. Activation free energy convergence as a function of the simulation time and distance to the TS in TICA space. Simulations were truncated at different time lengths, and configurations at distances lower than d from the TS region are considered to calculate activation free energies.

General Conclusions

1. In this thesis work, we have approached the problem of computational enzymatic improvement by developing a computational framework for enzymatic design and catalytic evaluation. We conclude that merging two state-of-the-art methodologies, like Rosetta for computational enzyme design and the EVB framework for catalytic assessment, is an excellent approach to address this problem since both methods complement each other by working at different protocol stages.

1. As a test case for applying our proposed methodology, we employed a resurrected ancestral beta-lactamase scaffold in which *de novo* catalytic activity for the KE reaction was designed in a secondary active site. We conclude that to correctly model the system's interactions, a first validation of the Rosetta score function is needed before employing it to obtain proper binding energies. Additionally, a flexible backbone approach is essential to produce a diversified set of catalytic proposals.

1. A sampling strategy was devised to rank the designed models before assessing their catalytic activities with the EVB method. This strategy was based on a local sampling of the enzyme conformational space, followed by calculating binding free energies based on a Boltzmann distribution of the sampled space. Despite not correlating well with the EVB catalytic assessment step, we confirmed its utility in a separate benchmark that used MHC-I-peptide complexes as model systems for protein-protein interactions. The success of this strategy in predicting experimental activities will be valuable for vaccine design efforts and, with a more in-depth exploration of its applicability, it could serve as a fast tool for ranking protein and enzymatic designs.

1. The EVB simulations were first validated to reproduce the reaction energies of many Kemp eliminase enzymes. This validation showed that EVB simulations could be a good tool for ranking enzymatic designs, although predicting absolute binding energies proved to be a more challenging task. Because of the latter problem, we could not be sure if the design algorithm successfully created improved enzymatic variants.

1. EVB simulations not only served us to rank the enzymatic designs. Since the assessment was based on molecular simulations of the reactive system, it aided in the physical interpretation of the catalytic phenomena. These simulations confirmed that electrostatic preorganisation could significantly affect the lowering of activation free energy barrier and that fast metrics to predict residue preorganisation could be applied to improve the design algorithm stage. In this regard, we validated the weighted contact number (WCN) metric to predict protein dynamic profiles, and because of its fast computation, it could be a good candidate metric for assessing residue preorganisation during protein design optimisation.

1. EVB sampling was carried out over a single configuration of the designed models. However, enzymatic systems can adopt other configurations relevant for predicting correct catalytic values. This unaccounted dynamics could be a reason why it was challenging to predict absolute activation free energies. Since traditional MD methods are very costly for protein configurational sampling, we have developed SBMOpenMM, a new library for setting up structure-based force fields, that can address this sampling problem in a more simplified and GPU-accelerated manner.

Still, many challenges lay ahead in the landscape of enzyme optimisation, with the catalytic and Michaelis constants as the main factors dominating enzyme proficiency. However, other parameters have essential roles when addressing enzyme engineering. Catalytic activity profiles depend on temperature, pH, pressure, solvent composition, and other environmental variables that are important to consider when generating new enzymes to be deployed for specific applications. Other scenarios can also include substrate promiscuity, multi-step reactions, product inhibition, and other properties of interest.

References

1. Ulusu, N. N. Curious Cases of the Enzymes / Neobiča Istorija Enzima. *J. Med. Biochem.* **34**, 271–281 (2015).
2. Carrea, G. & Riva, S. *Organic Synthesis with Enzymes in Non-Aqueous Media*. (John Wiley & Sons, 2008).
3. Singh, R. S., Singhania, R. R., Pandey, A. & Larroche, C. *Advances in Enzyme Technology*. (Elsevier, 2019).
4. Warshel, A. & Bora, R. P. Perspective: Defining and quantifying the role of dynamics in enzyme catalysis. *J. Chem. Phys.* **144**, 180901 (2016).
5. Pauling, L. Molecular Architecture and Biological Reactions. *Chemical & Engineering News Archive* vol. 24 1375–1377 (1946).
6. Amyes, T. L. & Richard, J. P. Specificity in transition state binding: the Pauling model revisited. *Biochemistry* **52**, 2021–2035 (2013).
7. Mader, M. M. & Bartlett, P. A. Binding Energy and Catalysis: The Implications for Transition-State Analogs and Catalytic Antibodies. *Chem. Rev.* **97**, 1281–1302 (1997).
8. Agarwal, P. K. A Biophysical Perspective on Enzyme Catalysis. *Biochemistry* **58**, 438–449 (2019).
9. Warshel, A. *et al.* Electrostatic basis for enzyme catalysis. *Chem. Rev.* **106**, 3210–3235 (2006).
10. Kamerlin, S. C. L. & Warshel, A. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins: Structure, Function, and Bioinformatics* vol. 78 1339–1375 (2010).
11. Boxer, S. G., Fried, S. D., Schneider, S. H. & Wu, Y. ELECTRIC FIELDS AND ENZYME CATALYSIS. *Catalysis in Chemistry and Biology* (2018) doi:10.1142/9789813237179_0039.
12. 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.* **54**, 6622–6634 (2018).

13. Villà, J. & Warshel, A. Energetics and Dynamics of Enzymatic Reactions. *The Journal of Physical Chemistry B* vol. 105 7887–7907 (2001).
14. Menger, F. M. & Nome, F. Interaction vs Preorganization in Enzyme Catalysis. A Dispute That Calls for Resolution. *ACS Chem. Biol.* **14**, 1386–1392 (2019).
15. Kamerlin, S. C. L. & Warshel, A. The EVB as a quantitative tool for formulating simulations and analyzing biological and chemical reactions. *Faraday Discuss.* **145**, 71–106 (2010).
16. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. Computational design of a biologically active enzyme. *Science* **304**, 1967–1971 (2004).
17. Hellinga, H. W. In the wake of two retractions, a request for investigation. *Nature* vol. 454 397 (2008).
18. Jiang, L. *et al.* De Novo Computational Design of Retro-Aldol Enzymes. *Science* **319**, 1387–1391 (2008).
19. Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329**, 309–313 (2010).
20. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
21. Richter, F. *et al.* Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–16206 (2012).
22. Bjelic, S. *et al.* Computational design of enone-binding proteins with catalytic activity for the Morita-Baylis-Hillman reaction. *ACS Chem. Biol.* **8**, 749–757 (2013).
23. Leaver-Fay, A. *et al.* Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol.* **523**, 109–143 (2013).
24. Bonneau, R. *et al.* Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* **Suppl 5**, 119–126 (2001).
25. Gorder, P. F. Top7: from computer-aided design, a new protein. *Computing in Science & Engineering* vol. 6 6–9 (2004).
26. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational enzyme design. *Angew.*

Chem. Int. Ed Engl. **52**, 5700–5725 (2013).

27. Bunzel, H. A., Anderson, J. L. R. & Mulholland, A. J. Designing better enzymes: Insights from directed evolution. *Curr. Opin. Struct. Biol.* **67**, 212–218 (2021).
28. Mak, W. S. & Siegel, J. B. Computational enzyme design: transitioning from catalytic proteins to enzymes. *Curr. Opin. Struct. Biol.* **27**, 87–94 (2014).
29. Williams, G. & Hall, M. *Modern Biocatalysis: Advances Towards Synthetic Biological Systems*. (Royal Society of Chemistry, 2018).
30. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3790–3795 (2012).
31. Warshel, A. Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem.* **273**, 27035–27038 (1998).
32. Kamerlin, S. C. L., Sharma, P. K., Chu, Z. T. & Warshel, A. Ketosteroid isomerase provides further support for the idea that enzymes work by electrostatic preorganization. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 4075–4080 (2010).
33. Cui, Q. Faculty Opinions recommendation of Catalysis by dihydrofolate reductase and other enzymes arises from electrostatic preorganization, not conformational motions. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature* (2011) doi:10.3410/f.13294041.14653181.
34. Lameira, J., Bora, R. P., Chu, Z. T. & Warshel, A. Methyltransferases do not work by compression, cratic, or desolvation effects, but by electrostatic preorganization. *Proteins: Structure, Function, and Bioinformatics* vol. 83 318–330 (2015).
35. Warshel, A. & Weiss, R. M. An empirical valence bond approach for comparing reactions in solutions and in enzymes. *Journal of the American Chemical Society* vol. 102 6218–6226 (1980).
36. Vardi-Kilshtain, A., Roca, M. & Warshel, A. The empirical valence bond as an effective strategy for computer-aided enzyme design. *Biotechnol. J.* **4**, 495–500 (2009).
37. Amrein, B. A. *et al.* : Computer-Aided Directed Evolution of Enzymes. *IUCrJ* **4**, 50–64 (2017).
38. Jindal, G., Ramachandran, B., Bora, R. P. & Warshel, A. Exploring the Development of Ground-State

Destabilization and Transition-State Stabilization in Two Directed Evolution Paths of Kemp Eliminases. *ACS Catal.* **7**, 3301–3305 (2017).

39. Chowdhury, R. & Maranas, C. D. From directed evolution to computational enzyme engineering—A review. *AIChE Journal* vol. 66 (2020).
40. Khersonsky, O. *et al.* Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. *J. Mol. Biol.* **407**, 391–412 (2011).
41. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).
42. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* vol. 503 418–421 (2013).
43. Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* **12**, 944–950 (2016).
44. Petrović, D., Risso, V. A., Kamerlin, S. C. L. & Sanchez-Ruiz, J. M. Conformational dynamics and enzyme evolution. *Journal of The Royal Society Interface* vol. 15 20180330 (2018).
45. Broom, A. *et al.* Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).
46. Risso, V. A. & Sanchez-Ruiz, J. M. Resurrected Ancestral Proteins as Scaffolds for Protein Engineering. *Directed Enzyme Evolution: Advances and Applications* 229–255 (2017) doi:10.1007/978-3-319-50413-1_9.
47. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
48. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
49. Marcos, E. & Silva, D. Essentials of de novo protein design: Methods and applications. *WIREs Computational Molecular Science* vol. 8 (2018).
50. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of

macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).

51. Risso, V. A. *et al.* De novo active sites for resurrected Precambrian enzymes. *Nat. Commun.* **8**, 1–13 (2017).
52. DeLuca, S., Khar, K. & Meiler, J. Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand. *PLoS One* **10**, e0132508 (2015).
53. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. De novo enzyme design using Rosetta3. *PLoS One* **6**, e19230 (2011).
54. Tyka, M. D. *et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405**, 607–618 (2011).
55. Purg, M. & Kamerlin, S. C. L. Empirical Valence Bond Simulations of Organophosphate Hydrolysis: Theory and Practice. *Methods Enzymol.* **607**, 3–51 (2018).
56. Lee, F. S., Chu, Z. T., Bolger, M. B. & Warshel, A. Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603. *Protein Eng.* **5**, 215–228 (1992).
57. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10383–10388 (2000).
58. Loshbaugh, A. L. & Kortemme, T. Comparison of Rosetta flexible-backbone computational protein design methods on binding interactions. doi:10.1101/674291.
59. Humphris-Narayanan, E., Akiva, E., Varela, R., Conchúir, S. Ó. & Kortemme, T. Prediction of Mutational Tolerance in HIV-1 Protease and Reverse Transcriptase Using Flexible Backbone Protein Design. *PLoS Computational Biology* vol. 8 e1002639 (2012).
60. Babor, M., Mandell, D. J. & Kortemme, T. Assessment of flexible backbone protein design methods for sequence library prediction in the therapeutic antibody Herceptin-HER2 interface. *Protein Sci.* **20**, 1082–1089 (2011).
61. Smith, S. T. & Meiler, J. Assessing multiple score functions in Rosetta for drug discovery. *PLoS One* **15**, e0240450 (2020).

62. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

63. Markin, C. J. *et al.* Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* **373**, (2021).

64. Fleishman, S. J. *et al.* RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One* **6**, e20161 (2011).

65. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

66. Frish, M. J. *et al.* Gaussian 09, revision A. 02. *Gaussian Inc. , Wallingford CT* (2009).

67. Dupradeau, F.-Y. *et al.* The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Phys. Chem. Chem. Phys.* **12**, 7821–7839 (2010).

68. Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*. (Wiley, 1991).

69. Feynman, R. P. Forces in Molecules. *Phys. Rev.* **56**, 340–343 (1939).

70. Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **93**, 2395–2417 (1993).

71. Beveridge, D. L. & DiCapua, F. M. Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.* **18**, 431–492 (1989).

72. Bauer, P. *et al.* Q6: A comprehensive toolkit for empirical valence bond and related free energy calculations. *SoftwareX* vol. 7 388–395 (2018).

73. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

74. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

75. 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.* **97**, 10269–10280 (1993).

76. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple

potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

77. Warshel, A. & King, G. Polarization constraints in molecular dynamics simulation of aqueous solutions: The surface constraint all atom solvent (SCAAS) model. *Chem. Phys. Lett.* **121**, 124–129 (1985).
78. Vreven, T. *et al.* Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* **427**, 3031–3041 (2015).
79. Han, L., Yang, Q., Liu, Z., Li, Y. & Wang, R. Development of a new benchmark for assessing the scoring functions applicable to protein–protein interactions. *Future Med. Chem.* **10**, 1555–1574 (2018).
80. Yan, Y. & Huang, S.-Y. Pushing the accuracy limit of shape complementarity for protein-protein docking. *BMC Bioinformatics* **20**, 696 (2019).
81. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–53 (2016).
82. Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A. & Warren, S. D. Essential considerations for using protein–ligand structures in drug discovery. *Drug Discov. Today* **17**, 1270–1281 (2012).
83. Parks, C. D. *et al.* D3R grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Des.* **34**, 99–119 (2020).
84. Roda, S. *et al.* Computationally Driven Rational Design of Substrate Promiscuity on Serine Ester Hydrolases. *ACS Catal.* **11**, 3590–3601 (2021).
85. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
86. Nesmiyanov, P. P. Antigen Presentation and Major Histocompatibility Complex. (2020).
87. van der Merwe, P. A. & Cordoba, S.-P. Late arrival: recruiting coreceptors to the T cell receptor complex. *Immunity* vol. 34 1–3 (2011).
88. Feltkamp, M. C., Vierboom, M. P., Kast, W. M. & Melief, C. J. Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity. *Mol. Immunol.* **31**, 1391–1401

(1994).

89. Mei, S. *et al.* A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief. Bioinform.* **21**, 1119–1135 (2020).
90. Antunes, D. A., Abella, J. R., Devaurs, D., Rigo, M. M. & Kavraki, L. E. Structure-based Methods for Binding Mode and Binding Affinity Prediction for Peptide-MHC Complexes. *Curr. Top. Med. Chem.* **18**, 2239–2255 (2018).
91. Ishizuka, J. *et al.* Quantitating T cell cross-reactivity for unrelated peptide antigens. *J. Immunol.* **183**, 4337–4345 (2009).
92. Sirin, S., Pearlman, D. A. & Sherman, W. Physics-based enzyme design: Predicting binding affinity and catalytic activity. *Proteins: Structure, Function, and Bioinformatics* vol. 82 3397–3409 (2014).
93. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
94. Armstrong, K. M., Piepenbrink, K. H. & Baker, B. M. Conformational changes and flexibility in T-cell receptor recognition of peptide–MHC complexes. *Biochem. J* **415**, 183–196 (2008).
95. Tynan, F. E. *et al.* A T cell receptor flattens a bulged antigenic peptide presented by a major histocompatibility complex class I molecule. *Nature Immunology* vol. 8 268–276 (2007).
96. Lin, C.-P. *et al.* Deriving protein dynamical properties from weighted protein contact number. *Proteins* **72**, 929–935 (2008).
97. Yeh, S.-W. *et al.* Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *Biomed Res. Int.* **2014**, 572409 (2014).
98. Shih, C.-H., Chang, C.-M., Lin, Y.-S., Lo, W.-C. & Hwang, J.-K. Evolutionary information hidden in a single protein structure. *Proteins* **80**, 1647–1657 (2012).
99. Chang, C.-M., Huang, Y.-W., Shih, C.-H. & Hwang, J.-K. On the relationship between the sequence conservation and the packing density profiles of the protein complexes. *Proteins* **81**, 1192–1199 (2013).
100. Huang, S.-W. *et al.* On the relationship between catalytic residues and their protein contact number.

Curr. Protein Pept. Sci. **12**, 574–579 (2011).

101. Nosrati, G. R. & Houk, K. N. Using catalytic atom maps to predict the catalytic functions present in enzyme active sites. *Biochemistry* **51**, 7321–7329 (2012).
102. Waller, I. Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen. *Zeitschrift für Physik* **17**, 398–408 (1923).
103. Debye, P. Interferenz von röntgenstrahlen und wärmebewegung. *Ann. Phys.* **348**, 49–92 (1913).
104. Stone, B. Evaluating experimental and theoretical measures of protein conformational dynamics. (Aston University, 2016).
105. Mulero, M. C. *et al.* Chromatin-bound IκBα regulates a subset of polycomb target genes in differentiation and cancer. *Cancer Cell* **24**, 151–166 (2013).
106. Jacobs, M. D. & Harrison, S. C. Structure of an IkappaBalpha/NF-kappaB complex. *Cell* **95**, 749–758 (1998).
107. Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–50 (2016).
108. Lovell, S. C. & Robertson, D. L. An Integrated View of Molecular Coevolution in Protein–Protein Interactions. *Mol. Biol. Evol.* **27**, 2567–2575 (2010).
109. Kobe, B. *et al.* Crystallography and protein–protein interactions: biological interfaces and crystal contacts. *Biochem. Soc. Trans.* **36**, 1438–1441 (2008).
110. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781–1791 (2004).
111. Sawle, L. & Ghosh, K. Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma. *J. Chem. Theory Comput.* **12**, 861–869 (2016).
112. Rovigatti, L., Sulc, P., Reguly, I. Z. & Romano, F. A comparison between parallelization approaches in molecular dynamics simulations on GPUs. *J. Comput. Chem.* **36**, 1–8 (2015).
113. Yang, Y. I., Shao, Q., Zhang, J. & Yang, L. Enhanced sampling in molecular dynamics. *The Journal of chemical* (2019).

114. Blaszczyk, M. *et al.* Protein Structure Prediction Using Coarse-Grained Models. in *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes: From Bioinformatics to Molecular Quantum Mechanics* (ed. Liwo, A.) 27–59 (Springer International Publishing, 2019).

115. Nymeyer, H., García, A. E. & Onuchic, J. N. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5921–5928 (1998).

116. Taketomi, H., Ueda, Y. & Gō, N. STUDIES ON PROTEIN FOLDING, UNFOLDING AND FLUCTUATIONS BY COMPUTER SIMULATION: I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* **7**, 445–459 (1975).

117. Noel, J. K. & Onuchic, J. N. The Many Faces of Structure-Based Potentials: From Protein Folding Landscapes to Structural Characterization of Complex Biomolecules. in *Computational Modeling of Biological Systems: From Molecules to Pathways* (ed. Dokholyan, N. V.) 31–54 (Springer US, 2012).

118. Noel, J. K. *et al.* SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput. Biol.* **12**, e1004794 (2016).

119. Whitford, P. C. *et al.* An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins* **75**, 430–441 (2009).

120. Clementi, C., Nymeyer, H. & Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and 'on-route' intermediates for protein folding? An investigation for small globular proteins. *arXiv [cond-mat.stat-mech]* (2000).

121. Floor, M. *et al.* SBMOpenMM: A Builder of Structure-Based Models for OpenMM. *J. Chem. Inf. Model.* **61**, 3166–3171 (2021).

122. Noel, J. K., Whitford, P. C., Sanbonmatsu, K. Y. & Onuchic, J. $1/2$ N. SMOG@ ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res.* **38**, W657–W661 (2010).

123. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).

124. Carugo, O. How large B-factors can be in protein crystal structures. *BMC Bioinformatics* **19**, 61

(2018).

125. 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).
126. Wu, H. & Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *Journal of Nonlinear Science* vol. 30 23–66 (2020).
127. Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **140**, 2386–2396 (2018).
128. Röblitz, S. & Weber, M. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Adv. Data Anal. Classif.* **7**, 147–179 (2013).
129. Metzner, P., Schütte, C. & Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **7**, 1192–1219 (2009).
130. Lane, T. J., Shukla, D., Beauchamp, K. A. & Pande, V. S. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **23**, 58–65 (2013).
131. Hruska, E., Abella, J. R., Nüske, F., Kavraki, L. E. & Clementi, C. Quantitative comparison of adaptive sampling methods for protein dynamics. *J. Chem. Phys.* **149**, 244119 (2018).
132. Sun, L., Noel, J. K., Sulkowska, J. I., Levine, H. & Onuchic, J. N. Connecting thermal and mechanical protein (un)folding landscapes. *Biophys. J.* **107**, 2950–2961 (2014).
133. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
134. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 015102 (2013).
135. Trendelkamp-Schroer, B., Wu, H., Paul, F. & Noé, F. Estimation and uncertainty of reversible Markov models. *J. Chem. Phys.* **143**, 174101 (2015).
136. Senn, H. M. & Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed Engl.* **48**, 1198–1229 (2009).
137. Himo, F. Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *J. Am. Chem. Soc.*

139, 6780–6786 (2017).

138. Leopoldini, M. *et al.* The role of quantum chemistry in the elucidation of the elementary mechanisms of catalytic processes: from atoms, to surfaces, to enzymes. *Theoretical Chemistry Accounts* vol. 117 765–779 (2007).
139. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).
140. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* vol. 25 1422–1423 (2009).
141. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* vol. 109 1528–1532 (2015).
142. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
143. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* vol. 30 772–780 (2013).
144. Camacho, C. *et al.* BLAST : architecture and applications. *BMC Bioinformatics* vol. 10 (2009).
145. Consortium, T. U. & The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research* vol. 43 D204–D212 (2015).
146. Noel, J. K., Whitford, P. C. & Onuchic, J. N. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. *J. Phys. Chem. B* **116**, 8692–8702 (2012)

Appendices

Appendix 1 - Methodologies

Computational methods for studying enzymatic reactivity

Quantum-mechanics based methods

There have been many attempts to relate protein structure with their catalytic activities using computer simulations. Given the electronic nature of any chemical change, it is reasonable to address this question using QC methods. Due to the cost of representing the electronic structure of large molecular systems, they are usually limited to treat only part of the system at a high *ab initio* theoretical level, while the remaining system's region is usually represented with a simpler Molecular Mechanics (MM) force field, as in the case of QM/MM,[136] or replaced with an implicit solvent, as in the case of the cluster approach[137] or more traditional quantum mechanical (QM) studies.[138] The advantages of these methods are that many proposals for the reaction mechanism can be studied without much prior experimental knowledge of the reaction. This is essential to understand the energetics of the electronic rearrangements, and also, to discriminate among several possible mechanistic proposals. Regarding enzymatic design, it is usually the only methodology capable of giving correct geometries and charge distributions to create TS models of the elementary steps of a target reaction and to give approximated reaction energies that can be used to calibrate other simulation methods.

The Empirical Valence Bond model

A very helpful tool for studying enzymatic reactions is the EVB method. Developed by A. Warshel in the '80s,[35] the method offered an excellent trade-off between computational efficiency and physical accuracy. Its principal advantage lies in that physical interactions, including chemical bonds being broken/formed, are represented by simplified energy functions that capture the energetics of the electronic rearrangements at the cost of standard MD force fields.

Using the formalism of a semi-empirical QC method, EVB physical representation of the reacting systems has its roots in Valence Bond theory, in which several adiabatic ground state potentials

are mixed to give rise to the diabatic description of the chemical reaction. Although the EVB method uses standard and generalizable force fields for representing and simulating chemical reactions, it needs parametric calibration to adjust the contribution of each adiabatic Hamiltonian to the diabatic energy surface. The data employed for this parameterisation are the thermodynamic and kinetic constants, specific for the reaction under study, which can come from wet-lab experimental measurements or *ab initio* QC calculations of the relevant chemical steps.

Given its more computationally tractable description, the EVB method has been employed to simulate the free-energy surface of enzymatic reactions, which is still deemed too costly to be carried out with *ab initio* methods such as QM or QM/MM approaches. This gives a more realistic picture of active-site dynamics and allows to interpret and contrast the physical interactions and reorganisational changes produced along the reaction coordinates of different solvents.

When studying the physical origins of enzymatic catalysis with EVB, it is necessary to run simulations in the uncatalysed (i.e., water solution) and the catalysed environment (i.e., the solvated enzymatic system). One of these simulations is calibrated to match available thermodynamic and kinetic data, for afterwards, using the same calibrated parameters, predict the corresponding thermodynamic constants in the enzymatic simulation. If the method is successful in predicting the kinetic change (in activation free energies) between these environments, the simulation is considered valid and can be queried to understand the physical origin of the simulated catalytic effects. Likewise, the reference simulation can be a low-activity enzyme, then, the adjusted parameters can be used to predict the catalytic trend among improved variants to understand the origins of the improved catalytic activity. This idea is straightforwardly extended to screen sets of computational enzymatic designed variants.

Computational methods for enzymatic design

Among several strategies to suggest amino acid changes for computational enzymatic design, they are divided into two strategies: *de novo* enzyme design and enzymatic optimisation or redesign.

The inside-out approach falls in the category of *de novo* enzyme design. It starts by building a model for a hypothetical active site that putatively catalyses a target reaction. This description, referred to as theozyme, is usually achieved at the quantum chemistry level by optimising the positions of disembodied side chains surrounding a model of the reacting molecule's TS. With these geometrical descriptions, the method then searches compatible pockets in a set of predefined protein scaffolds that are geometrically able to position the active-site proposal. Matches thus found are optimised searching for suitable sequences that stabilise the core catalytic proposal. Models are then selected by a combination of metrics that try to assess the stability of the protein and its active site complementary with the TS model. A batch of the most promising models is chosen for the assessment of their desired catalytic activity.

Methods for catalytic redesign can vary significantly. They start from an already active enzyme with low activity and, from there, they search the protein sequence landscape seeking catalytically improving mutations. The suggested changes by these methods depend mainly on the catalytic hypothesis on how to improve the reaction rate and they are usually guided by a score function able to discriminate the effect that mutations could have on the protein stability. Some approaches first assess the catalytic effect of individual residues by running QC, EVB or MD simulations. Depending on these findings, they propose new changes that can be tested experimentally.

The weighted Contact Number Metric to study the relationship of protein structure, dynamics, and evolution

The Weighted Contact Number (WCN) is a metric that quantifies the "crowdedness" that a specific atom or residue has in a particular protein molecule. Its mathematical form is based on a continuous quantification of the contact map, in which the contact specific contribution is the squared inverse of the contact distance.

When the WCN contributions of all the atoms are added up, the obtained profile has excellent agreement with the system's experimental dynamical profile (e.g., the B-factor profile).[96] It has also been shown that the WCN agrees with the sequence conservation (SC) profile of protein families.[97] These agreements entail exciting hypotheses on the relationship between protein structure, dynamics, and evolution which we can now approach quantitatively with the WCN.

Although the exact mathematical relationship between these profiles is unknown, a possible interpretation of its origin can be hypothesised.

On the dynamic side, the WCN is based entirely on the native structure of the protein and defines the magnitudes of the system's dynamics around the native free-energy basin. Thus, atoms or residues in more crowded regions are expected to deviate less from their native positions, while others can move more freely because they are in less crowded environments.

On the evolutionary side, when a set of evolutionarily related proteins (i.e., a protein family) explores the protein sequence/fitness landscape in search of better adaptability, or even as a consequence of neutral drift, there is a higher restriction to swap residues in crowded regions than residues located in less restricted ones. The logical interpretation is that well-packed regions contain a higher number of optimal interactions that are better fulfilled by specific residue types and that, by changing them, the probability of satisfying those interactions drops in proportion to the number of contacts surrounding its immediate environment.

Appendix 2 - Code

SBMOpenMM Python Simulation Code For All-atom SBM simulation

```
#Import SBMOpenMM library
import sbmOpenMM

# Import OpenMM library as in for OpenMM 7.6 release
from openmm.app import *
from openmm import *
from openmm.unit import *

# Create a system instance using the all-atom model in the models class
sbmAA = sbmOpenMM.models.getAllAtomModel(pdb_file, contact_file)

# Define simulation object
integrator = LangevinIntegrator(temperature*kelvin, 1.0/picosecond, 0.002*picoseconds)
simulation = Simulation(sbmAA.topology, sbmAA.system, integrator)
simulation.context.setPositions(sbmAA.positions)
simulation.reporters.append(DCDReporter('AAModel_traj.dcd', 10000))

# Define a SBM reporter to write the SBM energies into a file
sbmReporter = sbmOpenMM.reporter.sbmReporter('AAModel_energy.data', 100, step=True,
potentialEnergy=True, temperature=True, sbmObject=sbmAA)

# Add the reporter to the simulation object
simulation.reporters.append(sbmReporter)

# Run the simulation for N steps.
simulation.step(n_steps)
```

The PyCBBL library

Most code developed for setting up and analysing the calculations in this thesis can be found in the exclusively-developed library for the Computational Biochemistry and Biophysics Lab (CBBL),

directed by Professor Jordi Villà Freixa. The PyCBBL library is hosted in the GitHub platform and can be publicly accessed through the following link:

https://github.com/CompBiochBiophLab/pycbbl

The PyCBBL library contains heavy dependencies on programs like PyRosetta (https://www.pyrosetta.org/),[139] BioPython (https://biopython.org/),[140] MDTraj (https://www.mdtraj.org),[141] Pymol (https://pymol.org), sciPy (https://scipy.org), pandas (https://pandas.pydata.org), scrapy (https://scrapy.org), among other bioinformatic libraries, since it encompasses the many aspects of this work. The PyCBBL library contains specific modules addressing different methods and functions. We give a brief description of them here:

- **alignment** - Contains methods for working with multiple sequence alignments, based on different packages, such as BLAST,[142–144] MAFFT,[142,143] and CD-HIT.[142]
- **calculations** - Contain the methods used for deploying parallel calculations either locally, or on specific HPC clusters.
- **clustering** - Contains methods for hierarchical clustering analysis based on the hierarchical SciPy module.
- **databases** - different methods for web scraping the PDB and the UniProt[145] webpages based on scrapy.
- **MD** - Different methods based on MDTraj and PyMol for structural alignment, clustering, and interface analysis of MD trajectories.
- **PDB** - Adapted methods to work with PDB structures to analyse multiple PDB structures. It depends mainly on the BioPython[65,140,141,145] PDB module and Pymol.
- **protein_contacts** - Contain methods for generating contact and topology information for SBM. Uses the shadow map algorithm[146] implemented in the smog2[118] program for generating contact information.
- **rosetta** - Contains the functions used to calculate Boltzman averaged scores and methods to convert Rosetta[50] silent files into MD trajectory files for faster analysis using the MD module.
- **WCN** - A class for calculating the different WCN metrics implemented in this study.