Modelling Asparaginases for Leukemia Treatments: Multiscale Simulations as a Guide for Enzyme Design

Milorad Anđelković

Departamento de Química Física
Universitat de València

Programa de Doctorado en Química Teórica y Modelización Computacional

Dirigida por los doctores
Ignacio Nilo Tuñón García de Vicuña
José Javier Ruiz Pernía

March, 2025

Ignacio Nilo Tuñón García de Vicuña, Profesor Catedrático de Química Física del Departamento de Química Física de la Universitat de València y José Javier Ruiz Pernía, Profesor Titular de Química Física del Departamento de Química Física de la Universitat de València,

CERTIFICAN:

Que el trabajo con el título: "Modelling Asparaginases for Leukemia Treatments: Multiscale Simulations as a Guide for Enzyme Design" ha sido realizado por Don Milorad Andjelkovic bajo nuestra dirección, para optar al grado Doctor en Química.

Así, autorizamos la presentación de este trabajo a efectos de seguir los trámites correspondientes de la Universitat de València.

València, 11 de marzo de 2025

DIRECTORES

Fdo.: Ignacio Nilo Tuñón García de Vicuña Fdo.: José Javier Ruiz Pernía

Contents

List of Abbreviations

AEI	Acyl-enzyme Intermediate
AF	AlphaFold
ALL	Acute Lymphoblastic Leukemia
AM1	Austin Model 1
AMBER	Assisted Model Building with Energy Refinement
AML	Acute Myeloid Leukemia
ASM	Adaptive String Method
ASNase	Asparaginase
ASNS	Asparagine Synthetase
ATP	Adenosine Triphosphate
BA	Beta-amylase
BLAST	Basic Local Alignment Search Tool
CD	Circular Dichroism
CG	Conjugate Gradient
CLL	Chronic Lymphoblastic Leukemia
CML	Chronic Myeloid Leukemia
CV	Collective Variable
DDT	Dichlorodiphenyltrichloroethane
DFT	Density Functional Theory
DFTB3	Density-Functional Tight-Binding
DL	Deep Learning
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
DOF	Degrees of Freedom
EC	Enzyme Commission number
EH	Epoxide Hydrolase
EM	Electron Microscopy
FDA	Food and Drug Administration
FEP	Free Energy Perturbation
FES	Free Energy Surface
FF	Force Field
GB	Generalized Born

GBSA	Generalized Born Surface Area Solvation
GG	Golden Gate
HF	Hartree-Fock
HRE	Hamiltonian Replica Rxchange
IDT	Integrated DNA Technologies
IPD	Institute for Protein Design
IPTG	Isopropyl-ß-D-Thiogalactopyranoside
k_{cat}	Catalytic Constant
K_M	Michaelis-Menten constant
MD	Molecular Dynamics
MFEP	Minimum Free Energy Path
ML	Machine Learning
MM	Molecular Mechanics
MP	Mass Photometry
MPNN	Message Passing Neural Networks
MRE	Molar Residue Ellipticity
MSA	Multiple Sequence Alignment
MSG	Monosodium Glutamate
NMR	Nuclear Magnetic Resonance
NTA	Nickel-Charged Nitrilotriacetic Acid
OD	Optical Density
PAGE	Polyacrylamide Gel Electrophoresis
PB	Poisson-Boltzmann
PBSA	Possion-Boltzman Surface Area Solvation
PDB	Protein Data Bank
PEG	Polyethylene Glycol
PES	Potential Energy Surface
pLDDT	Per-residue Local Distance Difference Test
PMF	Potential of Mean Force
PMSF	Phenylmethylsulfonyl Fluoride
QM	Quantum Mechanics
RC	Reaction Coordinate
RE	Replica Exchange
RESP	Restrained Electrostatic Potential
RF	RosettaFold
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation

SASA	Solvent Accessible Surface Area
SCF	Self-consistent field
SD	Steepest Descent
SEC	Size Exclusion Chromatography
sEH	Soluble Epoxide Hydrolase
SNAC	Sequence-Specific Nickel-Assisted Cleavage
SOC	Super Optimal Broth
TCA	Trichloroacetic acid
TEV	Tobacco Etch virus
TI	Thermodynamic Integration
TS	Transition State
US	Umbrella Sampling
UV	Ultraviolet
VdW	Van der Waals
WHAM	Weighted Histogram Analysis Method

Chapter 1: Introduction

1.1 Acute Lymphoblastic Leukemia

1.1.1 Leukemia

Leukemia is a hematological malignancy affecting blood cells and bone marrow, characterized by uncontrolled proliferation and differentiation of white blood cells. This leads to a gradual displacement of normal, healthy blood cells, resulting in the body becoming overwhelmed by dysfunctional cells.[1, 2] In 2018, it was ranked as the fifteenth most diagnosed cancer, with an estimate over three hundred thousand deaths in the United States due to this disease (American Cancer Society). Some of the risk factors for leukemia include radiation exposure, hereditary syndromes, smoking, age, and some unknown factors.[3]

Based on the affected cells, leukemia is primarily categorized into four main types: Acute Myeloid Leukemia (AML), Chronic Myeloid Leukemia (CML), Acute Lymphoblastic Leukemia (ALL) and Chronic Lymphoblastic Leukemia (CLL). Namely, primary blood stem cells differentiate into either myeloid or lymphoid stem cells, which further develop into various specialized cell types (Figure **1.1**). Lymphoid stem cells give rise to lymphoblasts, which subsequently differentiate into natural killer (NK) cells, T lymphocytes or B lymphocytes. Myeloid stem cells ultimately develop into red blood cells, platelets (thrombocytes) or myeloblasts. Subsequently, myeloblasts differentiate into granulocytes (neutrophils, eosinophils or basophils). Additionally, B and T lymphocytes, NK cells, and granulocytes constitute white blood cells, which are fundamental to the immune system.

In myeloid leukemia (ML), myeloblasts divide uncontrollably, leading to the formation of abnormal granulocytic white blood cells that are incapable of performing their normal functions. This type of leukemia is also referred to as granulocytic or non-lymphocytic leukemia.[4] Acute myeloid leukemia progresses rapidly, whereas chronic myeloid leukemia develops more slowly.

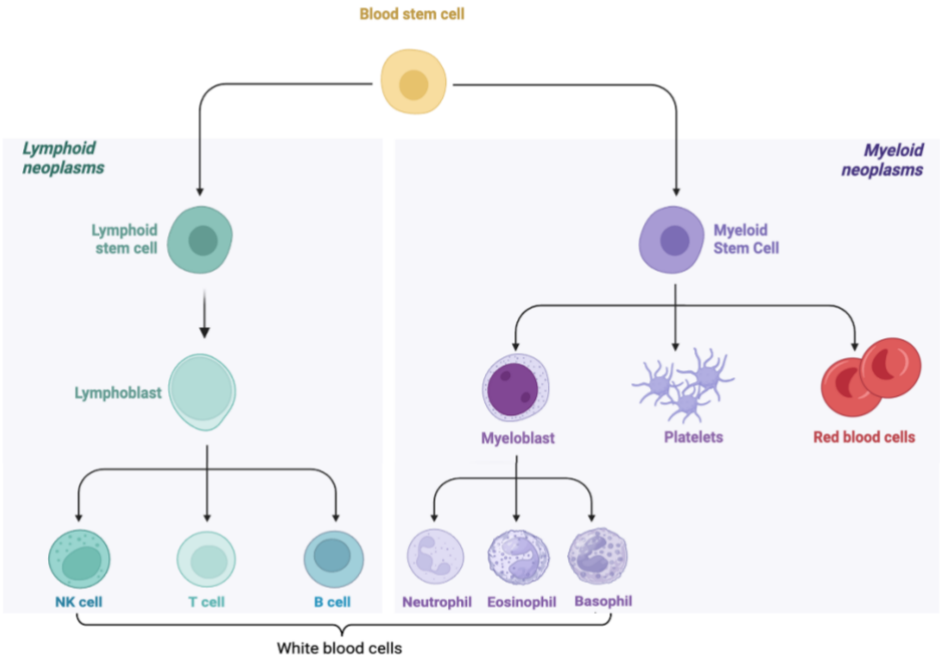


Figure 1.1. Differentiation of the blood stem cells.

Similarly, lymphoblastic leukemia (LL) is characterized by uncontrollable proliferation of the white blood cells in the bone marrow. Depending on which subtype of white blood cells are affected (B, T or NK), LL can further be classified, the most common subtypes being B and T lymphoblastic leukemia. Again, ALL develops very rapidly, thereby causing immediate and severe symptoms, while CLL develops slower. Even though it can occur in adults, ALL primarily affects children between the ages of 1 and 7 years.[5] The incidences have also been noted to be higher in children of American Indian, Native Alaskan, and Hispanic descents as well as almost twice as much higher in White than Black kids.[6]

The ALL-treatment goal is to destroy malignant cells in the bone marrow and the blood. This is typically achieved with the combination of radiation [7], chemotherapy [8] and stem cells transplantation. [9] Chemotherapy treatment for the ALL usually includes a mixture of FDA-approved drugs, which are often

administered to patients as a cocktail. Some common drugs are Mercaptopurine [10], Vincristine [11], Dexamethasone [12], Methotrexate [13] and Cyclophosphamide [14] in combination with the enzyme L-Asparaginase.

1.1.2 L-Asparaginase

L-Asparaginases (L-ASNases or simply ASNases, EC 3.5.1.1) are a class of hydrolase enzymes that catalyze the transformation of asparagine (Asn) to aspartate (Asp) and ammonia (Figure 1.2).

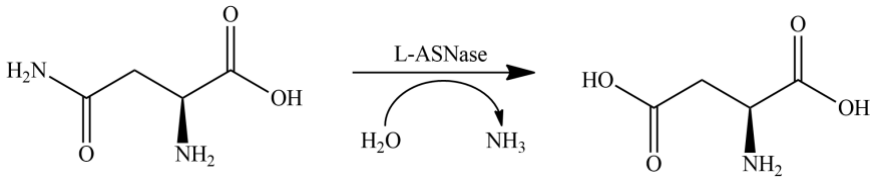


Figure 1.2. Schematic representation of L-asparaginase reaction.

These enzymes are a critical component in the treatment of ALL and non-Hodgkin lymphoma since the 1960s.[15] The connection between L-asparaginases and its anti-leukemic effects was first described in the 1950s when Kidd demonstrated that guinea pig serum could induce regression of transplanted lymphomas in mice.[16] Then, in the early 1960s, Broome identified ASNases as the key component behind the therapeutic activity of the serum. [17] A few years later, Yellin and Wriston successfully isolated a L-ASNase from guinea pig serum and provided direct evidence of its efficacy against leukemia.[18] This breakthrough led to its clinical application the very same year, marking a significant milestone in ALL therapy. The high importance of ASNases in clinical practice is also evidenced by their inclusion in the World Health Organization (WHO) List of Essential Medicines.

The therapeutic effect of ASNases in treating ALL comes from the inability of ALL blast cells to synthesize asparagine independently. Namely, blast cells exhibit little to no detectable Asparagine Synthetase (ASNS) enzyme. [19] ASNS catalyzes the synthesis of asparagine (Asn) from aspartate through an ATP-dependent reaction. As a result, the survival of malignant cells relies entirely on

the exogenous supply of L-asparagine from the patient's serum. Therefore, intramuscular or intravenous administration of an ASNase enzyme depletes the circulating Asn in the blood, depriving blast cells of this essential nutrient. The inhibition of protein synthesis due to the lack of Asn ultimately triggers the apoptosis of the blast cells. [20] Importantly, normal blood cells are unaffected by this treatment since they can synthesize L-asparagine. This has provided a strong foundation for the development of what is known as Amino Acid Depletion Cancer Therapy.[21]

1.2 Asparaginases: Classification and Kinetic Properties

1.2.1 Nomenclature issues: Class and type

Based on the source of the first enzyme discovered, ASNases were initially classified into three canonical classes: bacterial, plant and Rhizobium etli-class (see Figure 1.3).[22] However, this classification soon became misleading since various organisms produced various types ASNases that are distributed across various classes.[23] For example, even though some *E. coli* ASNases are really classified as a bacterial type, other *E. coli* ASNases are rather plant-type. For this reason, nowadays, these are renamed as class 1, 2 and 3, with references to the canonical classes above.[24] In addition, ASNases are also divided in types 1, 2, etc. based on their structural similarities.[23] In this doctoral thesis, this classification proposal, given by Silva et al [24], will be followed (Figure 1.3). This classification was originally established for *E. coli* enzymes and later extended to other proteins with a similar architecture to those of *E. coli*. For example, the enzyme with the highest structural similarity with *E. coli* ASNase type 1 is therefore classified as type 1 enzyme.

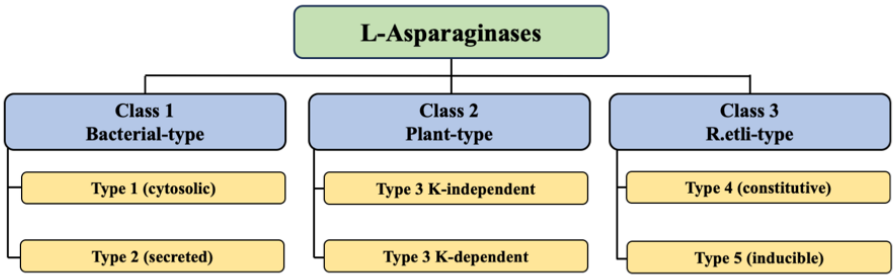


Figure 1.3. Structure-based classification of L-asparaginases. Adapted from [23].

There are different ways ASNases are being denoted in the literature. In this doctoral thesis, we will most of the times be referring to the ASNase types and not classes. Therefore, we will use the following rule: the name of an *x*ASNase is constructed in such a way that *x* (first letters) correspond to the source organism (e.g. *Escherichia Coli* – Ec, *human* – h, *Guinea Pig* – gp, etc.) and after the name,

a number (1, 2 etc.) is added to represent a belonging type of enzyme. For example, EcASNase3 stands for the *E. coli* asparaginase of type 3.

1.2.2 Class 1 L-ASNases

Class 1 enzymes (previously called bacterial ASNases) contain two ASNase types. Namely, *E. coli* produces these two types of ASNases encoded by the ansA and ansB genes. These enzymes are referred to as EcASNase1 (type 1, a cytosolic enzyme expressed in the cytoplasm) and EcASNase2 (type 2, periplasmic enzyme expressed under anaerobic conditions).[25] The structures of these two enzymes are given in Figure 1.4a and Figure 1.4b. Both EcASNase1 and EcASNase2 are crystalized as a homotetramers formed by two intimate dimers (see Figure 1.4).[26] However, up to this date there are no structural explanations to support the hypothesis that protein remains in a tetrameric configuration in solution neither whether it is active as a tetramer or a dimer.[23] EcASNase1 has a Michaelis constant (K_M) in the millimolar range that corresponds to a low affinity for the substrate.[26] Interestingly, EcASNase2 shares around 60% homology with EcASNase1, but it shows greater affinity towards asparagine (K_M in the micromolar range).[26]

Each of the monomers of these two types consist of a N-terminal and a C-terminal domain, connected by a 20-residue linker (Figure **1.5**), [27] with the active site being fully located in the N-terminal domain.

In comparison to EcASNase2, both the guinea pig and human homologs of the same type (gpASNase1 and hASNase1) differ in the lengths of the linker.[28] Also, these two mammal enzymes possess additional 200-residue C-terminal domains, which are likely to contain several ankyrin repeats.[28] These additional parts have no well-defined electronic density in the X-ray spectra and are not resolved.[28] Later studies showed that this additional part neither affect catalytic activity nor binding.[29] Authors speculated that this part is probably involved in the protein-protein interactions or other cellular processes, however its role remains unknown.[29]

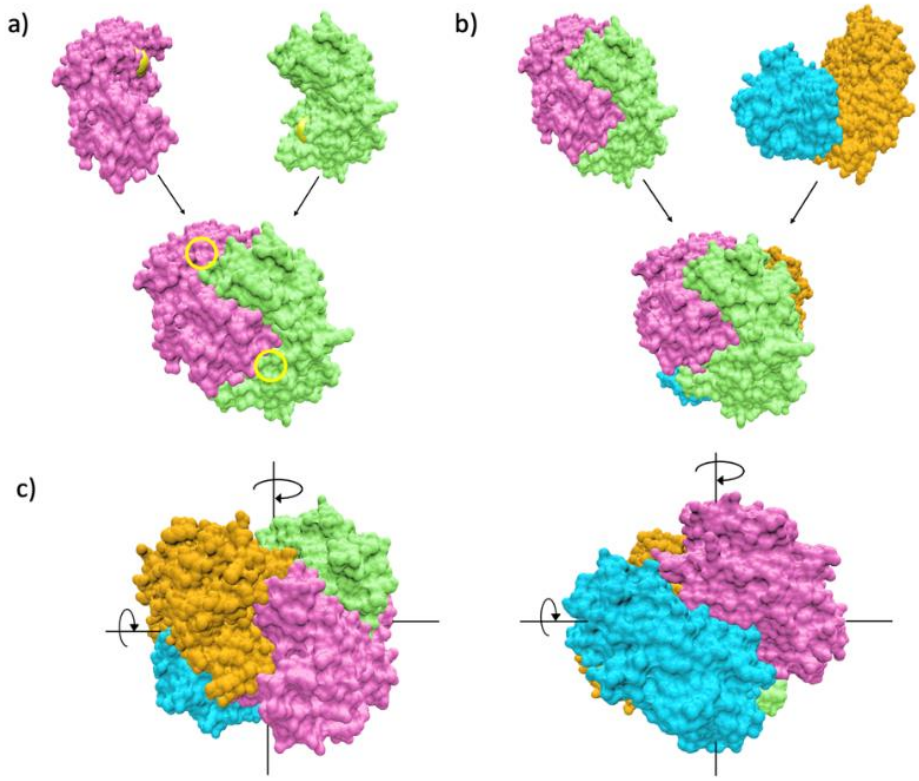


Figure 1.4. Crystal structures of EcASNase2 enzyme (PDB code: 3ECA [30]). a) Schematic representation of the intimate dimer formation. Yellow parts denote active sites; b) Formation of the homotetrameric structure out of the two intimate dimers; c) 3D view of the tetrameric structure of the EcASNase2.

An additional structural element in class 1 ASNases worth mentioning is a flexible loop (green color Figure **1.5**) in the proximity of the active site. In the product-bound structure, this loop is observed to act as a lid sterically restraining the substrate into the active site.[31] Additionally, there are also some proposals where some residues of this flexible loop also participate in the reaction mechanism.[32] (see section Previous Theoretical Work on L-asparaginases). Interestingly, in the case of type 2 ASNases, this loop is located within the same monomer (see Figure **1.5**). However, in type 1 ASNases, this loop originates from the adjacent monomer that forms the intimate dimer (see Figure 1.6).[28]

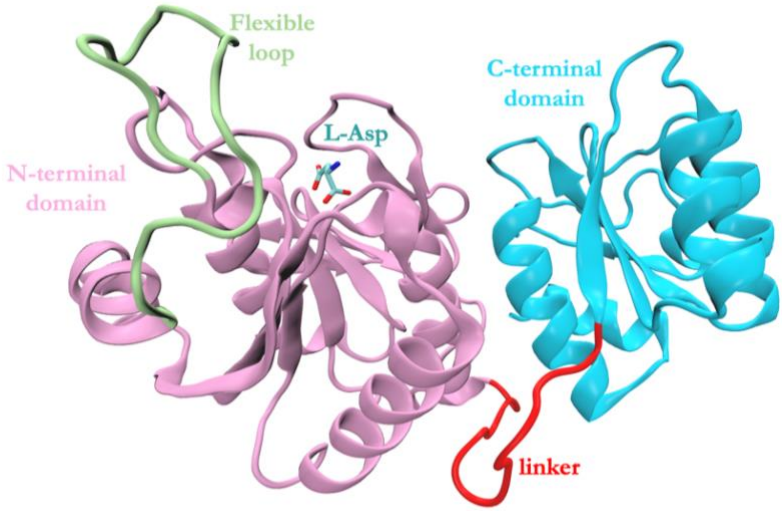


Figure 1.5. Structure of the monomer EcASNase2 (PDB code: 3ECA [30]). N-terminal is given in pink and C-terminal domain is colored in blue color. The substrate is given in the stick model in the green color. Linker connecting two domains is given in red color. A flexible loop is given in light green color.

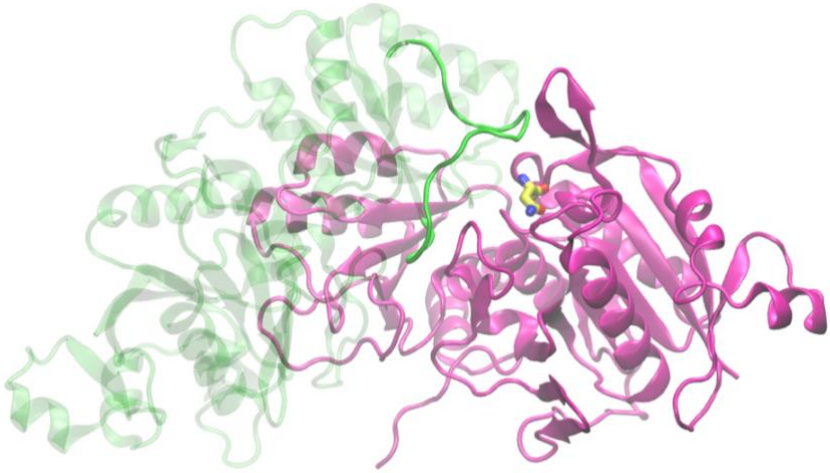


Figure 1.6. Structure of the intimate dimer gpASNase1 (PDB code: 5DNC [33]) with the two protomers given in green and pink color. The substrate in the active site of the pink protomer is given in the stick model and yellow color. A flexible loop of the green protomer is giving in opaque style while the rest of the green protomer is transparent.

There are some differences within this loop region between hASNase1 and gpASNase1. In hASNase1, this loop contains two additional residues in comparison to gpASNase1.[29] Lavie and coworkers also speculated that this can be linked to the differences in the dynamics of the loop between hASNase1 and gpASNase1, ultimately provoking the differences observed for the affinity towards asparagine between hASNase1 (K_M =2960 μM) and gpASNase1 (K_M =57.7 μM) (see Table 1.1).[29] There is an additional difference between gpASNase1 and hASNase1 that could be responsible for the differences found in the binding efficiencies: while gpASNase1 follows Michaelis Menten kinetics, hASNase1 is allosteric and follows Hill kinetic equation instead (Table 1.1).[29]

Table 1.1. Kinetic data of some wild-type class 1 ASNases[1].

Enzyme	K_M (μM)	k_{cat} (s^{-1})	k_{cat}/K_M (s^{-1} μM^{-1})	Hill
EcASNase1 [34]	400	7.4	0.028	3.5
EcASNase2 [28]	14.9	48.9	4.4	*NA*
hASNase1 [28]	2960	14.4	0.005	2.1
gpASNase1 [28]	57.7	38.6	0.8	*NA*
ErASNase2 [28]	47.5	207.5	4.4	*NA*

In general, some type 1 ASNases are allosteric (e.g. *E. coli type 1*, human type 1, *S. cerevisiae*), while other class 1 ASNases show no detectable allosteric regulation (Guinea pig type 1, *E. coli* type 2, *P. horikoshii, P. furiosus*). Interestingly, the allosteric regulator is the same substrate molecule binding to the allosteric site close to the active site (see Figure 1.7a). Signal transmission from the allosteric site to the active site likely involves subtle structural rearrangements at the dimer interface, accompanied by the relocation of the residues in the active site, placing them in the correct conformation.[35] However, the underlying reason why certain class 1 ASNases do exhibit allosteric behavior and others do not, despite retaining the allosteric site, remains elusive and poorly understood.

[1] The values in the Table 1.1 correspond to the kinetic parameters determined at the pH=8 and 37°C. The NA stand for the "Not Applicable", as these enzymes behaved according to the classical Michaelis Menten mechanism. In cases of allosteric enzymes, Hill equation was used, and K_M is actually $[S]_0$.

In addition, this allosteric regulation seems to be connected with the large- scale conformational changes taking place upon substrate binding (Figure 1.7b).[28] A "breathing-like" motion involves the monomers coming closer together, ultimately resulting in a more compact holo structure of the homotetramer. Lavie and co-workers also speculated that this conformational changes could be responsible for the binding efficiency differences within the class 1 ASNases,[28] especially considering that both allosteric EcASNase1 and hASNase1 suffer from worse binding efficiency than non-allosteric EcASNase2 and gpASNase1 (see Table 1.1). To date, hASNase1 has not been crystallized, and as a result, these large-scale conformational changes remain speculative.

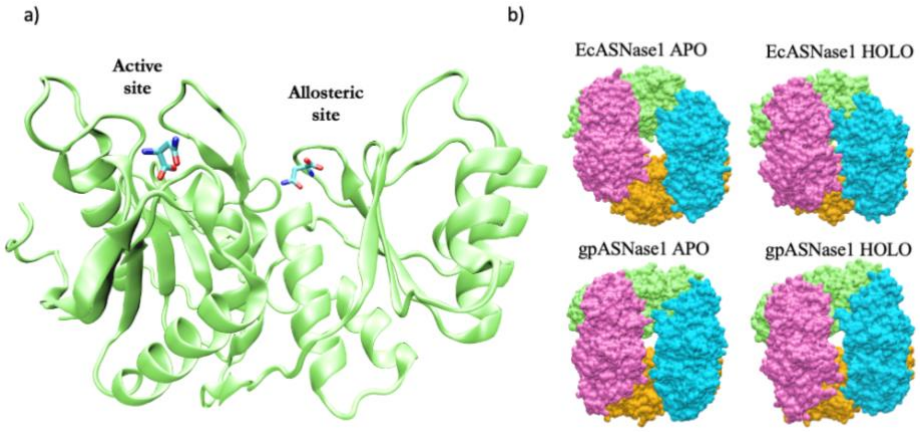


Figure 1.7. a) Active site and allosteric site within the monomer of the EcASNase1 (PDB code: 2P2N [35]); b) On the top: Conformational changes accompanying the allosteric regulation of EcASNase1. In the apo form, EcASNase1 forms an open "donut" shape (PDB code: 2P2D [35]) while holo form seems more closed (PDB code: 2P2N [35]). On the bottom: No similar conformational changes are visible when gpASNase1 switches from apo (PDB code: 4R8K [28]) to the holo form (PDB code: 4R8L [28]).

1.2.3 Class 2 L-ASNases

Class 2 enzymes, formerly known as plant type ASNases, has a dual isoaspartyl aminopeptidase/l-asparaginase activity (EC 3.5.1.1/3.4.19.5). Unfortunately they present quite low affinity towards asparagine (K_M in the millimolar range) (see Table 1.2).[23] Overall, there are fewer PDB structures available for this class

compared to class 1 ASNases and less structural diversity (RMSD between all the PDB structures within 0.6–0.8 Å).[23]

Table 1.2. Kinetic data of some wild-type class 2 ASNases[2]

Enzyme	K_M (mM)	k_{cat} (s^{-1})	k_{cat}/K_M (s^{-1} mM^{-1})	Hill
gpASNase3 [36]	2.24	3.95	1.76	*NA*
hASNase3	2.09	3.19	1.52	*NA*
EcASNase3	3.90	0.28	0.072	*NA*

Some significant representatives of this class are human (hASNAse3)[37], guinea pig (gpASNase3)[36] and *E. coli* asparaginase (EcASNase3)[39]. In contrast to the class 1 ASNases, they all elute as homodimers (see Figure **1.8**a). However, it remains unclear whether the active form is a dimer or if a monomer could also be physiologically relevant.[37]

Class 2 ASNases also belong to the N-terminal nucleophile (Ntn) family that are characterized by the enzymatically inactive precursor (uncleaved), that becomes activated upon the autoproteolytic maturation (cleavage).[38] Autoproteolysis involves the removal of a pro-peptide or the cleavage of a precursor chain, with the autocleavage mechanism varying among different types.[40] As a result of the cleavage reaction, α and β subunits are formed within each protomer and the enzyme adopts an αββα sandwich structure (see Figure **1.8**c). Once cleaved, the enzyme gets the N-terminal residue of the β-subunit to become free to act as a nucleophile in the hydrolysis reaction. As depicted schematically in Figure **1.8**b and Figure **1.8**c, in the case of gpASNase3 cleavage occurs immediately after the flexible loop region, specifically between residues Gly167 and Thr168. Unfortunately, this flexible loop region (grey part on the Figure **1.8**b) is not resolved in the PDB structures due to the lack of well-defined electronic density. Further implications of the cleavage to the enzymatic mechanism are given in the next chapter.

[2] The values in the Table 1.2 correspond to the kinetic parameters determined at the pH=7.5 at 37°C.

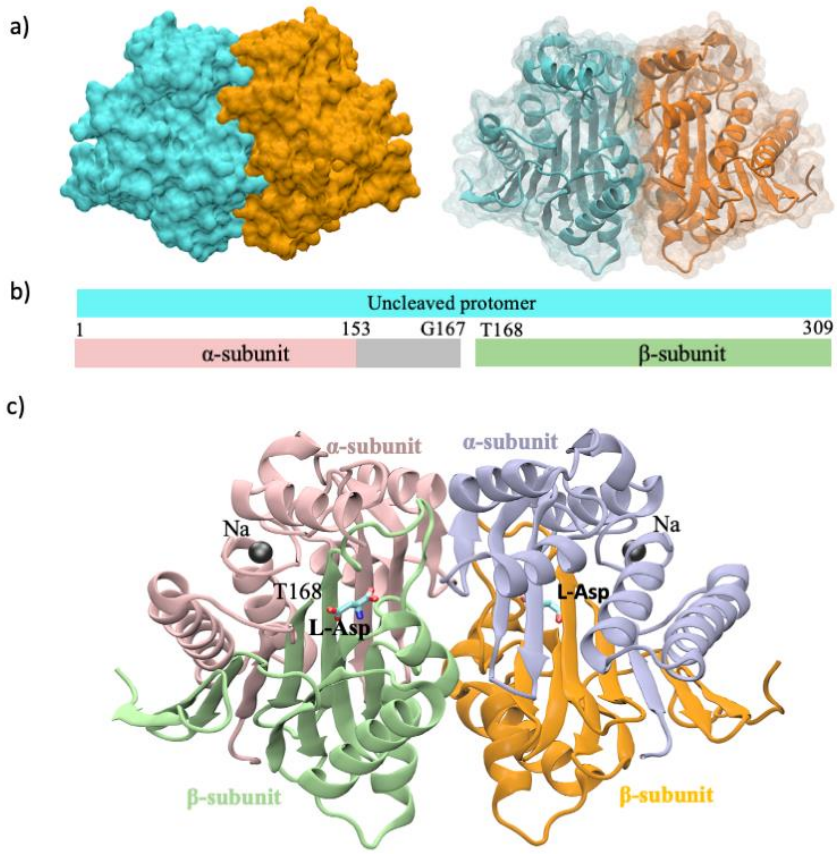


Figure 1.8. a) Homodimeric structure of the mature gpASNase3. PDB code: 4O48 [36]. b) Schematic representation of the sequence for the uncleaved and cleaved protomer. Colors of the rectangles correspond to the colors of the subunits of protomer A given on the panel c, while the gray part corresponds to the residues spanning the flexible loop region (153−167). c) Two subunits formed within each protomer along with the active sites and two sodium ions in the gpASNase3. Cleavage site is labeled as T168.

Two alkali-metal binding loops (stabilization and activation loop) also form an important structural part of class 2 ASNases (Figure **1.9**). All class 2 ASNases have a stabilization loop that coordinates a Na^+ ion (black sphere in Figure **1.8**c). However, they are further classified into K^+ dependent and K^+ independent ASNases based on the role of the activation loop (see Figure 1.3). In simple

terms, this activation loop can switch the enzyme between ON/OFF states in the presence/absence of potassium cation. [41] When K^+ is coordinated within the activation loop, the side chain of active site Arg, crucial for substrate binding, adopts a conformation that anchors the substrate or product in the active site. In the absence of K^+ or when replaced by Na^+, the activation loop rearranges, shifting the enzyme into the OFF state.[41] Interestingly, in potassium-independent ASNases, a structural element similar to the activation loop is present. However, it lacks the ability to coordinate metal ions.

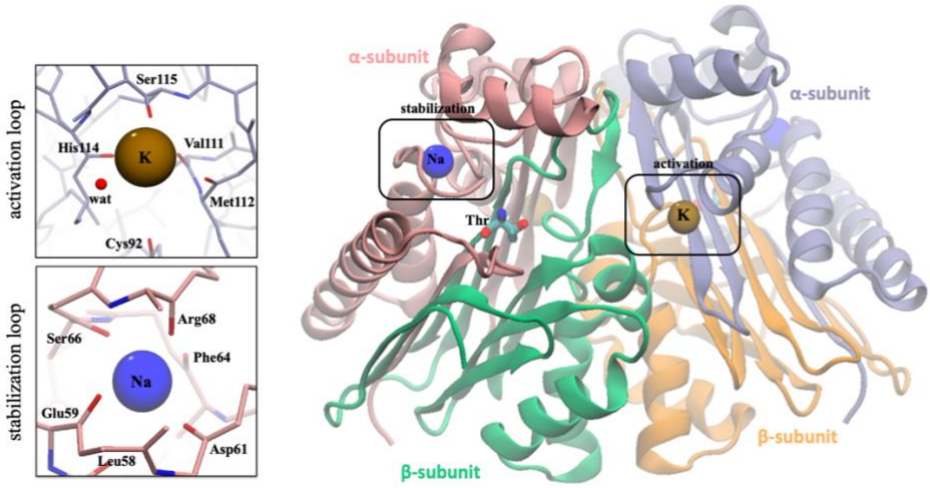


Figure 1.9. The crystal structure of the mature potassium-dependent ASNase from P. vulgaris (PDB code: 4PV2 [41]). Detailed view of the Na^+ (blue sphere) and K^+ (brown sphere) coordination is given on the left.

1.2.4 Class 3 L-ASNases

Class 3 ASNases (formerly known as *R. etli* ASNases) have been significantly less explored compared to the extensively studied class 1 and class 2 enzymes. *R. etli* is a soil bacteria that colonize the root of the legume *Phaseolus vulgaris* (common bean), in which they fix atmospheric nitrogen for its ultimate use in the form of ammonium by the plant cells.[23]

Two types of class 3 ASNases can be differentiated: constitutive and inducible ASNases. Sequences of both these ASNases differ so significantly from class 1 and 2 ASNases (sequence identities around 15-16%) that are labeled as type 4 and type 5 ASNases, respectively.[23] Moreover, the absence of homologs for these two types in the *E. coli* genome further substantiates their distinct classification. This likely indicates that these two types have evolved separately from the typical bacterial and fungal asparaginases. Constitutive (thermostable) form, exhibits high stability and migrates faster on native gel, while inducible (thermolabile) form has a life-time of only 11 minutes at the same temperature (50^{o}C) and is less mobile on the native electrophoretic gel.[42]

1.3 What makes a good therapeutic L-asparaginase?

1.3.1 Glutaminase activity

The close structural similarity between glutamine and asparagine, differing by only a single carbon atom, makes it challenging for ASNases to be selective in their activity. Therefore, in addition to depleting asparagine, these enzymes can also deaminate L-glutamine to L-glutamate and ammonia. The consequences of this secondary activity for therapeutic efficacy in treating ALL remains a topic of debate. Some studies suggest that glutamine depletion may be vital for the effectiveness of L-asparaginase against ALL cells [43, 44], while others caution that glutamine depletion could lead to harmful side effects. [24-26] The only consensus seems to be that asparaginases with higher glutaminase activity than *E. Coli* and *Erwinia Chrysanthemi* ASNase have been linked to more severe side effects and toxicity in patients.[48] Similarly, some studies have shown benefits of the glutamine supplementation to chemotherapy [43, 49, 50], while others found that there was no benefit or even harmful effects, due to the compromised gut integrity, such as fever or toxicity. [51, 52] These contradictory results are likely due to the publication and funding bias and/or the limitations of small-scale studies with few participants.[53] Anyhow, in recent years, there is high growing interest in exploring glutaminase-free therapies as potentially safer alternatives to conventional treatments.[33-36]

1.3.2 Catalytic efficiency

Apart from the selectivity towards asparagine, what are the qualities that make a good therapeutic ASNase? As in other cancer treatments, the effective chemotherapeutics are those that rapidly kill as many malignant cells as possible and achieve a successful remission without severely affecting the healthy cells. In case of ALL, this means that an effective L-asparaginase therapeutic must possess sufficient activity to deplete serum asparagine while maintaining a practical dosage level.[48] The activity-to-dose relationship of an enzyme is described by the Michaelis-Menten kinetics. The kinetics is described by the Michaelis constant (K_M) and the catalytic rate constant, also called turnover

number, (k_{cat}). Namely, for the enzymatic reaction of an enzyme E binding to the substrate *S* to form a substrate-enzyme complex $E{\cdot}S$ that further releases a product *P* regenerating the original form of the enzyme, we can write:

$$S + E \underset{k_{-1}}{\overset{k_1}{\leftrightharpoons}} E{\cdot}S \xrightarrow{k_{cat}} E + P \tag{1.1}$$

where k_1 and k_{-1} stand for forward and reverse rate constants of the binding step, respectively. Under the assumption of the steady state, the reaction rate is given by:

$$v = \frac{v_{max}[S]}{K_m + [S]} \tag{1.2}$$

where *[S]* stands for the substrate concentration and v_{max} is the maximum reaction rate achieved by the system. At low substrate concentrations, such as L-asparagine in this context, the enzyme's catalytic sites remain available to bind the substrate, causing the reaction rate to increase rapidly as substrate concentration rises. Once all catalytic sites are fully occupied, the reaction rate reaches a plateau, defined as the maximum rate ($v_{max} = k_{cat}[E]_0$). An enzyme with a low k_{cat} can achieve significant reaction rates at low substrate levels, demonstrating high affinity for the substrate. When it comes to the Michaelis constant ($K_M = (k_{cat} + k_{-1})/k_1$), enzymes with low values are more efficient at low substrate concentrations, whereas the high K_M enzymes perform less effectively under these conditions. Given the concentration of Asn in human blood (around 50 μM [58]), a good ASNase candidate must have a sufficiently low K_M (approximately ≤ 50 μM) to act efficiently at human blood L-asparagine levels.

1.3.3 Immunogenicity and hypersensitivity

For almost 50 years now, bacterial ASNase enzymes have actively been used in ALL treatments. *Escherichia coli* type 2 enzyme (EcASNase2, tradename Elspar ®) and *Erwinia chrysantemi* (ErASNase2, tradename Erwinase ®), both FDA

approved drugs, have shown to have good kinetics parameters that were previously described.[59, 60] However, being of bacterial origin, treatment with both of these enzymes has been noted to cause immunogenetic, hypersensitivity and toxic reactions in patients.[61, 62] Therefore, the standard therapeutic procedure is that once the patient develops a strong immunogenic reaction to asparaginase derived from *E. coli*, their treatment is transitioned to asparaginase sourced from ErASNase [63] as it does not cross react with EcASNase antibodies.[48] This is mainly due to the fact that anti-asparaginase antibodies can lead to the rapid clearance of the enzyme from the bloodstream, reducing half-life of ASNases *in vivo*.

1.3.4 Antigenicity

To reduce antigenicity, EcASNase epitopes have been masked using polyethylene glycol (PEG). Pegylation is a common technique involving the covalent and noncovalent attachment of polyethylene glycol (PEG) to a biopharmaceutical, enhancing the drug's hydrodynamic radius, extending its plasma retention time and shielding antigenic determinants from immune detection.[64] In 2006, FDA approved pegylated EcASNase2 (trade name Oncaspar ®). Results shown that pegylated EcASNase exhibits a significantly longer half-life of 5.7 days compared to both ErASNase and unpegylated EcASNase (half-lives of 0.65 days and 1.28 days, respectively).[48] Attempts to pegylate ErASNase have been less successful, as hypersensitivity reactions persisted, and the clearance of PEG- ErASNase from the bloodstream turned out to be faster than without pegylation. [65] Consequently, research shifted towards asparaginases of fungal plants and mammalian origin, which are discussed in greater detail in the following section. Some mammalian enzymes, such as guinea pig ASNase type 1 (gpASNase1), have shown remarkable catalytic properties and seemed to be a promising alternative to the bacterial enzymes.[28]

1.3.5 Thermal stability

Thermal stability also plays an important role for therapeutic asparaginases. Since asparaginases are also used in industries (e.g. reducing acrylamide in food

[66]), thermostable ASNases are also very beneficial for industrial use. When it comes to clinically relevant ASNases, thermal stability can be correlated with the long-term storage stability and increased shelf life. In the pursuit of developing more effective and durable therapeutic asparaginases, two main approaches have been explored to enhance thermal stability: structural/rational and biological approaches.

The rational approach involves a targeted modification of the structure of the enzyme. One of the most employed strategies is to increase the number of hydrogen bonds, disulfide bonds, salt bridges, etc. within the enzyme.[67–69] These additional interactions can help to stabilize the tertiary and quaternary structures of the enzyme, reducing the likelihood of denaturation under heat stress. For example, Li and co-workers introduced a single point mutation within the flexible lid of the *Bacillus subtilis* ASNase (BsASNase). These mutations enhanced thermal stability, while catalytic efficiency remained unaffected.[70]

On the other side, a biological approach refers to the use of thermophile bacteria as a source of ASNases. Thermophilic enzymes are naturally adapted to function at elevated temperatures and possess inherent stability under such conditions. By sourcing asparaginase from these organisms, researchers obtain enzymes with exceptional heat tolerance that could be used directly or as a foundation for further engineering.[71, 72] However, the majority of studies on thermostable ASNases focus on temperatures that can be relevant for industrial applications (> 65^{o}C) while significantly exceeding those clinically relevant (human body temperature and common storage temperature).

1.3.6 Possible alternatives

In the search for clinically relevant ASNases, apart from the wild-type enzymes sourced from natural organisms, several alternative strategies have been explored. One notable approach involves the development of *in silico* driven engineering of asparaginases.[73–75] Some of these studies have led to a reduced enzyme flexibility and enhanced substrate binding [75], while others achieved reduced immune response.[73, 74] The most promising attempt was the

humanization of asparaginases coming from other organisms. Lavie and colleagues, employed directed evolution and DNA shuffling techniques with guinea pig asparaginase type 1 (gpASNase1) and human asparaginase type 1 (hASNase1), successfully generating two chimeric enzymes that exhibited high catalytic activity and sequence similarity to the human enzyme.[29] These chimeras have been patented recently, highlighting their potential for clinical use (patent number US10821160B2).

1.4 Previous Theoretical Work on L-asparaginases

Theoretical studies on ASNases have predominantly focused on elucidating the catalytic mechanisms underlying their enzymatic activity. Even though some contributions have proposed mechanisms based on structural data without exploring the corresponding reaction pathways, this chapter will briefly summarize only theoretical contributions as well as their limitations and shortcomings.

Two distinct reaction mechanisms have been proposed for class 1 ASNases: the direct displacement mechanism and the ping-pong or double displacement mechanism (Figure **1.10**).

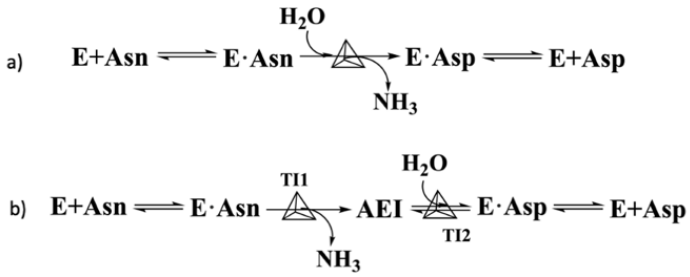


Figure 1.10. a) Single-displacement mechanism. Direct nucleophilic substitution by a water molecule and liberation of ammonia; b) The double-displacement mechanism involves a sequence of two nucleophilic substitution reactions, each encountering an energy barrier due to the formation of tetrahedral intermediates (TI1 and TI2). These steps are separated by the formation of a covalent acyl-enzyme intermediate (AEI). Adapted from [76].

In the direct displacement mechanism, a water molecule (nucleophile) directly attacks the substrate asparagine (Asn). This results in the cleavage of the amide bond, liberation of ammonia and the release of aspartate (Asp). In contrast, the double displacement mechanism involves a two-step process. Initially, the enzyme displaces the amino group of the Asn side chain, leading to the formation of a covalent enzyme-substrate intermediate. Subsequently, a water molecule attacks and displaces this intermediate, resulting in the release of Asp as the

product. This mechanism is more complex and involves the transient formation of a covalent intermediate that plays a key role in the reaction.

One of the first theoretical studies on ASNases is the one carried out by Maria João Ramos and collaborators in 2013.[77] Their work focused on the reaction mechanism of EcASNase2 using hybrid quantum mechanics/molecular mechanics (QM/MM) approaches, specifically using the ONIOM methodology. They employed a cluster model containing around 30 residues surrounding the substrate (asparagine), using a two-layer approach to balance computational efficiency and accuracy. In the geometry optimization, the inner layer, was described at the B3LYP/6-31G(d) level, while the outer layer was treated with lower accuracy, AM1 level of theory.[77] Similarly, single-point energy calculations were then performed at the M06-2X/6-311++G(2d,2p) level for the inner layer and the M06-2X/6-31G(d) level for the outer layer.[77] Lavie and coworkers later on provided experimental data supporting this mechanistic proposal given by João Ramos for gpASNase1.[33] The mechanism suggests an unprotonated lysine in the active site acting as the base that deprotonates a water molecule that attacks the substrate (see Figure 1.11). However, the validity of the protonation state of Lys residue remains highly questionable.

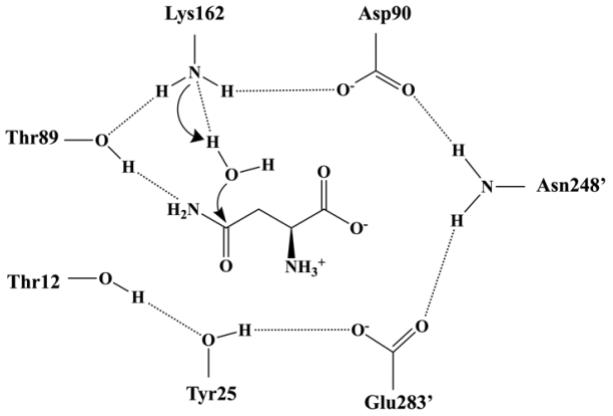


Figure 1.11. Mechanistic proposal for the EcASNase2. The residues with the prime denote residues from the other protomer of the intimate dimer. Adapted from [77].

Lubkowski and collaborators published a combined theoretical and experimental study, offering strong evidence for a double-displacement mechanism in EcASNase2.[76] As a primary evidence in support for the ping-pong mechanism authors uses a structural study of the T89V mutant of the *E. coli* type 2 enzyme.[78] Namely, X-ray diffraction data from the T89V mutant grown in the presence of aspartate revealed the formation of a covalent bond between the enzyme and the side chain of this amino acid. Overall, the mechanistic proposal for the double-displacement mechanism suggested by Lubkowski provides detailed insights into the transition states and intermediates.[76] However, this theoretical work was done employing DFT calculations on a cluster model comprising only of the few amino acids near the active site.

Type 1 ASNases were suggested to follow the same reaction mechanism as type 2 ASNases.[33] However, just recently, Sanchez and colleagues, published a different mechanistic proposal from the one given by Lavie et al. for gpASNase1.[33] Interestingly, their mechanistic proposal is quite similar to that proposed by Lubkowski [76] and it is in quite good agreement with experimentally measured parameters.[32] These authors used a QM/MM multiscale methodology where the energy and structure optimizations were determined at the M06-2X+D3(0)/6-311+G(2d,2p)//CHARMM36 level of theory. The mechanistic proposal involves the activation of a tyrosine residue (Tyr308'), present in the active site loop belonging to the neighbor monomer (see Figure 1.6), through a network of water molecules bridging it to Asp117 (see Figure 1.12).

Once deprotonated, Tyr308' facilitates the activation of a nucleophile (Thr19), which subsequently attacks the Asn ligand (Figure 1.12). A proton is transferred from the nearby Thr116/Lys188 dyad to the amino group of the substrate, leading to the formation of an acyl-enzyme complex and ammonia. Finally, a water molecule, activated by the Thr116/Lys188 dyad, hydrolyzes the complex attacking the carbonyl carbon atom. Interestingly, this mechanism also explains the necessity of closing the active site loop for the catalytic activity. Only when the loop is closed Tyr308' gets close enough to be able to activate the nucleophile (Thr19). Although plausible and reasonable, this study raises some doubts: only

dimer (instead of the homotetramer) was considered and significant restraints were imposed in the MD simulations in order to maintain catalytic distances.[32]

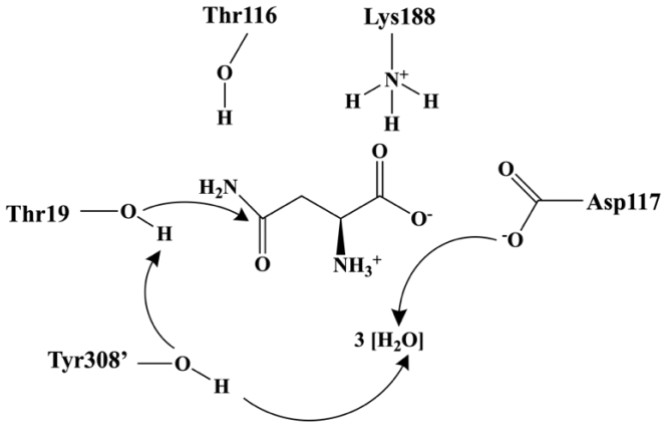


Figure 1.12. Mechanistic proposal for the enzymatic reaction in the gpASNase1. Adapted from [32]. The prime sign denotes the residue belongs to the other protomer.

In 2021, Guimarães and collaborators advanced the understanding of class 1 ASNases by performing classical molecular dynamics simulations on a model of human ASNase1 (hASNase1).[79] Their study aimed to elucidate the structural determinants of enzymatic activity and provided valuable insights into the dynamic behavior of the enzyme. This work represented a significant step forward in understanding the unique features of class 1 ASNases. However, in the absence of an hASNase1 crystal structure, this study relies on a model built using homology modeling, with gpASNase1 as a template. Despite the high sequence identity between hASNase1 and gpASNase1, hASNase1 appears to undergo additional large-scale conformational changes similar to those of EcASNase1 and does not follow classic Michaelis-Menten kinetics (Figure 1.7b). This raises questions about the reliability of using gpASNase1 as a template for homology modelling of the human version.

Class 2 ASNases are believed to follow the mechanistic framework proposed for Ntn hydrolases [78]. However, to the best of our knowledge, the mechanism remains unexplored. Specifically, the absence of a resolved flexible loop region in the X-ray structure complicates the efforts to build a proper model.

Furthermore, while the N-terminal residue is capable of autoactivation[80], it relies on the unprotonated state of a threonine, a condition for which the associated energetic cost has yet to be addressed.

To this date, there has been no detailed investigation into the active forms of either class 1 or class 2 ASNases using reliable computational models that combine a trustable QM level with a flexible environment, leaving a significant gap in the understanding of their functionality. Moreover, existing mechanistic proposals for these enzymes are divided and lack consensus. Furthermore, no theoretical research has been conducted to explore the glutaminase activity of ASNases or to investigate the origins of their affinity towards glutamine. These unsolved questions make ASNase systems particularly intriguing and well-suited for theoretical studies, offering a valuable opportunity to provide insights that experimental approaches have yet to uncover.

Chapter 2: Objectives

The overall objective of this doctoral thesis is to investigate the catalytic properties, reaction mechanisms and dynamics of L-asparaginases (ASNases) of different types. By rationalizing the origins of their catalytic properties and identifying factors that impair enzymatic activity, the results of this thesis aim to guide both rational engineering and *de novo* designs of ASNases with enhanced therapeutic efficacy.

Using classical MD simulations, we aim to explore the dynamic behavior, conformational changes and physiologically active forms of various ASNase types, which remain unexplained to date. Through multiscale simulations, we will analyze the key factors influencing ASNase activity, with a focus on resolving discrepancies in existing mechanistic proposals for different ASNase types. Additionally, we plan to investigate the pK_a values of key catalytic residues using free energy methods, as these are crucial for the mechanistic understanding. Furthermore, we will examine chimeras recently developed through directed evolution to provide evidence that our methods can accurately predict the amino acid positions retained by directed evolution for preserving optimal enzymatic activity.

This thesis also aims to integrate these insights into cutting-edge methods for redesigning native sequences. The objective is to redesign the native sequence of the most promising ASNases to generate variants with enhanced expressibility, catalytic efficiency, stability and selectivity, ultimately yielding variants with improved therapeutic properties. Finally, once experimentally validated, we will revisit the computational chemistry methods and provide a theoretical framework to explain the observed variations in the stability and activity of the redesigned variants.

Chapter 3: Methods

3.1 Classical Molecular Dynamics Simulations: The Basic Idea

"Everything that living things do can be understood in terms of the jigglings and wigglings of atoms."

R. Feynman

Molecular Dynamics (MD) simulations are a statistical mechanics-based numerical approach rooted in Newtonian laws of motion, designed to simulate molecular systems. This method is particularly powerful for studying large systems and complex processes, such as biochemical interactions, as it offers a detailed, atomic-level perspective of molecular behavior with temporal resolution.

The basic principle behind these simulations is Newton's equation of motion. A force $\boldsymbol{F_i}$ acting on each atom of mass m_i causes the atom to accelerate according to:

$$\boldsymbol{F_i} = m_i \cdot \frac{d^2\boldsymbol{r}_i}{dt^2} \tag{3.1}$$

where $\boldsymbol{r}_i$ stands for the position vector of each atom, while its second derivative with respect to time *t*, represents the acceleration. In MD simulations, this force is derived from the negative gradient of the potential energy (E) of the system:

$$\boldsymbol{F_i} = -\boldsymbol{\nabla_i} \cdot E \tag{3.2}$$

Mathematical models of potential energy functions, also known as force fields, account for both bonded interactions (bond stretching, angle bending, torsions) and non-bonded interactions (van der Waals and electrostatics). By combining equations (3.1) and (3.2) and integrating Newton's motion equations over small time steps, both velocities and positions of each atom can be derived at each time step. The atomic positions are then updated accordingly to simulate the dynamics of the system.

Due to the complexity of potential energy functions, Newton's equations of motion are not integrated analytically, but rather applying some of the commonly used numerical integration algorithms such as the Verlet method.[81] This algorithm is designed to efficiently compute the trajectories of atoms, while ensuring stability and conserving energy to a reasonable extent. The integration is carried out at each time step, which is chosen to be small enough to capture the fastest motions in the system and avoid energy drift (usually 1 fs).[82] Sometimes, however, constraint algorithms can be used to restrict fastest vibrations (usually X-H bonds) allowing to increase the timestep to 2 fs.[83]

In MD simulations, periodic boundary conditions (PBCs) are employed to eliminate non-physical effects caused by the finite size of the simulated system. Typically, the system size in MD simulations is of the order of 10^3 nm^3. Without PBCs, boundaries could introduce artifacts, such as unrealistic reflections or distortions in particle behavior. By using PBCs, the simulation effectively mimics an infinite system: when a particle moves beyond one boundary, it re-enters the simulation box from the opposite side. This approach creates the illusion of an infinite three-dimensional grid, preventing boundary effects and providing a more realistic representation of bulk-phase properties.

MD simulations provide a numerical framework for studying the average behavior of a system applying the ergodic hypothesis. In simple terms, the ergodic hypothesis posits that an ensemble average of an observable A, $\langle A \rangle$ - what is typically measured experimentally, is equivalent to a temporal average of the same observable, $\bar{A}$ – which can be obtained by simulating the dynamics of a system over time:

$$\langle A \rangle = \bar{A} \tag{3.3}$$

Namely, the assumption is that a single trajectory will, over a sufficiently long simulation time, explore all the accessible microstates of a system, specifically those that satisfy the constraints imposed to the simulation (such as the conservation of energy or temperature). Therefore, the average value of the observable A over the simulation time t_f, is calculated as:

$$\bar{A} = \lim_{t\to\infty} \frac{1}{t_f} \int_0^{t_f} A(\boldsymbol{p}, \boldsymbol{q}, t)\, dt \tag{3.4}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ represent momentum and position vectors, respectively.

The ergodic hypothesis, however, may not be held in all cases. Particularly when dealing with non-equilibrium or out-of-equilibrium systems.[84] In such situations, the assumption that a single trajectory will explore all accessible microstates over time may not be valid. Factors such as external driving forces, varying temperature conditions, or complex interactions between components can lead to dynamic behaviors that do not allow the system to equilibrate properly.[85, 86] In these cases, special methods and considerations (such as enhanced sampling methods, which will be described in the next chapter) are required to accurately describe the behavior of a system.

Moreover, while classical MD simulations are powerful tools for simulating large biological systems based on classical principles, they may also fall short, especially when it comes to capturing processes such as bond breaking and bond forming during the chemical reactions. In these cases, methods like multiscale quantum mechanics/molecular mechanics (QM/MM) simulations are often used to describe the quantum mechanical nature of the reactive sites. However, the dynamics of the system, including the motion of atoms, remains treated classically in QM/MM simulations. These methods will be further discussed in the corresponding chapter.

3.2 Molecular Mechanics Free Energy Methods

"I'm really interested in the idea of absence and presence, and the tension between the two."

Cy Twombly

3.2.1 Free Energy Perturbation

Free energy differences between two states are more accessible through simulations than absolute free energy. In many cases, these relative free energy differences are sufficient to address the specific thermodynamic or kinetic questions of interest, such as ligand binding, conformational changes, or reaction pathways. This chapter focuses on the calculation of relative free energy differences.

In thermodynamics, the choice between Helmholtz free energy (A) and Gibbs free energy (G) depends on the conditions of the system. Gibbs free energy is used when studying processes at constant pressure and temperature and it is the natural thermodynamic potential for the isothermal-isobaric, NTP, ensemble. This makes it particularly relevant for biochemical reactions, as biological systems, including the human body, that operate under approximately constant temperature and pressure. Helmholtz free energy is, on the other hand, used for systems at constant volume and temperature and it is the natural thermodynamic potential of the canonical, NVT, ensemble. In the MD simulations, the system can be simulated in a fixed-volume simulation box (NVT) ensemble and therefore Helmholtz free energy (A) is the thermodynamic potential used. However, changes in the Gibbs free energy can be approximated by Helmholtz free energy for condensed phases, as variations in the volume are usually negligible.

Having two states I and II, with Helmholtz free energies A_I and A_{II}, the free energy difference in between these states (ΔA) can be defined as:

$$\Delta A = A_{II} - A_I = -kTln\frac{Q_{II}}{Q_I} \tag{3.5}$$

where Q_I and Q_{II} stand for the canonical partition function. The classical expression of this partition function for the set of coordinates and momenta of all particles ($\boldsymbol{r}^N$, $\boldsymbol{p}^N$):

$$Q(N,V,T) = C\int ...\int e^{-\frac{H(\boldsymbol{r}^N,\boldsymbol{p}^N)}{kT}} d\boldsymbol{r}^N d\boldsymbol{p}^N \tag{3.6}$$

where H is the energy, C is a constant ($C = (N!\ h^{3N})^{-1}$), and $d\boldsymbol{r}^N = \prod_{i=1}^{N} d\boldsymbol{r}_i$ and $d\boldsymbol{p}^N = \prod_{i=1}^{N} d\boldsymbol{p}_i$. By introducing (6) into (5) and rearranging the expression we get:

$$\Delta A = -kTln\int ...\int e^{-\frac{H_{II}}{kT}} e^{\frac{H_I}{kT}} \frac{e^{-\frac{H_I}{kT}}}{\int ...\int e^{-\frac{H_I}{kT}} d\boldsymbol{r}^N d\boldsymbol{p}^N} d\boldsymbol{r}^N d\boldsymbol{p}^N \tag{3.7}$$

Or simply:

$$\Delta A = -kTln\int ...\int e^{-\frac{H_{II}-H_I}{kT}} \rho_I\ d\boldsymbol{r}^N d\boldsymbol{p}^N \tag{3.8}$$

The term ρ_I represents the probability density of finding the system in a particular microscopic state (with coordinates $\boldsymbol{r}^N + d\boldsymbol{r}^N$ and momenta $\boldsymbol{p}^N + d\boldsymbol{p}^N$) corresponding to state I. Furthermore, having in mind the definition of the ensemble average, equation (3.8) can be rewritten as:

$$\Delta A = -kTln\,\langle e^{-\frac{H_{II}-H_I}{kT}}\rangle_I \tag{3.9}$$

The last equation states that the free energy difference between states I and II can be calculated as an ensemble average of $e^{-\frac{H_{II}-H_I}{kT}}$ over a simulation run at state I. A similar expression can be derived for a simulation run at state II.

If the energy difference between the two states is significantly larger than the thermal energy, equation (3.9) will converge very poorly. Most of the evaluation

of the exponential energy difference will therefore result in very small contributions to the average. To overcome this issue, intermediate states, with small energy differences between consecutive states, can be introduced, effectively enabling the free energy difference to be calculated as the sum of smaller, more manageable increments:

$$\Delta A = A_{II} - A_I = (A_{II} - A_{N-1}) + (A_{N-1} - A_{N-2}) + \cdots (A_2 - A_I) \tag{3.10}$$

Substituting (3.9) into equation (3.10) we obtain:

$$\Delta A = -kT \sum_{i=1}^{N-1} ln \langle e^{-\frac{H_i - H_{i+1}}{kT}} \rangle_{i+1} \tag{3.11}$$

Intermediate states can be defined by introducing a perturbation parameter λ, that will ensure nonlinear interpolation between two states, by continuous transformation from the Hamiltonian of state I to the Hamiltonian of a state II:

$$H(\lambda) = (1 - \lambda)H_I + \lambda H_{II} \tag{3.12}$$

In this way, by changing the coupling parameter λ from 1 to 0, the Hamiltonian switches gradually from $H(\lambda = 0)$ corresponding to the initial state H_I to the $H(\lambda = 1)$ describing the second state H_{II}. It is interesting to note that the thermodynamics path from I to II does not have to be a "real" physical path, that is why this method is usually called alchemical transformation. Illustration of the alchemical transformation of phenol to the benzene is given in Figure 3.1.

In practice, this transformation can be done via single topology and dual topology approaches. These two approaches define how atoms are treated during the transformation between states. The single topology approach smoothly interpolates between two states by modifying atomic interactions, typically mutating one set of atoms into another, with non-existing atoms becoming "dummy" particles with no interactions.[87] In contrast, the dual topology approach represents both end states simultaneously, turning off the interactions of one set of atoms from the rest of the system. Even though more intermediate steps might be required as more atoms need to be altered along the transformation

in the dual topology approach, it is often more efficient due to its easier convergence.[87]

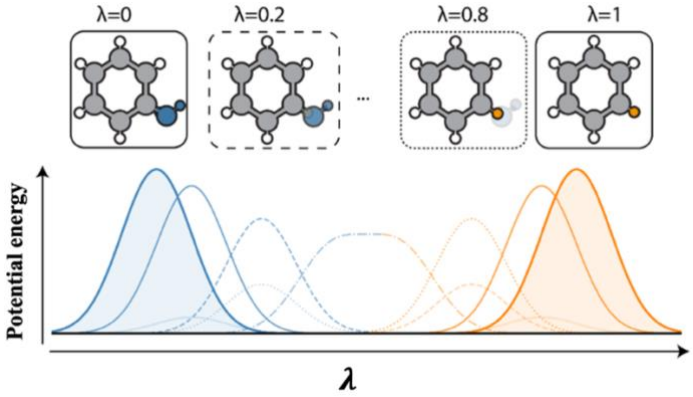


Figure 3.1. (Top) Illustration of the transformation of phenol to benzene and the alchemical intermediates interpolated between the two end points. (Bottom) The probability distribution functions of the potential energies as the switching function takes values from $\boldsymbol{\lambda = 0}$ to $\boldsymbol{\lambda = 1}$. Adapted from [87].

In addition to this, the transformation of the Hamiltonian can be done using two different approaches as well: "one-step" or a "two-step" approach. In the one-step approach, electrostatic and van der Waals forces are modified simultaneously, offering a simpler and more computationally efficient pathway. Two-step method separately modifies these interactions, which can improve sampling in certain cases but increasing the computational effort.[87]

It is important to mention that these transformations can sometimes encounter challenges, especially with extremely low or high values of λ, when creating or annihilating particles. The pairwise interaction between two particles using Lennard-Jones 6-12 potential changes too steeply at small distances, causing convergence problems when the interaction disappears. To address this issue, soft-core potential functions, similar to the one given in (3.13) are commonly employed.[88]

$$E_{soft}(r,\lambda) = 4\varepsilon(1-\lambda)\left(\frac{1}{(\alpha\cdot\lambda+(r/\sigma)^6)^2}-\frac{1}{\alpha\cdot\lambda+(r/\sigma)^6}\right) \qquad (3.13)$$

where α is the soft-core parameter (usually 0.5). These types of soft-core potentials modify the interaction energies to remain finite across all distances, ensuring smooth free energy curves and preventing abrupt changes during the transformation process.

3.2.2 Thermodynamic Integration

By assuming that the Helmholtz free energy, $A = -kTlnQ(\lambda)$, is a smooth and differentiable function of λ, we can write:

$$\frac{dA(\lambda)}{d\lambda} = -kT\frac{1}{Q(\lambda)}\frac{dQ(\lambda)}{d\lambda} \tag{3.14}$$

Since the partition function Q depends on λ exponentially through the Hamiltonian (see equation (3.6)), we can write:

$$\frac{dQ(\lambda)}{d\lambda} = \int \dots \int \frac{\partial H(\lambda)}{\partial \lambda} e^{-\frac{H(\lambda)}{kT}} d\boldsymbol{r}^N d\boldsymbol{p}^N \tag{3.15}$$

Substituting (3.15) into (3.14) we get:

$$\frac{dA(\lambda)}{d\lambda} = -kT\frac{1}{Q(\lambda)}\int \dots \int \frac{\partial H(\lambda)}{\partial \lambda} e^{-\frac{H(\lambda)}{kT}} d\boldsymbol{r}^N d\boldsymbol{p}^N \tag{3.16}$$

Which is nothing but the ensemble average of the derivative of the Hamiltonian with respect to a coupling parameter:

$$\frac{dA(\lambda)}{d\lambda} = \langle\frac{\partial H}{\partial \lambda}\rangle_\lambda \tag{3.17}$$

Finally, the free energy difference can be expressed as an integral of the ensemble average of the derivative of the Hamiltonian with respect to a coupling parameter:

$$\Delta A = \int_0^1 \langle\frac{\partial H}{\partial \lambda}\rangle_\lambda \, d\lambda \tag{3.18}$$

The free energy change can therefore be calculated by running simulations at various values of λ and computing the average value of the Hamiltonian derivative. Integration is usually performed using some common numerical techniques, such as Simpson's rule or Gaussian quadrature.

Thermodynamic Integration (TI) and Free Energy Perturbation (FEP) are powerful tools, particularly in biochemical simulations dealing with residue pK_a calculations, ligand binding free energies, free energy of a single residue mutation, etc.[87, 89] However, they do not easily give insights into how free energy changes along specific coordinates, such as distances or dihedral angles, which are often critical in biochemical studies (e.g. in bond breaking or dihedral conformational changes). For these purposes, the concept of the Potential of Mean Force (PMF) becomes essential.

3.2.3 Potential of Mean Force

A Potential of Mean Force (PMF) allows the determination of the free energy change as a function of a reaction coordinate. In relation to the previous section, the PMF can be calculated by simply assigning the coupling parameter in (3.12) to the coordinate ξ as:

$$\xi = (1 - \lambda)\xi_I + \lambda\xi_{II} \tag{3.19}$$

In general, since the free energy difference between two states, corresponds to the ratio of their partition functions (see equation (**3.8**), the PMF (ΔW) can be obtained by evaluating the ratio just for those configurations presenting particular values of the selected coordinate:

$$\Delta W(\xi_I \to \xi_{II}) = -kTln\frac{\int \dots \int \delta(\xi(\boldsymbol{r}^N) - \xi_{II})e^{-\frac{H}{kT}}d\boldsymbol{r}^N d\boldsymbol{p}^N}{\int \dots \int \delta(\xi(\boldsymbol{r}^N) - \xi_I)e^{-\frac{H}{kT}}d\boldsymbol{r}^N d\boldsymbol{p}^N} \tag{3.20}$$

Given the definition of the averaged probability density of finding the system at a particular value of $\xi(\boldsymbol{r}^N)$:

$$\rho(\xi) = \int \dots \int \rho_{NVT} \delta(\xi(r^N) - \xi_{II})\, dr^N dp^N \tag{3.21}$$

the expression (3.21) can be written as:

$$\Delta W(\xi_I \rightarrow \xi_{II}) = -kTln\frac{\langle \rho(\xi_{II}) \rangle}{\langle \rho(\xi_I) \rangle} \tag{3.22}$$

Or simply as:

$$W(\xi) = C' - kTln\langle \rho(\xi) \rangle \tag{3.23}$$

Namely, one can simply measure the probability that a simulated system visits a configuration with the selected coordinate taking a value in between ξ and $\xi + \Delta\xi$ during the molecular dynamic simulation and from there obtain the PMF.

Similar to the free energy perturbation method, a PMF can also come across difficulties when the free energy between the two values of the coordinate is significantly larger than kT, as both values will not be properly sampled throughout a single simulation. Therefore, some enhanced sampling techniques have been developed, such as the Umbrella Sampling (US) method.

3.2.4 Umbrella Sampling and Weighted Histogram Analysis Method

Umbrella Sampling is an enhanced sampling method based on applying a bias, an additional energy term to the potential energy of the system.[90] In that way, energetically separated regions of phase space can overlap and an efficient sampling across the entire range of the coordinate is facilitated. This can be achieved either within a single simulation or through multiple simulations (windows) with overlapping distributions (see Figure 3.2).

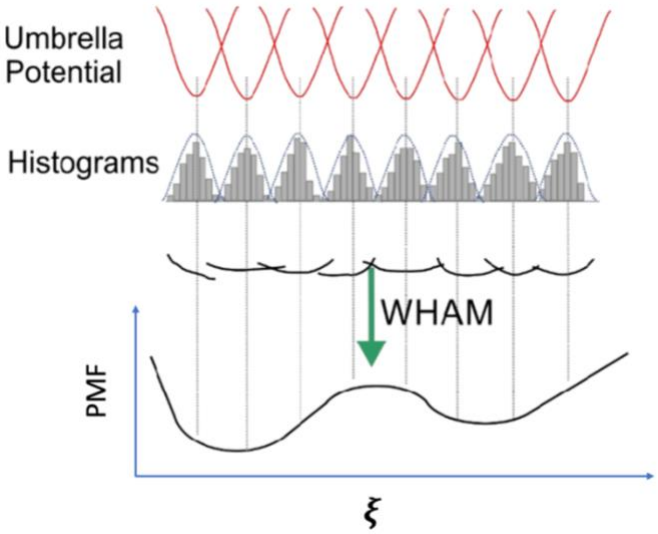


Figure 3.2. Schematic representation of the US and WHAM methods.

A bias, an additional energy term that depends only on the coordinate ξ, is applied to enhance the sampling in the neighborhood of a particular value of the coordinate. The total new energy function corresponds to the sum of the unbiased potential and the biasing potential V_{umb}:

$$H_{biased}(r) = H(r) + V_{umb}(\xi) \tag{3.24}$$

The probability distribution of the biased system is therefore:

$$\rho(\xi)_{biased} = \frac{\int e^{-\frac{H+V_{umb}(\xi)}{kT}} \delta[(\xi(\boldsymbol{r}^N) - \xi] d\boldsymbol{r}^N d\boldsymbol{p}^N}{\int e^{-\frac{H+V_{umb}(\xi)}{kT}} d\boldsymbol{r}^N d\boldsymbol{p}^N} \tag{3.25}$$

To obtain the free energy change along the coordinate, the unbiased probability density distribution must be obtained. This unbiased probability distribution can be related to the biased distribution (3.25):

$$\langle \rho(\xi) \rangle_{biased} = e^{-\frac{V_{umb}(\xi)}{kT}} \frac{\int e^{-\frac{H}{kT}} \delta[(\xi(\boldsymbol{r}^N) - \xi] d\boldsymbol{r}^N d\boldsymbol{p}^N}{\int e^{-\frac{H}{kT}} e^{-\frac{V_{umb}(\xi)}{kT}} d\boldsymbol{r}^N d\boldsymbol{p}^N} \tag{3.26}$$

Given the definition of the ensemble average, equation (3.26) is nothing but:

$$\langle \rho(\xi) \rangle_{biased} = \frac{e^{-\frac{V_{umb}(\xi)}{kT}} \langle \rho(\xi) \rangle_{unbiased}}{\langle e^{-\frac{V_{umb}(\xi)}{kT}} \rangle_{unbiased}} \tag{3.27}$$

The unbiased PMF is then related to the biased distribution by:

$$A_{unbiased}(\xi) = C' - \frac{1}{kT} ln \langle \rho(\xi) \rangle_{biased} - V_{umb}(\xi) + F(\xi) \tag{3.28}$$

where $F(\xi)$ stands for the free energy associated with introducing the biasing potential:

$$F(\xi) = -kT ln \langle e^{-\frac{V_{umb}(\xi)}{kT}} \rangle_{unbiased} \tag{3.29}$$

The biasing potential is chosen to restrict the coordinate within a narrow range around a specific value, facilitating more efficient configurational sampling within a defined range, commonly referred to as a window. The most typical biasing potential is the harmonic, parabolic type function of the form:

$$\omega_i(\xi) = \frac{1}{2} k_{f,i} \left(\xi - \xi_{ref,i}\right)^2 \tag{3.30}$$

where $k_{f,i}$ is the force constant and $\xi_{ref,i}$ is the reference value of the coordinate whose value of the coordinate is changed at the *i* simulation window. This approach enhances configurational sampling efficiency within a particular range. Therefore, a series of simulations are run applying a bias potential to force the configurational sampling around different values of the coordinate. Each of these simulations is called a simulation window. This produces a set of biased distribution functions, each centered on different reference values of the coordinate (see parabolas in the Figure 3.2).

One of the method most commonly used to recover the whole unbiased distribution from the biased simulation windows is the Weighted Histogram Analysis Method (WHAM).[91] In this method, the PMF is computed in such a way that the error of the unbiased probability ($\rho_{unbiased}$) is minimized. The overall distribution is determined by calculating the weighted average of the distributions from the M individual windows:

$$\langle \rho(\xi) \rangle_{unbiased} = \sum_{\alpha=1}^{M} w^{\alpha} \langle \rho(\xi) \rangle_{unbiased}^{\alpha} \tag{3.31}$$

The weights are required to be normalized $\sum_{i}^{M} w^{\alpha} = 1$ and to minimize the statistical error of the total probability distribution:

$$\frac{\partial \sigma^2(\rho(\xi))}{\partial w^{\alpha}} = 0 \tag{3.32}$$

The weights satisfying the previous condition are also satisfying:

$$w^{\alpha} = \frac{n^{\alpha} e^{-\frac{V_{umb}(\xi) - F^{\alpha}}{kT}}}{\sum_{\beta=1}^{M} n^{\beta} e^{-\frac{V_{umb}(\xi) - F^{\beta}}{kT}}} \tag{3.33}$$

where n^{α} is number of independent data points used for the generation of the distribution function of α^{th} window. Therefore, the full distribution function can be written as:

$$\langle \rho(\xi) \rangle_{unbiased} = \sum_{\alpha=1}^{M} w^{\alpha} \langle \rho(\xi) \rangle_{biased}^{\alpha} e^{-\frac{V_{umb}(\xi) - F^{\alpha}}{kT}} \tag{3.34}$$

And F_i can be calculated from:

$$e^{-\frac{F^{\alpha}}{kT}} = \int e^{-\frac{V_{umb}(\xi)}{kT}} \langle \rho(\xi) \rangle_{unbiased} d\xi \tag{3.35}$$

Given that both distribution function and participates F_i are not known initially, equations (3.34) and (3.35) need to be solved iteratively until convergence is reached.

3.2.5 Molecular Mechanics Poisson−Boltzmann and Generalized Born Surface Area Methods

While the previously described free energy, methods offer powerful tools for exploring detailed physical and unphysical pathways, there are cases where the focus is primarily on the properties of the initial and final states. In such scenarios, free energy differences can be approximately estimated solely from the equilibrium conformations of the system without the need to traverse the entire energy landscape. These approaches are typically referred as end-point methods, as they rely on sampling of the initial and final states of a system. End-point methods are computationally cheaper than pathway-based methods, at the cost of introducing approximations that can limit their applicability. The most well-known end-point free energy approaches are Molecular Mechanics Poisson−Boltzmann Surface Area (MM/PBSA) and Molecular Mechanics Generalized Born Surface Area (MM/GBSA), developed by Kollman et al. in the late 90s.[92]

Let's consider the case of a ligand (*L*) binding to a protein receptor (*R*), forming a receptor-ligand complex (*RL*):

$$L + R \rightarrow RL \tag{3.36}$$

In MM/PBSA or MM/GBSA approach, the free energy of binding is obtained as:

$$\Delta G_{bind} = G_{RL} - G_L - G_R \tag{3.37}$$

where G_{RL} is the free energy of the ligand complex, G_L is the free energy of the ligand and G_R is the free energy of the receptor. The free energy in the solution (G_X) can be written as the sum of the free energy in gas phase (ΔG_{gas}) and the solvation free energy (ΔG_{sol}):

$$G_X = \Delta G_{gas,X} + \Delta G_{sol,X} \tag{3.38}$$

The free energy in gas phase can be further decomposed into an enthalpic and an entropic contribution as $\Delta G = \Delta H - T\Delta S$, which makes (3.37):

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{sol} - T\Delta S \tag{3.39}$$

where ΔE_{MM} is the change in the gas phase molecular mechanics energy terms, $-T\Delta S$ is the conformational entropy upon ligand binding and ΔG_{sol} is the free energy change of solvation. Having in mind different interaction contributions, each term can further be rewritten as:

$$\Delta E_{MM} = \Delta E_{int} + \Delta E_{ele} + \Delta E_{vdW} \tag{3.40}$$

$$\Delta G_{sol} = \Delta G_{PB/GB} + \Delta G_{SA} \tag{3.41}$$

$$\Delta G_{SA} = \gamma \cdot SASA + b \tag{3.42}$$

The change in the molecular mechanics energy includes the changes in the internal energies ΔE_{int} (bond, angle, and dihedral energies), electrostatic energies ΔE_{ele} and the van der Waals energies ΔE_{vdW}. The term ΔG_{sol} is a sum of the electrostatic solvation energy ($\Delta G_{PB/GB}$) and non-polar contribution between the solute and the solvent (ΔG_{SA}). The polar contribution is calculated using either the Poisson-Boltzmann (PB) or Generalized Born (GB) continuum models, while the nonpolar energy is typically estimated through the solvent-accessible surface area (SASA).[92] SASA is the area accessible to a probe sphere (representing the solvent) around the solute atoms, with each atom approximated as a sphere with its van der Waals radius.

Practically, snapshots from well-equilibrated molecular dynamic simulations of the complex, ligand and receptor are selected in such a way that are sufficiently separated to avoid autocorrelation. And then, each of the before mentioned terms are calculated in each snapshot. In this way, ΔG_{bind} is estimated as an ensemble of averages:

$$\Delta G_{bind} = \langle G_{RL} \rangle_{RL} - \langle G_L \rangle_L - \langle G_R \rangle_R \quad (3.43)$$

Strictly, averages should be determined from separate simulations of the complex, the free ligand and the unbound receptor. However, it is a common practice to determine the free energies from the simulation of a receptor-ligand complex by removing atoms, and therefore obtaining binding free energy as:

$$\Delta G_{bind} = \langle G_{RL} - G_L - G_R \rangle_{RL} \quad (3.44)$$

In many cases, the internal energy term can cancel out, as it represents the contributions from internal degrees of freedom (DOF) within the system. This simplification reduces the computational cost, as only one simulation is needed. In addition, results converge faster because just electrostatic and van der Waals interactions are needed. Obviously, the method is based in severe approximations and can only be used for well-defined purposes. In this sense, it is advisable to read the recent guidelines published in the editorial guidelines.[93]

3.3 Electrostatic Potential and Electric Field Analysis

Electrostatic forces can be crucial to many biological processes. In the case of enzymes, they are often responsible for maintaining the overall protein structure and facilitating the protein-ligand binding. Most importantly, electrostatic forces often fine-tune the active site environment to stabilize transition states and intermediates.[94–96] That is why the electric field analysis of MD simulations can often provide valuable insights into the origins of the stabilization of the transition state (TS) and driving forces behind enzymatic catalysis.[94, 95, 97]

The electric field ($\boldsymbol{E}$) exerted by N point charges at a specific point in the simulation box can be calculated as:

$$\boldsymbol{E} = \sum_{i=1}^{N} \frac{1}{4\pi\varepsilon_0} \frac{Q_i}{r_i^2} \cdot \boldsymbol{u_r} \tag{3.45}$$

where ε_0 stands for the electric permittivity, Q_i is partial charge of atom i, r is the distance between atom i and the point in space and $\boldsymbol{u_r}$ is the unit vector of the distance.

By defining a point in space (so called probe) and iterating over all *N* atoms in each trajectory frame, one can calculate the magnitude and direction of electric fields. A point in space is usually chosen as a certain atom (atom probe) or a certain coordinate (coordinate probe). A probe can also be chosen as a midpoint of two selected atoms, $\boldsymbol{E}$ as the electric field vector can in that case be projected onto the defined bond ($\boldsymbol{r_{bond}}$ – bond axis vector) as:

$$E_{proj} = \frac{\boldsymbol{E} \cdot \boldsymbol{r_{bond}}}{|\boldsymbol{r_{bond}}|} \tag{3.46}$$

This analysis can also be performed on a per-residue basis, which is highly effective for identifying the "most important" residues determining the electrostatic properties of an enzyme.[98] Furthermore, this feature also allows individual contributions of residues to be summed up within the structural motif

they belong to (e.g., alpha helix, beta sheet), providing valuable insights into the role of protein structural motifs in stabilizing the transition state.

3.4 Quantum Mechanics / Molecular Mechanics Simulations (QM/MM)

Quiero dejar mi huella con profunda
sinceridad, quiero cruzar barreras

P. Almodovar

3.4.1 Multiscale methods in biomolecular systems

In previous sections, we explored classical MD simulations and free energy methods, which can effectively describe large biomolecular systems. However, the analysis of chemical reactions needs of an explicit description of the electrons involved in the bond breaking/forming process and thus a quantum mechanical (QM) treatment. QM calculations for a full enzymatic system can be extremely computationally expensive, especially when many configurations must be sampled to get statistical averages. And while reactive forcefields can sometimes capture certain aspects of enzymatic processes, they are still not as accurate describing the quantum behavior within the active site of an enzyme.[99]

Therefore, one of the adopted solutions is the hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) approach, a concurrent multiscale scheme that combines the strengths of both QM and MM methods.[100–103] In this framework, the quantum mechanical description focuses on the reactive subsystem, including the active site of an enzyme and the substrate. The rest of the system, including other parts of the enzyme, solvent, etc., is treated using classical molecular mechanics, as detailed in the previous section. An illustration of the division of the full system into QM and MM subsystems is given in the Figure 3.3.

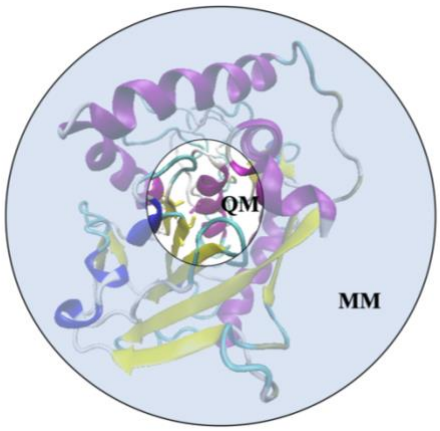


Figure 3.3. Scheme of a QM/MM model.

There are two QM/MM approaches: subtractive and additive schemes.[104] The subtractive approach calculates the energy by subtracting the energy of the QM region, as calculated by a MM method ($E_{MM(QM)}$), from the energy of the total system calculated at the MM level ($E_{MM,total}$) and then adding the energy of the QM region, calculated using a QM method ($E_{QM(QM)}$):

$$E_{total} = E_{MM,total} - E_{MM(QM)} + E_{QM(QM)} \tag{3.47}$$

In the additive approach, the total energy is calculated as the sum of the energy of the QM region (E_{QM}), the energy of the MM region (E_{MM}), and the interaction between the QM and MM region ($E_{QM/MM}$) as:

$$E_{total} = E_{QM} + E_{MM} + E_{QM/MM} \tag{3.48}$$

The QM/MM coupling term can be calculated by taking into account electrostatic ($E_{QM/MM}^{el}$), van der Waals ($E_{QM/MM}^{vdW}$) and covalent (bonded) ($E_{QM/MM}^{bonded}$) interactions between the QM and MM regions:

$$E_{QM/MM} = E_{QM/MM}^{el} + E_{QM/MM}^{vdW} + E_{QM/MM}^{bonded} \tag{3.49}$$

Coupling of the quantum mechanical (QM) and classical molecular mechanics (MM) methods, also known as embedding, ensures that the electronic properties of the QM region are influenced by the classical environment, leading to a more accurate representation.[105] Three main embedding schemes can be used: (i) mechanical embedding scheme, (ii) electrostatic embedding scheme and (iii) polarization embedding scheme.[103, 105] In the mechanical embedding scheme, all interactions between the two subsystems are handled at the force field level. In the mechanical embedding approach, the electronic wave function of the QM subsystem is evaluated without accounting for polarization effects of the surrounding MM environment. The electrostatic embedding scheme includes polarization effects of the QM subsystem. The MM charges are incorporated into the QM Hamiltonian as one-electron operators, allowing QM electrons to perceive MM atoms as pseudo-nuclei, which polarizes the QM electron density.[101] As a result, this approach includes additional terms to the Hamiltonian. A further refinement of electrostatic embedding is a polarization embedding, which explicitly accounts for the mutual polarization between the QM and MM subsystems. This approach allows the QM electron density to polarize in response to the MM environment while simultaneously enabling MM atoms to adjust their polarization in response to changes in the QM and MM subsystems. Various methods have been developed to model MM polarization.[106–108] The total QM/MM energy in this approach requires the MM polarizations to be recalculated at every step of the self-consistent field (SCF) iteration of the QM wave function. Although the most realistic, the polarization embedding using polarizable MM regions have so far mostly been restricted to non-biological systems.[109]

Another challenge arises when both the QM and MM subsystems are connected by chemical bonds. Simply cutting through a QM/MM bond leaves an unpaired electron in the QM region. To address this issue, the simplest and most common approach is to introduce a monovalent link atom (most often hydrogen) at an appropriate point along the bond vector between the QM and MM atoms.[105, 110–112]. Link atoms are included in the QM calculation and are not visible for the MM subsystem. Although a link atom theoretically adds three degrees of freedom to the system, in practice, it is placed at a fixed distance of the QM region

along the bond vector with the MM subsystem at each simulation step, effectively removing these extra degrees of freedom.[111] Forces acting on the link atom are distributed between the QM and MM atoms of the bond using the lever rule, ensuring continuity of forces across the QM/MM boundary.[113]

The final consideration in QM/MM simulations is the choice of the force field to describe the MM subsystem and QM method describing the QM region. This last is crucial as it directly influences the accuracy and efficiency of the electronic calculation that mostly determines the energetics of the process. Due to the computational cost of reliable ab initio calculations, the time scales achievable in QM/MM simulations are limited. Therefore, instead of the ab initio, choices usually fall into the DFT levels e.g., B3LYP [114–117], M06-2X [118], PBE [119] etc., where the maximum timespan is in the order of a few picoseconds, or semi-empirical methods, e.g., AM1 [120–122], DFTB3 [123], GFN2-xTB [124] etc., where the simulation time scale can be extended to approximately two orders of magnitude longer.

In this thesis three QM Hamiltonians have been used: (*i*) B3LYP, a widely used hybrid density functional that combines three-parameter exchange functional with the LYP correlation functional, providing a good balance between accuracy and computational efficiency [125]; (*ii*) DFTB3, that combines elements of DFT and Tight Binding offering a balance between accuracy and computational efficiency.[123] It is particularly effective for large systems, including proteins, DNA, and metal-organic frameworks, offering a practical choice for studying dynamics and reaction pathways;[123, 126–129] (*iii*) GFN2-xTB is a fast semi-empirical tight-binding method designed for large molecular systems.[130, 131] It provides good estimates of geometries and electronic properties, making it suitable for studying complex organic and inorganic molecules, including transition metal complexes.[124, 130, 131]

Finally, when selecting a molecular mechanics (MM) method, several biomolecular force fields are commonly utilized. Popular examples include AMBER force field [132, 133], CHARMM force field [134, 134, 135], GROMOS force field [136, 137] and OPLS-AAforce field [138, 139]. In this

doctoral thesis, we employed the AMBER ff14SB force field [140] for proteins and TIP3P forcefield for water.[141] These force fields are widely used for biomolecular simulations due to their well-validated parameterization and reliable performance in capturing protein dynamics.[140, 142]

3.4.2 Exploring reaction free energy paths: The Adaptive String Method

Enzymatic reactions are generally complex and often involve more than a simple bond being formed/broken. Exploring the free energy landscape typically requires more than two distinguished coordinates to capture the full complexity of the system. One possible approach is to combine two distances (e.g., $r_2 - r_1$) into a single coordinate. However, additional coordinates may still be needed to describe properly the chemical process.[143] As a result of being spanned by more than two coordinates, free energy surface becomes multidimensional. However, the exploration of a high dimension free energy surface significantly affects computational costs, as the cost of exploring the surface grows exponentially with the number of coordinates, making it impractical to include many coordinates. This issue, often referred to as the "curse of dimensionality," limits the applicability of standard free energy calculation techniques like umbrella sampling [90] and metadynamics.[144]

Path-based methods address this issue by projecting the high-dimensional free energy surface (FES) onto a one-dimensional reaction path or a few key collective variables that capture the essential progress of the system. These methods assume that a reaction is most likely to proceed along a specific path (reaction tube). The "reaction tube" refers to a narrow region connecting reactants and product basins, which includes most reactive trajectories.[145] This allows the projection of the free energy space onto a single path collective variable (path CV) that defines the progress along the path. There are slight differences in different path definitions, and the methods used to determine a path vary accordingly. Generally, for a narrow reaction tube and a smooth FES, the minimum free energy path (MFEP) and other path definitions are expected to be similar.[146] One of the widely used and well-established path optimization method is string method. In this method, dynamics are run for each intermediate

state, with the "string" being progressively pulled (converging) towards the Minimum Free Energy Path (MFEP). Once converged, the PMF along the obtained path CV is calculated to estimate the free energy profile. In this doctoral thesis, the Adaptive String Method (ASM) is used to explore free energy profiles of the enzymatic reactions.[147] Therefore, in this section, a brief overview of the theoretical framework underlying ASM will be given.

Firstly, a set of *D* CVs ($\{\theta_1, \ldots, \theta_D\}$) describing all important DOF for the process are defined. Typical CVs include the lengths of those bonds being broken and formed during the reactions, as well as the angles and distances between key atoms or groups. Then a string, vector function $\mathbf{z}(\alpha)$ is defined and discretized in the set of *n* equidistant string nodes $\{\boldsymbol{z}_i, \ldots \boldsymbol{z}_n\}$ (white circles on the Figure **3.4**):

$$z_i = z(\alpha_i) = z\left(\frac{i-1}{n-1}L\right) \tag{3.50}$$

where α represents a parametrization arc-length of the string and L is the total arc-length of the string.

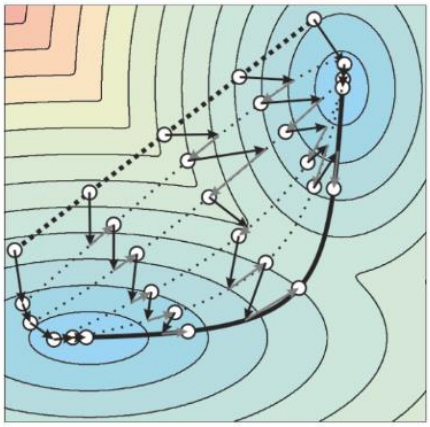


Figure 3.4. Schematic representation of the string evolution. White circles represent the string nodes. Dotted line represents the initial guess, full bold line represents the MFEP, while black and grey arrows represent the evolution and reparameterization of the string.

Independent MD simulations are run at each string node, with a harmonic biasing potential that restrains the system close to the position of the node in the collective variable (CV) space:

$$V_i(x) = \sum_j \frac{K_j}{2}\left(\theta_j(x) - z_{ij}\right)^2 \tag{3.51}$$

where K_j is the force constant for the j CV and z_{ij} is the value of j CV in the i node. This biasing potential ensures that the system remains within a narrow range around the node. This biasing potential can be seen as an ellipsoid with the axis aligned along the basis coordinates of the selected CV space. However, in the ASM, the bias is constructed in such a way to orient the ellipsoid along the string. The nodes evolve according to the:

$$z_i(t + \Delta t) = z_i(t) + \gamma^{-1}\boldsymbol{M}(x_i\,(t))\boldsymbol{K}(\theta(x_i(t)) - z_i)\Delta t \tag{3.52}$$

where $\boldsymbol{K}$ is the diagonal biasing matrix (of a dimension D x D) of force constants, $\boldsymbol{M}$ is the metric tensor, γ is the damping coefficient that determines the "speed" of the node and $\boldsymbol{x}_i\,(t)$ is the trajectory associated with the node $\boldsymbol{z}_i$. Given $\boldsymbol{K}$ and γ are large enough, each node converges led by the free energy gradient to a MFEP (dark bold line on the Figure **3.4**) [148]:

$$\frac{dz(\alpha)}{d\alpha}\Big\| M(z(\alpha))\nabla A(z(\alpha)) \tag{3.53}$$

After every dynamical step, the string nodes are kept equidistant to avoid the nodes to collapse to the free energy minima by interpolating the arc-length as:

$$z_i' = \mathcal{Z}_{z_1,\ldots z_n}\left(\frac{i-1}{n-1}L\right) \tag{3.54}$$

where $\mathcal{Z}_{z_1,\ldots z_n}$ represents a continuous interpolation of the string nodes. In order to enhance sampling, Hamiltonian replica exchange (HREX) between neighboring nodes is attempted every certain number of steps, allowing exchanges of configurations.[149] Instead of calculating the ensemble averages, the dynamics are performed on fly:

$$\alpha_i(t+\Delta t) = \alpha_i(t) + \gamma_\alpha^{-1} K_i^{\parallel} \Delta l_i x(t))\Delta t \tag{3.55}$$

$$K_i^{\parallel}(t+\Delta t) = K_i^{\parallel}(t) + \gamma_K^{-1}\beta^{-1}(\sigma_i^{-2} - (\overline{\Delta l_i^2(t)})^{-1} + \kappa(\overline{\Delta l_\iota(t)}))^2))\Delta t \tag{3.56}$$

The progress along the path ($s(\theta(x))$) and the distance from it $z(\theta(x))$ are given by:

$$s(\theta(x)) = \frac{\sum_{i=1}^{n} \frac{i-1}{\lambda} e^{-\lambda|\theta(x)-z_i|_{M_i}}}{\sum_{i=1}^{n} e^{-\lambda|\theta(x)-z_i|_{M_i}}} \tag{3.57}$$

$$z(\theta(x)) = -\lambda^{-1} \ln \sum_{i=1}^{n} \frac{i-1}{\lambda} e^{-\lambda|\theta(x)-z_i|_{M_i}} \tag{3.58}$$

where λ is the inverse distance between the two nearby z points.

The convergence of the string is assessed using the root-mean-square deviation (RMSD) of the CVs, with respect to the structures of the previous step. Once the string is converged, the path-collective variable *s* is defined along which US method is used to evaluate the free energy profile along this reaction coordinate. Additionally, sampling from each string node during the string optimization can also be used to reweight a corresponding US window along the path CV to obtain PMFs. However, these PMFs are only semiquantitative because the path CV changes during the convergence and an error is introduced in the reweighting scheme.

3.5 Protein structure prediction and *de novo* protein design

"Cherchez ce qui est bon, fort et beau dans votre société,
puis développez à partir de cela. Tournez-vous vers l'extérieur.
Créez toujours à partir de ce que vous possédez déjà."

M. Foucault

Proteins are synthesized by ribosomes as linear chains of amino acid residues, as dictated by genetic information. However, their functionality relies on more than just their sequence; these chains fold into specific three-dimensional structures to acquire their unique properties and perform their biological roles.[150] Proper folding determines their stability, activity and interactions within the cellular environment, enabling them to carry out functions such as catalyzing reactions, providing structural support, or facilitating communication between cells. Misfolding, on the other hand, can lead to loss of function and the development of diseases, highlighting the critical importance of the folding process. Some folding experiments have demonstrated that the information required to determine a protein's folded, native structure is fully encoded within its linear amino acid sequence.[150, 151] According to Anfinsen's thermodynamic hypothesis, the folding information is embedded in the protein's energy landscape, where the native state corresponds to the conformation with the lowest Gibbs free energy.[152] For decades, this concept has underpinned a general strategy for protein structure prediction, combining the sampling of alternative conformations, evaluating their energies, and identifying the structure with the lowest free energy as the native state. This approach, called energy-guided approach, refers to the optimization of the interactions avoiding energetically unfavorable atomic overlaps. These interactions include burying of the hydrophobic residues within the core of the protein, away from the solvent, minimizing the cavity size in water, maximizing van der Waals and electrostatic forces needed the close packing of side chains in the core, respecting of the torsional preferences of the amino acids, etc.

The rapid increase in the number of known protein sequences, coupled with the challenges of crystallization, has posed significant challenges for structural determination using traditional methods and experimental approaches such as X-ray crystallography (X-ray), nuclear magnetic resonance (NMR) spectroscopy and cryo-electron microscopy (Cryo-EM). Additionally, regions of proteins like flexible loops often lack electron density, rendering them particularly difficult to resolve. However, recent advancements in computational tools, particularly machine learning (ML) techniques, have opened new possibilities in the protein structure determination. Groundbreaking approaches such as AlphaFold (AF) [153] and RoseTTAFold (RF) [154] have achieved remarkable accuracy and efficiency, fundamentally transforming our understanding of protein structure prediction, providing new insights in biological and biomedical research.

3.5.1 Structure prediction software

Homology modeling (template-based modeling) uses techniques for protein structure determination that leverage sequence similarity spanning most of the modeled sequence and at least one known structure.[155] These methods are based on the principle that the 3D structures of proteins are more conserved than their amino acid sequences.[156] In practice, this approach allows a structure of a desired sequence to be modeled using resolved structures of sequences with high similarity, identified through methods like BLAST [157] or targeting sequence profiles.

However, a significant portion of sequence data does not share significant homology with well-studied protein families. In such cases, ab-initio protein structure prediction (template-free modelling) can often provide some advantages.[158] These methods rely on physics-based principles to determine protein structures, generating structural models without known structural homologs. They utilize algorithms designed to quickly locate the global energy minimum and a scoring function to select the best conformation among the generated models. A sub method of template-free modelling methods, that is sometimes classified separately, is fragment-based assembly.[159] It involves constructing fragment libraries of varying lengths, each representing a pseudo-

structure. This method uses fragment information rather than entire templates to build protein models. By limiting the number of possible folding patterns, it reduces computational cost. One of the most renowned fragment-based approaches is Rosetta.[160] It follows the classical Metropolis criterion to decide whether a structural fragment should replace a part of the current conformation of the target protein, with the aim of finding the structure that minimizes the global free energy. Hybrid approaches, which combine both template-based and template-free prediction methods, are also worth mentioning.[161]

The most intriguing advancements come from deep-learning (DL) techniques. These methods can (*i*) either rely on multiple sequence alignments (MSAs) for training and prediction or (*ii*) adopt an end-to-end approach that directly predicts protein structures from the target sequence without relying on pre-existing templates or co-evolutionary data. The requirement for high-quality MSAs can often be challenging, especially for novel sequences. Additionally, MSAs can often be computationally very demanding due to the need for extensive database searches. One of the well-known tools of this type is DeepMind's AlphaFold.[153] This method uses a two-step process for protein structure determination that also incorporates coevolutionary profiles to guide model building. On the other side, there are several approaches such as AlQuraishi's deep learning-based tool [162] that employs an end-to-end differential deep learning strategy, achieving state-of-the-art results without utilizing co-evolutionary data or existing templates. Additionally, RF [154] can be used with or without multiple sequence alignments. Despite the fact that these deep learning approaches offer powerful alternatives in protein structure prediction, they also come with some limitations such as their dependence on high-quality MSAs, the risk of overfitting and the "black-box" nature of these methods.[163–165]

3.5.2 Why *de novo* protein design?

Proteins, composed of sequences of n amino acids, present an astronomical diversity of 20^n possible combinations. Yet, nature has only sampled a small fraction of this vast sequence space. This natural limitation opens the door for *de novo* protein design, which allows to explore and create novel protein sequences

beyond those shaped by evolution. Additionally, evolution often prioritize traits for survival rather than pharmaceutical or industrial utility, leaving gaps in attributes such as solubility, stability, or catalytic function. The contributions that came from the Baker Lab have set an extraordinary benchmark in the field of protein design and engineering.[166–169] The application of *de novo* methods gave rise to a variety of newly designed enzymes presenting K_M/k_{cat} values comparable with those of natural enzymes.[166, 167, 170] Additionally, some enzymes have been designed for entirely new chemical reactions without counterparts in nature.[171, 172] Even the native sequence redesign, performed while preserving the native backbone has shown promising results in terms of enzyme expressibility, stability and function of designed enzymes.[168, 169]

The general pipeline for the *de novo* protein design consists of several steps graphically represented in the Figure **3.5**.[173]

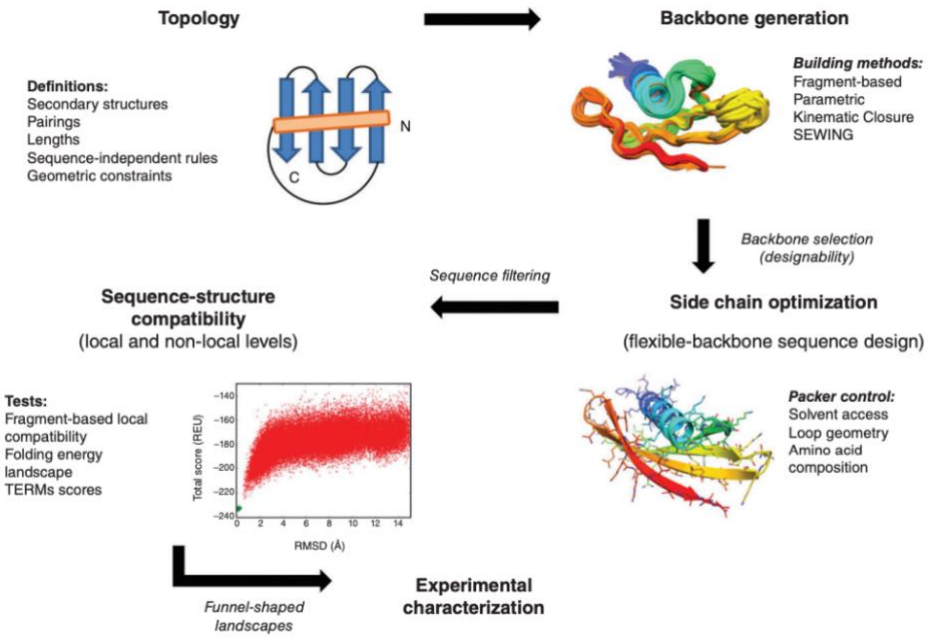


Figure 3.5. *De novo* protein design workflow. Adapted from [173].

The sequence design problem refers to the determination for the sequence with the energy landscapes that favor a previously designed fold. The workflow starts

by defining the target protein topology and generating backbones based on the protein size and folding pattern (see Figure **3.5**). Multiple backbone models are created, refined, and filtered to ensure they align with the design goals. Sometimes, as an additional step, loop regions can optimized before the sequence design is performed.[174] The sequences are then folded using tools like AlphaFold [153] and the structures are compared with the previously obtained ones. The final structures are filtered, and promising designs are then experimentally validated to confirm their structure and function (see Figure **3.5**). The filtering process ensures that only the most promising designs are selected for further experimental validation and are usually case dependent. Some of the filters used are detailed in the section 3.5.2.2. Some of these methods have been used in the present doctoral thesis and will therefore be detailed in the following sections.

3.5.2.1 Backbone generation with RFdiffusion

Whether it incorporates the preexisting experimental information or it starts from scratch (*ab initio* backbone generation methods), backbone generation is a critical first step in the *de novo* protein design. This step creates the structural framework to support the desired functions or topologies. The process typically begins by defining the target protein topology and desired size (amino acid length). Then the backbones are generated either through fragment assembly, which combines structural fragments from existing proteins, or through parametric methods. Parametric approaches involve arranging secondary structure elements followed by a loop closure step to connect these elements into a continuous amino acid chain.

A recent breakthrough in this field is the RFdiffusion method, developed by Baker Lab that provides a significant advance in *de novo* protein design.[175] RFdiffusion is a class of machine learning method based on denoising diffusion probabilistic models (DDPMs), designed to progressively refine data through a series of denoising steps, which enables the generation of high-quality outputs, such as protein structures, from a noisy input.[176] These models are trained to stochastically denoise data (e.g. text or images) that has previously been

corrupted with Gaussian noise. DDPMs can create very diverse outputs and the model can be guided at each step of the iterative generation process towards a specific output. RFdiffusion is built on the previously mentioned RF structure prediction network with minimal architectural adjustments, fine-tuning it for the denoising process. The model was trained by minimizing a mean-squared error loss between frame predictions and the true protein structure, where the protein structures were sampled from the Protein Data Bank (PDB) with the Cα coordinates perturbed with 3D Gaussian noise. A schematic representation of the diffusion model trained to recover a corrupted protein structure is given in Figure 3.6.

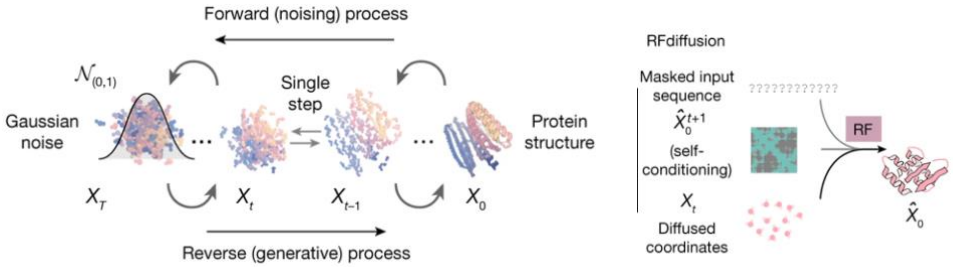


Figure 3.6. (Left) Illustration of the denoising protein backbone corrupted by Gaussian noise. (Right) Schematic representation of the self-conditioning in the RFdifussion model. Adapted from [175].

Diffusion models for proteins transform random noise (X_T) into a realistic structure (X_0), see Figure 3.6. RFdiffusion takes diffused residue frames (coordinates and orientations) as the input and predicts the final 3D coordinates of the structure. RFdiffusion employs a self-conditioning approach, where the model uses its previous prediction as a template input at each step of the denoising trajectory, which typically spans 200 steps. Practically this means that in each time step t, $\hat{X}_0^{t+1}$ and X_t are taken from the previous step and an updated structure X_0 is predicted ($\hat{X}_0^t$).

An important aspect of the RFdiffusion model is the ability to condition outputs on specific positional, distance and functional constraints. This feature enables the incorporation of predefined active sites or a specific structural motif into a *de novo* fold. Additionally, by constraining the position of monomers in a coordinated manner, RFdifussion can generate highly symmetrical oligomers.

RFdifussion also provides a feature for positioning and aligning the generated protein structures relative to a predefined "tip" or anchor atom in the protein (so called centering). This approach can, for example, guide the positioning of the constrained part of the protein to ensure that the constrained motif, for example the active site, is buried within the designed de novo structure. Furthermore, the randomness in the generated outputs can be controlled by adjusting the sampling temperature, allowing the sampling of new designs.

Once backbones are generated and before proceeding to the generation of a sequence that suits the generated folds, RFdiffusion backbones are filtered based on various criteria, such as the radius of gyration, in order to assess the compactness and stability of the protein fold. In case that an active site or another structural motif had been preserved, the position of this motif can also be filtered upon. The ability of a potential substrate to penetrate the active site is assessed to ensure that the generated structure can effectively bind the ligand and/or catalyze a particular reaction. Computational filters are defined based on necessity and tailored to meet specific functional and structural specifications for each case.

3.5.2.2 Sequence generation with ProteinMPNN

The protein sequence design problem refers to the identification of amino acid sequence(s) that will fold into a given protein backbone structure. This process ensures that the designed sequence adopts the desired conformation, maintains stability, and potentially fulfills a specific function. [177] One example of this methodology is Rosetta, which addresses protein sequence design by searching for the optimal combination of amino acid identities and conformations that minimize the energy of an input structure, explicitly considering sidechain rotameric states to achieve accurate predictions. Deep learning methods, which have revolutionized protein sequence design, bypass the computationally intensive evaluation of sidechain conformations, leveraging neural networks to predict sequences directly from structural features.

ProteinMPNN is a deep learning approach for sequence optimization of a given fold built upon the message-passing neural network (MPNN) architecture.[178]

This method has a graph-based neural network architecture that predicts protein sequences in an autoregressive manner from the N to C terminus. The protein structure is represented as a graph, with nodes corresponding to residues and edges representing spatial relationships between them. The neural network consists of three encoder and three decoder layers with 128 hidden dimensions and that uses protein backbone features such as distances between Cα atoms, relative orientations and rotations and backbone dihedral angles as inputs. The model has been trained on the recovery of native sequences of structures from the Protein Data Bank (PDB). Distances between N, Cα, C, O atoms are processed through the encoder to derive graph node and edge features. These encoded features, combined with a partial sequence, are then used to iteratively generate amino acids in a random decoding order. ProteinMPNN has several outstanding features that make it a powerful tool for protein sequence design. For example, certain residues can be fixed while others are designed, facilitating partial sequence redesign. ProteinMPNN also supports positional weights to prioritize certain residues, making it ideal for optimizing active sites or functional domains. Additionally, residues at certain positions can be restricted to a predefined list of amino acids, such as any residue, only charged residues, or other specific categories. This allows for precise control over the design process, enabling functional or structural requirements. Moreover, positional restraints (e.g. distances) can also be added, ensuring that the designed sequences meet desired constraints.

Finally, the ProteinMPNN framework has been generalized to incorporate non-protein atoms, leading to the development of LigandMPNN, a deep learning-based method specifically designed for protein-ligand interface sequence design.[179] LigandMPNN has two additional protein-ligand layers to encode protein-ligand interactions to explicitly model all non-protein components of biomolecular systems, including small molecule ligands, cofactors, or other non-standard residues. By incorporating these components into the sequence design process, it ensures that the generated protein sequences are not only compatible with the backbone structure but also optimized for interactions with non-protein elements. To include the information from ligand atoms to protein residues, a protein-ligand graph is added where protein residues and ligand atoms serve as

nodes. Edges are then created between each protein residue and its closest ligand atoms. Additionally, a fully connected ligand graph is constructed for each protein residue, with the nearest ligand atoms as nodes.

In practice, both ProteinMPNN and LigandMPNN are typically run coupled with Rosetta to refine the design prediction process iteratively. The workflow begins with ProteinMPNN predicting an initial sequence based on a given protein backbone. This sequence then serves as an input for the RF to generate an initial 3D structure. Rosetta subsequently performs structure minimization to refine the predicted conformation. After the minimization step, a new sequence is predicted based on the refined structure. This iterative loop continues, until a satisfactory design or number of designs is reached. Both ProteinMPNN and LigandMPNN generate not only protein sequences but also a structure, predicted by RF. To verify the fold, an additional *in silico* structure prediction method, typically AF, is used. The final designs are then filtered according to the lower root mean square deviation (RMSD) between the two folds and high per-residue local distance difference test (pLDDT). RMSD measures the average deviation of atomic positions in the predicted structure compared to the reference structure, while pLDDT quantifies the confidence metric that reflects the local accuracy of the predicted residue positions in the 3D structure. Higher pLDDT values indicate higher confidence in the predicted accuracy of a fold. Additionally, the structures predicted by AF could also be fed back to ProteinMPNN for additional sequence prediction. This feedback loop between the structures predicted by AF and ProteinMPNN allows for further refinement of the sequence prediction enhancement of the reliability of the sequence prediction.

Chapter 4: Results and discussion

Structure and organization of the Results section

The results section of this doctoral thesis is organized by research projects. At the end of each section, conclusions are provided, followed by a Technical Details subsection. All references are compiled at the end of the thesis.

The research on L-Asparaginases began with the investigation of hASNase3, the findings of which are presented in Section 4.1. At that time, hASNase3 appeared to be a promising candidate for acute lymphoblastic leukemia (ALL) therapy, as it did not trigger the strong immunogenic response observed with the actual treatment with *E. coli* and *Erwinia chrysanthemi* asparaginases, even though the efficacy of type 3 ASNases was limited by poor binding affinities.

As that project progressed, a patent on humanized chimeras, derived from gpASNase1, was published. This patent prompted a shift in focus to gpASNase1 as a promising alternative to the known ALL treatments. As a result, our research line shifted towards the study of gpASNase1 and patented chimeras. The findings related to the gpASNase1, hASNase1 and the chimeras are exposed in Section 4.2.

Finally, to apply previously gained insights to the redesign of the native gpASNase1 backbone, a five-month research stay was conducted at the Baker Lab, Institute for Protein Design (IPD), University of Washington (Seattle, WA, USA). Before directly working on the ASNase redesign project, training in *de novo* design techniques was undertaken through a project on epoxide hydrolases. The results of this training project are presented in Section 4.3, followed by the results of the gpASNase1 sequence redesign in Section 4.4.

4.1 Human Asparaginase type 3 (hASNase3)

Human ASNase type 3 belongs to the N-terminal nucleophile (Ntn) hydrolase family, characterized by enzymatic activation through autoproteolytic cleavage of an initially inactive precursor. This cleavage reaction results in the formation of α and β subunits within each protomer of the dimeric structure (see Figure **1.8**a). More importantly, the cleavage is generating Thr168 as the N-terminal residue (see Figure **1.8**b), which can then act as the nucleophile hydrolyzing the substrate. Structural evidence from the X-ray structure of the acyl-enzyme complex (PDB ID: 4O0H [180]) confirms that Thr168 forms a covalent bond with the substrate. While the reaction mechanisms of some Ntn enzymes have been explored, to our knowledge, no previous theoretical studies on hASNase3 have been conducted.

The structure of hASNase3 also features a conserved sodium-binding loop (residues 55–65) located near the active site within the α chain (Figure **1.8**). This loop has been proposed to stabilize the catalytic nucleophile in several ASNases type 3.[37] However, no theoretical studies have been conducted to confirm this proposal. Finally, despite all available X-ray structures of type 3 ASNases are dimers, the active site residues of hASNase3 originate from a single protomer. This opens the question of the importance of dimerization in the enzymatic activity and whether the monomeric form alone can be catalytically active. Even though this question has been raised by other researchers as well [37], to the best of our knowledge, the correlation between the oligomeric form and reactivity in type 3 ASNases has not yet been addressed.

This section presents the results published by the Andjelkovic et al [181]. First, the results on the dynamic behavior and the active form of hASNase3 are presented, followed by the analysis of the full enzymatic cycle (acyl-enzyme formation, hydrolysis and enzyme regeneration). Then a brief summary of key findings is given. Technical details of this section are given at the end of this section, including comprehensive information about system preparation, molecular dynamics simulations, thermodynamic integration and QM/MM methods.

4.1.1 Dynamic behavior and the active form of hASNase3

To investigate the catalytic relevance of the oligomeric form of hASNase3, molecular dynamics simulations were performed on both the dimeric and monomeric enzyme forms. Interestingly, all three 1 μs replicas of the monomeric enzyme in the Michaelis complex revealed that the substrate consistently left the active site, while the substrate remained bound the whole simulation time in the dimeric enzyme.[181] This behavior can be correlated with the change in the orientation of Arg196 in the active site of hSNase3. Namely, the X-ray structure of the bound substrate (PDB code: 40OH [37]) reveals that the side chains of residues Arg196 and Asp199 form salt bridges with the oppositely charged moieties of the zwitterionic substrate them. Figure 4.1 illustrates the time evolution of the main-chain dihedral angle ψ of Arg196 and the distance between the N-terminal nucleophile, Thr168Oγ, and the electrophilic carbon atom of the substrate, AsnCγ, in the dimeric and monomeric hASNase3 systems during 1 μs of simulation.

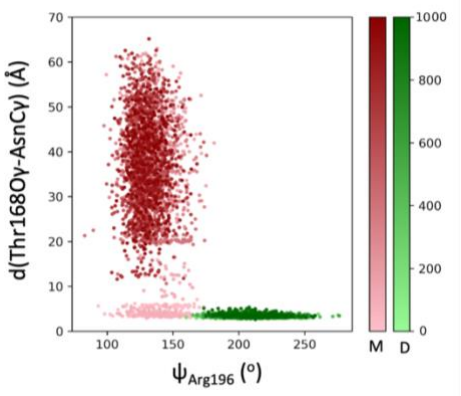


Figure 4.1. Distance between Thr168Oγ and AsnCγ atoms vs. ψArg196 dihedral angle for the monomeric (M, red) and dimeric (D, green) structures. The color intensity represents the time axis in nanoseconds (ns).

This behavior can also be understood considering the increased flexibility of the loop containing Arg196 and Asp199 in the monomeric hASNase3 in comparison to the dimeric hASNase3 (see RMSF comparison in the Figure 4.2).

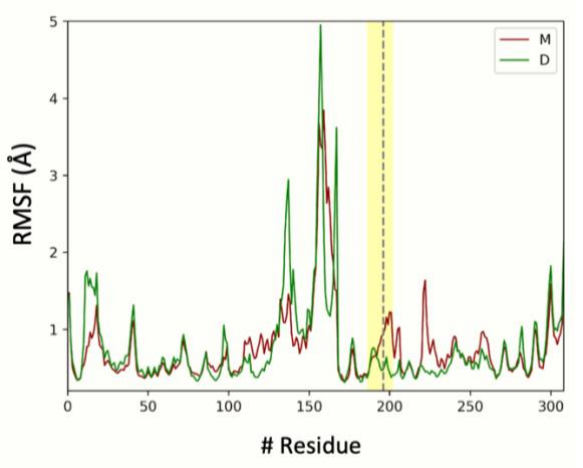


Figure 4.2. Root mean square fluctuation (RMSF) of Cα atoms for all residues within hASNase3. The red line represents the monomeric form (M), the green line represents the dimeric form (D) of the enzyme, the gray dashed line highlights residue Arg196, and the yellow region indicates the entire flexible loop region.

This result can be better understood having in mind the positioning of Arg196 in the case of monomeric and dimeric forms of hASNase3 (Figure 4.3). In the dimer, Arg196 is positioned at the interface between the two monomers, with its side chain consistently oriented toward the substrate. In contrast, in the monomer, Arg196 side chain exhibits rotational flexibility, weakening its interaction with the substrate and allowing the substrate to leave the active site (see Figure 4.3).

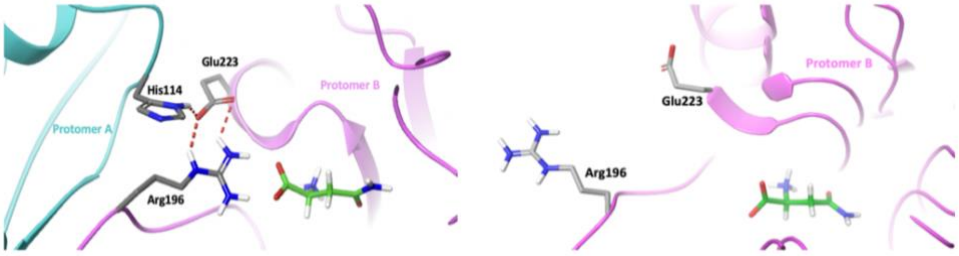


Figure 4.3. Structure of the ON (left side) and OFF state (on the right) of hASNase3, defined by the rotation of the Arg196 and Glu223 side chains.

The stabilization of the Asn-bound state by the neighboring protomer can be explained through a switch between two conformational states, referred to as "ON" and "OFF" states. In the "ON" state, Arg196 forms a hydrogen bond with

Glu223, maintaining the orientation required to coordinate the substrate (see Figure 4.3 left). Additionally, Glu223 is also stabilized in this position by an interaction with residue His114 from the second protomer (see Figure 4.3 left). When the second protomer is absent, Glu223 can move away from Arg196, adopting the "OFF" state. This shift allows Arg196 to reorient from a substrate-facing to a solvent-facing conformation (see Figure 4.3 right). As proposed by Loch and coworkers, the transition between these states in hASNase3 may also be influenced by changes in the protonation states of His114 and Glu223.[23]

All further analysis on this system were done considering the dimeric form of the hASNase3. Key interactions between the enzyme and the substrate, along with the probability distributions of the corresponding distances are depicted in Figure 4.4. As seen from the distributions, active site residues maintained their positions relative to the substrate in all replicas, closely aligning with the observed conformations of the product (PDB code: 4PVS [37]) and the acyl-enzyme intermediate (PDB code: 4O0H [37]). The positively charged amino group of the substrate is stabilized by the carboxylate group of Asp199 and the carbonyl group of Gly220. Additionally, the negatively charged α-carboxylate group of the substrate forms a double salt bridge with Arg196 and a hydrogen bond with the main-chain NH of Gly222.

The carbonyl oxygen of the substrate interacts via hydrogen bonds with the side chain of Thr219 and the main-chain NH of Gly220, forming the so-called oxyanion hole. These interactions stabilize the negative charge on the substrate oxygen atom developed during catalysis and align the NH_2 group of the substrate for an effective leaving group (NH_3) elimination. The substrate's carbonyl carbon atom (Cγ) is located close to the hydroxyl oxygen (Oγ) of Thr168 with an average distance of 3.46 ± 0.44 Å. Thr168 acts as the nucleophile attacking the carbonyl carbon to form the acyl-enzyme complex, supported by Thr186, which orients Thr168 via a persistent hydrogen bond.[37] Additionally, the N-terminal (Thr168) and C-terminal (Gly167) residues, generated after enzyme cleavage, also maintain a strong salt bridge throughout the simulation (evidenced by the last distribution presented in Figure 4.4).

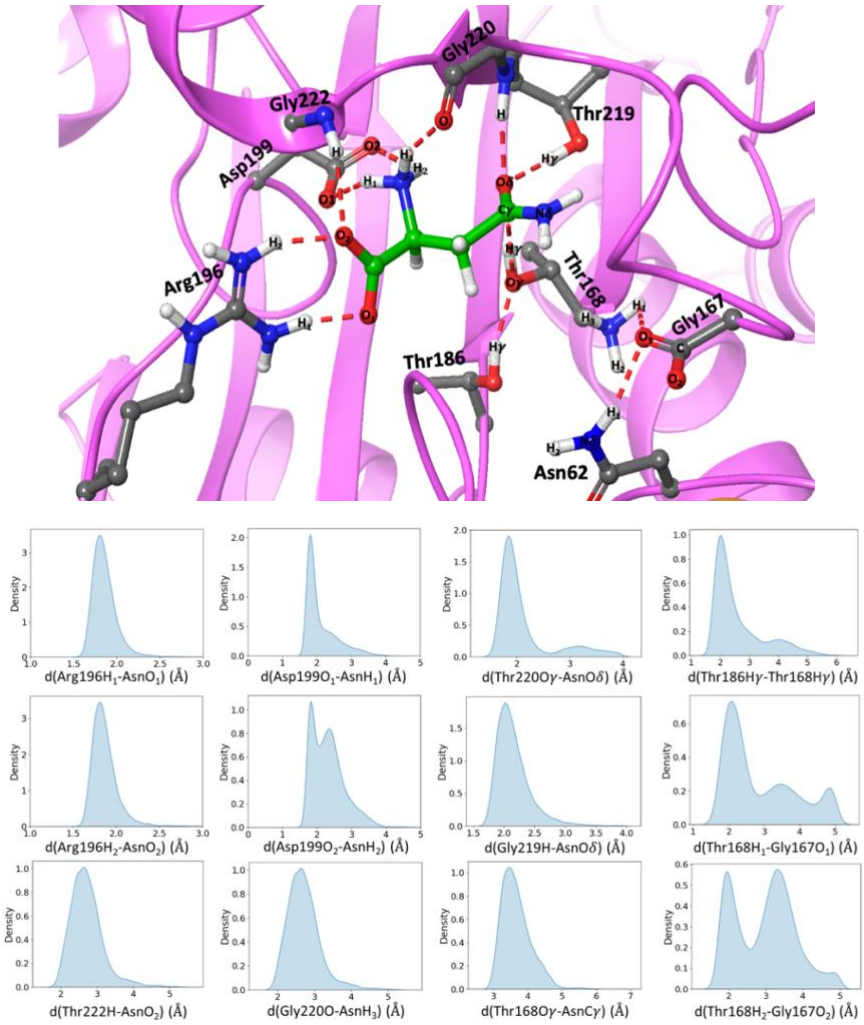


Figure 4.4. Representation of the active site of hASNase3with the Asn substrate (green stick representation) from MD simulations. Distributions of the important distances in (Å) obtained over three replicas of 1 μs of classical MD simulation run on the Michaelis complex.

Importantly, sodium ions were consistently retained within the binding loops (residues 55−65) of both protomers across all replicas in both monomer and dimer forms. These ions stabilize the Na-binding loops, as evidenced by low root-mean-square fluctuations (RMSFs) in comparison to other motifs depicted in the Figure 4.2. The stabilization of the Na-binding loop ensures that Asn62, that

forms a part of the loop and that it is highly conserved in Ntn-type enzymes [37], remains oriented toward the active site (Figure 4.4). In some previous publications, Asn62 was suggested to lock Thr168 in place.[23] However, our results suggest that rather than locking Thr168, Asn62 maintains Gly167 in a position near the active site after the cleavage. Gly167 further stabilizes the protonated amino group of Thr168 when it becomes activated as a nucleophile (see Figure 4.5).

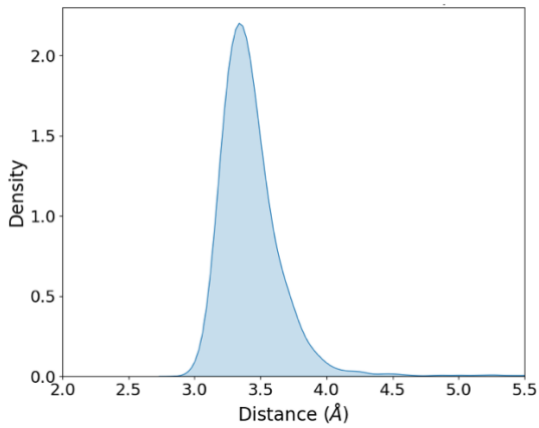


Figure 4.5. Probability density for the distance between the Asn62Nδ and the Gly167C during the simulation of the Michaelis complex.

4.1.2 The reaction mechanism in hASNase3

4.1.2.1 The protonation state of the nucleophile

Due to the conserved fold, with catalytic residues occupying equivalent positions across members,[38] the Ntn-hydrolase family is thought to share a common reaction mechanism, outlined in Figure 4.6.

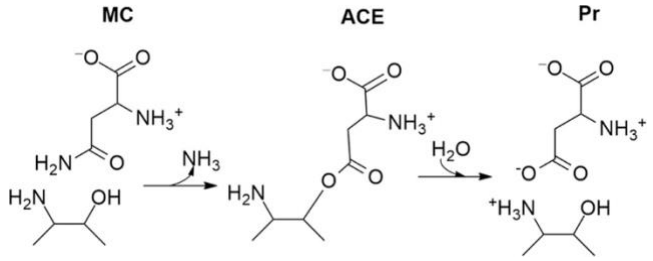


Figure 4.6. General mechanism of asparagine hydrolysis by Ntn-hydrolase enzymes.

In this mechanism, the hydroxyl oxygen of the N-terminal threonine first needs to be activated by proton removal. Once activated, the nucleophilic Oγ attacks the Cγ atom of the substrate, forming a covalent acyl-enzyme bond. This step is accompanied by the cleavage of the Cγ–Nδ bond in the substrate, resulting in the release of ammonia. The acyl-enzyme intermediate (ACE) is subsequently hydrolyzed by a water molecule, yielding the final product (Pr), aspartate (Asp).

Analysis of Michaelis complex simulations indicates that the hydroxyl group of Thr168 is a suitable candidate for the nucleophile attack (Figure 4.4). In order to be activated, as suggested for other Ntn-hydrolases, its own N-terminal amino group could act as a base to activate the nucleophile.[23] To evaluate this hypothesis, the pK_a of the terminal amino group was calculated using a thermodynamic cycle that relates the enzymatic pK_a to the value in aqueous solution (Figure 4.7).

The pK_a shift was related to the change in the interaction free energy of the N-terminal threonine between its protonated and unprotonated states using the following equation:

$$pK_{a,prot} = pK_{a,aq} + \frac{1}{2.303\ RT}\Delta\Delta G \qquad (4.1)$$

Where ΔΔG represents the difference in free energy changes when the amino group of threonine is deprotonated in the protein environment and in aqueous solution, given by $\Delta\Delta G = \Delta G_{(B\text{-}D)} - \Delta G_{(A\text{-}C)}$ (see Figure 4.7). R is the gas constant, T is the temperature, and $pK_{a,prot}$ and $pK_{a,aq}$ refer to the pK_a values of the terminal group in the protein environment and in aqueous solution, respectively.

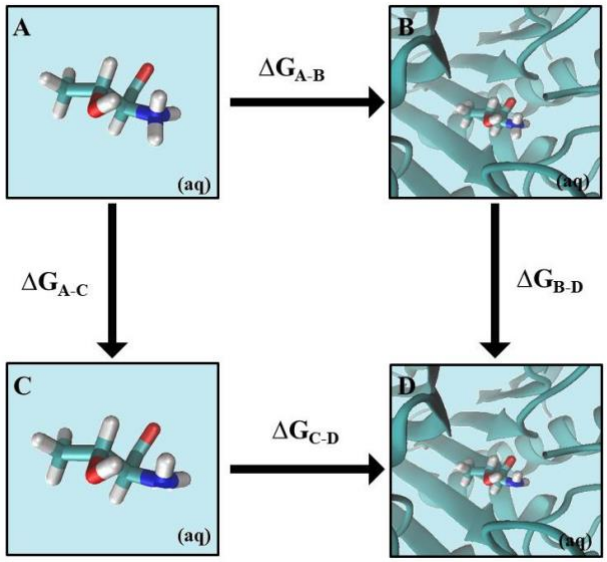


Figure 4.7. Thermodynamic cycle depicting the change in protonation state of amino group of a threonine in water and protein environments. States A and C represent protonated and deprotonated threonine in water, respectively, while states B and D correspond to protonated and deprotonated threonine within the hASNase3 active site. Alchemical transformations were carried out along the vertical axes from A to C and from B to D.

The free energy changes were calculated from the alchemical transformations of the neutral to the protonated residue, both in aqueous solution and in the protein environment. More details are given in the Technical Details section. The calculated pK_a values for the apo and holo forms of the enzyme, as well as for free threonine in water, are provided in Table 4.1.

Table 4.1. Free energy changes computed from five independent replicas (rep) of alchemical transformations between the protonated and unprotonated states of threonine. The calculations were carried out in aqueous solution (aq), as well as in apo and holo protein environments. The indices of the free energy changes ΔG are explained in the Technical Details section. Both the free energy values and the associated standard deviations (std) are given in kcal·mol^{-1}.

	rep	$\Delta G_{A\text{-}C}$		rep	$\Delta G_{B\text{-}D}$		rep	$\Delta G_{B\text{-}D}$
aq	1	-166.34	apo protein	1	-166.05	holo protein	1	-166.71
	2	-166.42		2	-167.02		2	-165.90
	3	-166.58		3	-165.00		3	-165.34
	4	-166.51		4	-165.58		4	-164.70
	5	-166.42		5	-165.12		5	-165.14
mean		-166.45	mean		-165.75	mean		-165.56
std		0.18	std		1.41	std		1.39

From the values given in the Table 4.1, and using for $pK_{a,aq}$ a value of 9.1[182], the final $pK_{a,prot}$ value of the threonine residue are calculated within the apo and holo protein environments. These values are given in the Table **4.2**.

Table 4.2. Free energies of the alchemical transformation of the protonation states of the threonine residue in apo and holo hASNase3 relative to water and corresponding pK_a shifts and pK_a values. Free energy costs of threonine deprotonation at pH 7.5 and T = 310 K and the calculated probabilities of its protonated and deprotonated states are given under the same conditions.

Form	apo	holo
ΔG_{B-D} (kcal·mol^{-1})	-165.75 ± 0.63	-165.56 ± 0.62
$\Delta\Delta G = \Delta G_{B-D} - \Delta G_{A-C}$ (kcal·mol^{-1})	-0.70 ± 0.45	-0.89 ± 0.48
$pK_{a,prot} - pK_{a,aq}$	-0.48 ± 0.33	-0.62 ± 0.34
$pK_{a,prot}$	8.62 ± 0.33	8.47 ± 0.34
$\Delta G(\text{deprotonation})_{pH=7.5,T=310\ K}$ (kcal·mol^{-1})	1.58 ± 0.06	1.38 ± 0.06
$P(\text{protonated})_{pH=7.5,T=310\ K}$	0.92	0.87
$P(\text{deprotonated})_{pH=7.5,T=310\ K}$	0.08	0.13

As seen from the Table **4.2**, the N-terminal group of Thr168 was found to have slightly lower pK_a values in the enzyme ($pK_{a,apo} = 8.6 \pm 0.3$, $pK_{a,holo} = 8.5 \pm 0.3$) compared to the reference value for the free threonine in water ($pK_{a,aq} \approx 9.1$).[182] Despite the presence of the negatively charged C-terminal Gly167 in the enzymatic environment, which could favor the protonated form, the pK_a of Thr168 decreases relative to the aqueous value. To understand the cause of the pK_a shift, additive effects of different groups on the electrostatic potential on the amino group were quantified (see Table 4.3). The shift can be explained by the combined effects of the protein environment and the reduction in the number of water molecules near the amino group, which stabilize the protonated form. The radial distribution function (Figure 4.8) highlights a reduced hydration shell around the N atom of Thr168 in the enzyme compared to the bulk solution. These factors result in a less negative electrostatic potential on the N-terminal group of Thr168 in the enzyme and a slightly lower pK_a.

Table 4.3. Main contributions to the electrostatic potential ($J \cdot C^{-1}$) experienced by the hydrogen atoms of the protonated amino group of Thr168 in the Michaelis complex and in aqueous solution.

Group	Michaelis complex	Aqueous solution
Total potential	-5.60 ± 0.02	-5.95 ± 0.02
Waters	-0.68 ± 0.04	-5.95 ± 0.02
Gly167 COO^-	-1.48 ± 0.05	
Substrate NH_3^+	+0.48 ± 0.04	
Substrate COO^-	-0.53 ± 0.04	

At pH 7.5, the most populated form of the N-terminal group of Thr168 is still predicted to be the protonated one. However, the free energy cost required to deprotonate this group under these conditions is small (1.38 ± 0.06 kcal·mol^{-1}). This result supports the hypothesis that its own amino group could act as a base to activate the hydroxyl group of Thr168 with minimal free energy penalty. To explore this hypothesis, the reaction mechanism was studied assuming an unprotonated N-terminal group for Thr168, incorporating the free energy cost of deprotonation into the reaction free energy profile. The mechanism was studied

in two stages: formation of the acyl-enzyme complex and subsequent deacylation by a water molecule.

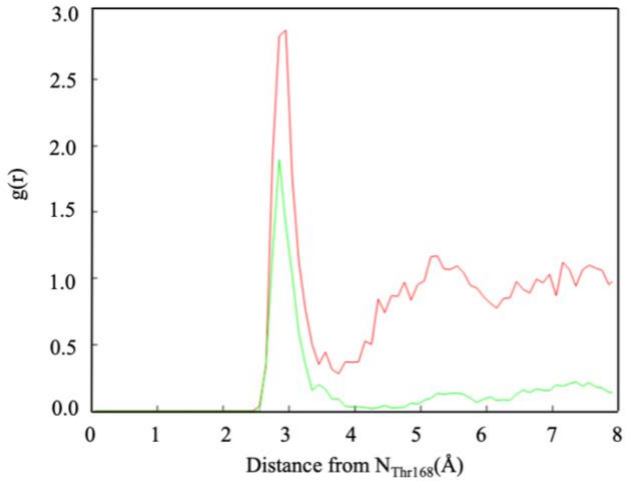


Figure 4.8. Radial Distribution Function (RDF) of water surrounding Thr168 in the Enzyme holo form (Green) and free threonine in solution (Red).

4.1.2.2 Acyl-Enzyme formation

The acyl-enzyme complex formation was explored using the Adaptive String Method [147] at the DFTB3/MM level of theory.[123] The most promising mechanism, in terms of activation free energy, was then recalculated at the B3LYP-D3/MM [125] level with the 6-31+G(d) basis set. Additionally, for this mechanistic proposal, several QM regions were tested that showed no significant differences in the obtained PMFs. Further details are given in the Technical Details section. As shown in Figure 4.9a, the acyl-enzyme formation reaction mechanism involves the N-terminal group of Thr168 acting as a base to deprotonate its own hydroxyl group (Thr168Oγ), followed by the nucleophilic attack of hydroxyl oxygen and finally the proton transfer to the amino group of the substrate. Collective variables (CVs) used to explore the mechanism are given in the Figure 4.9b. Finally, the free energy profile and the evolution of the CVs are given in the Figure 4.9c and Figure 4.9d, respectively. An illustrative movie of the reaction mechanism is provided in the supplementary information material of the original paper or the following link

(pubs.acs.org/doi/suppl/10.1021/acs.jcim.3c00900/suppl_file/ci3c00900_si_002.mp4).[181]

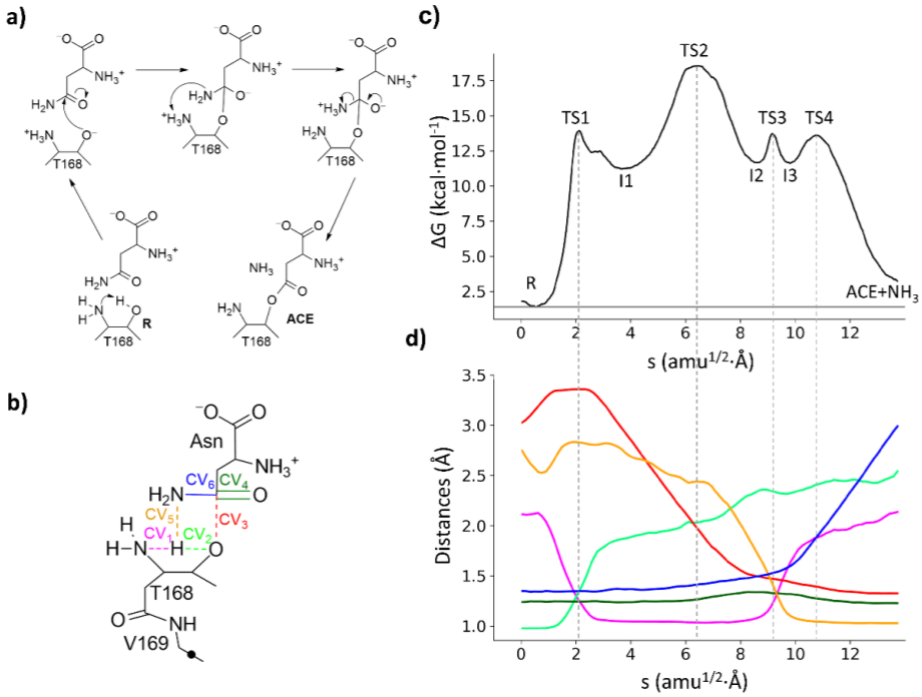


Figure 4.9. (a) Illustration of the mechanism of the acyl-enzyme (ACE) formation in hASNase3. (b) Visualization of the quantum mechanical (QM) region and collective variables (CVs) employed to determine the minimum free energy path (MFEP), with the link atom depicted as a black dot. (c) Free energy landscape along the path collective variable (s), computed at the B3LYP-D3/6-31+G(d)/MM level of theory. (d) Evolution of the chosen CVs along the MFEP, with color codes matching panel (b). Dashed light-grey lines indicate transition state positions.

The previously determined free energy cost for the deprotonation of Thr168 N-terminal group (1.38 kcal·mol^{-1}) must be added to the free energy profile given in Figure 4.9c. The resulting free energies (in kcal·mol^{-1}) of the stationary structures associated with the acyl-enzyme formation and the evolution of key distances along the MFEP (in Å) are presented in the Table 4.4.

Table 4.4. Free Energies (kcal·mol⁻¹) of stationary structures corresponding to acyl-enzyme formation and key distances along the MFEP (Å).

	R	TS1	I1	TS2	I2	TS3	I3	TS4	ACE+ NH_3
Free energy	1.4	14.0	11.2	18.6	11.7	13.7	11.7	13.6	2.7
CV1 (Thr168Nδ-Thr168Oγ)	2.20	1.30	1.89	2.03	2.33	2.33	2.39	2.38	2.10
CV2 (Thr168Hγ- Thr168Oγ)	0.97	1.20	1.07	1.05	1.02	1.20	1.80	2.00	2.53
CV3 (Thr168Oγ-AsnCγ)	3.07	3.36	3.00	1.86	1.43	1.47	1.46	1.36	1.30
CV4 (AsnCγ-AsnOδ)	1.22	1.25	1.23	1.30	1.36	1.32	1.32	1.27	1.23
CV5 (Thr168Hγ- AsnNδ)	2.89	2.93	2.71	2.43	2.00	1.44	1.05	1.00	1.00
CV6 (AsnCγ-AsnNδ)	1.36	1.38	1.38	1.43	1.47	1.50	1.58	1.80	3.00

The free energy barriers relative to the Michaelis complex are 14.0 ± 0.4 kcal·mol⁻¹, 18.6 ± 0.7 kcal·mol⁻¹, 13.7 ± 0.9 kcal·mol⁻¹, and 13.6 ± 0.9 kcal·mol⁻¹ for TS1-TS4, respectively, identifying TS2 as the rate-determining step in the acylation process. The geometries of these transition states are illustrated in Figure 4.10.

The acylation process begins with nucleophile activation via proton transfer from the hydroxyl oxygen to the amino group of Thr168, as shown by the evolution of CV1 and CV2 in Figure 4.9d. This step is assisted by the hydroxyl group of Thr186, which forms a hydrogen bond with Thr168Oγ, reducing their distance from 2.19 Å in the reactant state to 1.70 Å at TS1 (Figure 4.10a). This strong hydrogen bond is maintained throughout the acylation, consistent with experimental observations that the Thr186Val mutant lacks asparaginase activity.[180]

The highest energy barrier in this mechanism corresponds to the TS2, that involves the nucleophilic attack by Thr168 on the Cγ of the substrate atom. In TS2, the Thr168Oγ-AsnCγ distance (CV3) contracts to 1.76 Å, while the

carbonyl bond (CV4) elongates. TS2 stabilization is achieved through hydrogen bonds from the oxyanion hole residues, Thr219 and Gly220 (see Figure 4.10b).

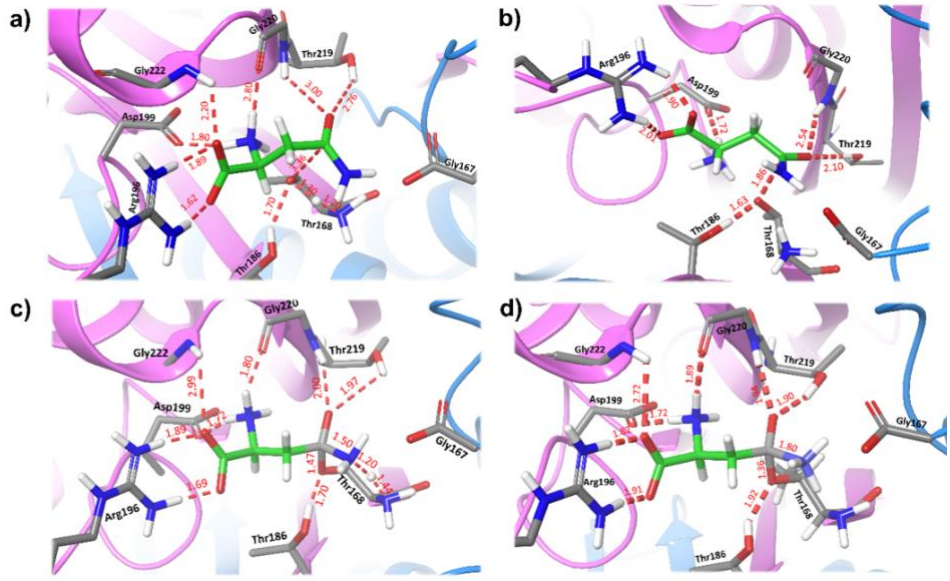


Figure 4.10. Structural representations of transition states involved in the acyl-enzyme complex formation in hASNase3. Panels: a) TS1, b) TS2, c) TS3, and d) TS4, with all distances provided in Å.

The carbonyl oxygen atom forms stronger hydrogen bonds with Thr219 (2.76 Å to 2.10 Å) and Gly220 (3.00 Å to 2.54 Å) during this step. These interactions are disrupted in the Thr219Val and Thr219Ala mutants, which can explain the absence of asparaginase activity in these variants.[180]

Transition states TS3 and TS4 correspond to the proton transfer from the amino group of the nucleophile to the substrate and the subsequent breaking of the Cγ-Nδ bond, leading to ammonia release. CV1 and CV5 track the proton transfer, with the proton positioned in between donor and acceptor nitrogen atoms at TS3 (1.32 Å and 1.20 Å, Figure 4.10c). CV6 monitors the Cγ-Nδ bond breakage, which reaches 1.58 Å post-proton transfer and 1.80 Å at TS4 (Figure 4.10d), before fully dissociating at ACE. Simultaneously, the Cγ-Oδ bond shortens to a double-bond length as shown by CV4. The free energy steadily decreases after

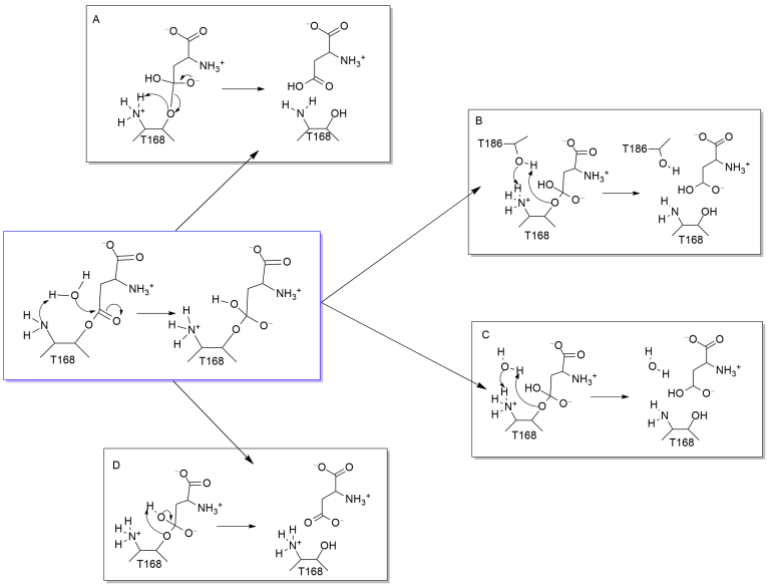
Cγ-Nδ bond cleavage, with the product state being 2.7 ± 0.5 kcal·mol^{-1} more stable than the Michaelis complex.

To complete the enzymatic cycle, the free energy for ammonia release from the active site was calculated using a thermodynamic cycle that accounts for the free energy of moving an ammonia molecule from the active site to the bulk. Detailed information is given in the Technical Details section. The Gibbs free energy for NH_3 removal from the active site is -4.47 ± 0.04 kcal·mol^{-1}, consistent with similar values found for other enzymatic systems.[183] The release of ammonia is driven by the additional hydrogen bonds that NH_3 forms in bulk water with respect to the active site. Including this contribution, the overall reaction free energy for the acylation stage relative to the Michaelis complex with unprotonated terminal amino group is -1.8 ± 0.6 kcal·mol^{-1}.

4.1.2.3 Acyl-Enzyme hydrolysis

The second phase of the enzymatic cycle involves the hydrolysis of the acyl-enzyme (ACE). Initial structures for the string method were prepared by relaxing the acyl-enzyme structure, through 1 μs of classical MD simulations. Parameters for the acyl-enzyme, which features a covalent bond between Thr168 and the substrate, were derived as described in the Technical Details section.

Various hydrolysis mechanisms for the acyl-enzyme complex were evaluated using free energy calculations with the Adaptive String Method at several levels of theory (see Figure **4.11**). These mechanisms considered the participation of one or two water molecules and/or different catalytic residues. The mechanism with the lowest activation energy was further refined at the B3LYP-D3/MM level. A scheme of the most favorable pathway is shown in Figure 4.12a, with the QM region and collective variables (CVs) in Figure 4.12b. The resulting free energy profile (Figure 4.12c) and the evolution of the CVs (Figure 4.12d). Free energy values for stationary points on the PMF (Figure 4.12d) and their corresponding CVs are listed in Table 4.5 while the transition states are depicted in Figure 4.13.



Mechanism	Level of theory	Free energy barrier (kcal·mol^{-1})	QM region
A	DFTB3/MM	19.2	ACE + 1wat
	GFN-xTB2/MM	21.6	
	B3LYP-D3/6-31+G*	20.9	
B	DFTB3/MM	27.9	ACE + 1wat + Thr186 (C_α-C_β)
	GFN-xTB2/MM	/	
	B3LYP-D3/6-31+G*	/	
C	DFTB3/MM	21.1	ACE + 2wat
	GFN-xTB2/MM	/	
	B3YP-D3/6-31+G*	/	
D	DFTB3/MM	20.1	ACE + 1wat
	GFN-xTB2/MM	/	
	B3LYP-D3/6-31+G*	27.2	

Figure 4.11. Alternative reaction pathways for the hydrolysis of the acyl-enzyme investigated across various theoretical levels, along with the computed free energy barriers (kcal·mol^{-1}).

A scheme of the most favorable pathway is shown in Figure 4.12a, with the QM region and collective variables (CVs) in Figure 4.12b. The resulting free energy profile (Figure 4.12c) and the evolution of the CVs (Figure 4.12d). Free energy values for stationary points on the PMF (Figure 4.12d) and their corresponding CVs are listed in Table 4.5. An illustrative movie of the hydrolysis step of the mechanism is provided in the supplementary information material of the original paper or the following link (pubs.acs.org/doi/suppl/10.1021/acs.jcim.3c00900/suppl_file/ci3c00900_si_003.mp4).[181]

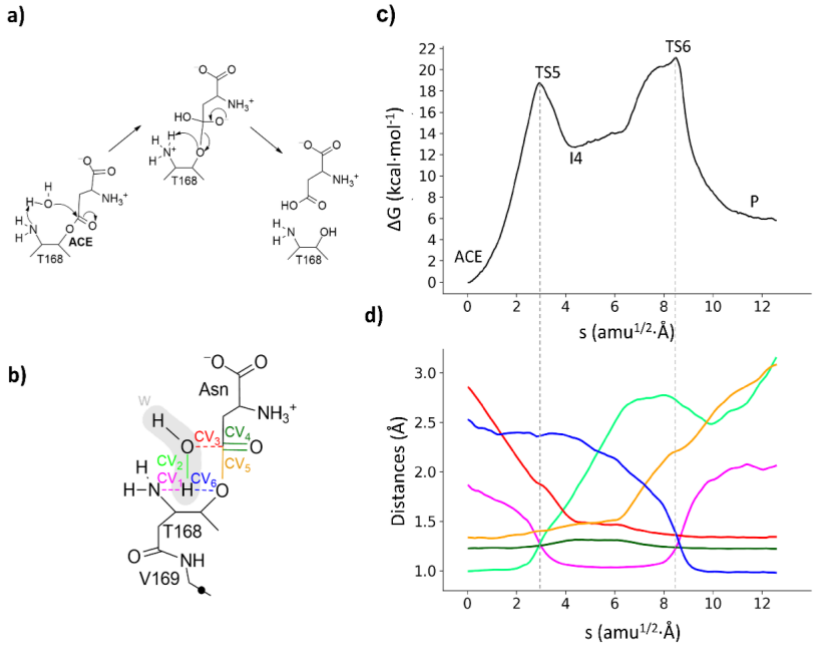


Figure 4.12. (a) Illustration of the mechanism of the acyl-enzyme hydrolysis in hASNase3. (b) Visualization of the quantum mechanical (QM) region and collective variables (CVs) employed to determine the minimum free energy path (MFEP), with the link atom depicted as a black dot. (c) Free energy landscape along the path collective variable (s), computed at the B3LYP-D3/6-31+G(d)/MM level of theory. (d) Evolution of the chosen CVs along the MFEP, with color codes matching panel (b). Dashed light-grey lines indicate transition state positions.

The deacylation stage begins when the Thr168 amino group deprotonates the hydrolytic water molecule. This initiates a concerted process where the water oxygen performs a nucleophilic attack on the Cγ of the atom of the substrate, as evidenced by changes in the distances (CV1, CV2, CV3 in Figure 4.12d). The associated transition state (TS5, Figure 4.13a) has a free energy of 18.6 ± 0.7 kcal·mol^{-1} relative to the acyl-enzyme complex and is stabilized by the oxyanion hole residues, as a negative charge accumulates on the substrate's carbonyl oxygen atom.

An illustrative movie of the hydrolysis step of the mechanism is provided in the supplementary information material of the original paper or the following link (pubs.acs.org/doi/suppl/10.1021/acs.jcim.3c00900/suppl_file/ci3c00900_si_003.mp4).[181] The deacylation stage begins when the Thr168 amino group deprotonates the hydrolytic water molecule. This initiates a concerted process where the water oxygen performs a nucleophilic attack on the Cγ of the atom of the substrate, as evidenced by changes in the distances (CV1, CV2, CV3 in Figure 4.12d). The associated transition state (TS5, Figure 4.13a) has a free energy of 18.6 ± 0.7 kcal·mol^{-1} relative to the acyl-enzyme complex and is stabilized by the oxyanion hole residues, as a negative charge accumulates on the substrate's carbonyl oxygen atom.

Table 4.5. Free Energies (kcal·mol^{-1}) of stationary structures associated with acyl-enzyme hydrolysis and key distance along the MFEP (Å).

	ACE	TS5	I4	TS6	P
Free energy	0.0	18.6	12.7	20.9	5.8
CV1 (Thr168Nδ-Hw)	1.87	1.26	1.05	1.19	2.02
CV2 (Hw-Ow)	0.99	1.27	1.81	2.74	2.90
CV3 (Ow-AsnCγ)	2.85	1.87	1.50	1.36	1.33
CV4 (AsnCγ-AsnOδ)	1.23	1.25	1.30	1.25	1.22
CV5 (AsnCγ-Thr168Oγ)	1.34	1.40	1.46	2.19	2.99
CV6 (Thr168Oγ-Hw)	2.53	2.36	2.33	1.41	1.00

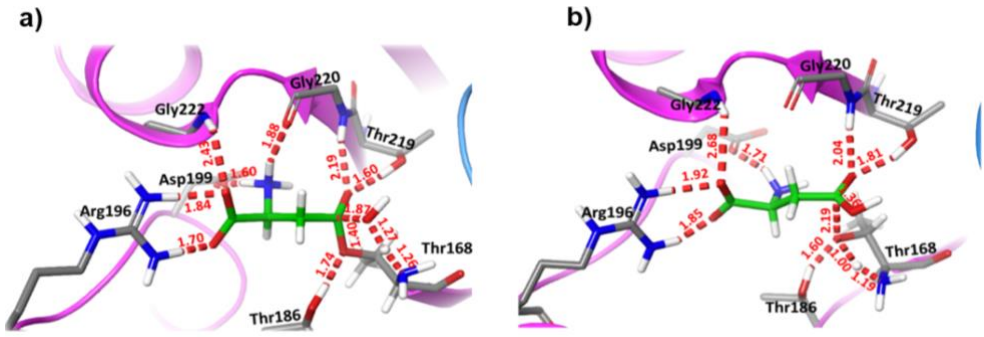


Figure 4.13. Structural representations of transition states involved in the acyl-enzyme complex hydrolysis in hASNase3. Panels: a) TS5 and b) TS6, with all distances provided in Å.

Subsequently, a tetrahedral intermediate (I4) is formed, leading to the rate-determining step: a proton transfer from the amino group of Thr168 to the acyl oxygen atom and the breaking of the acyl-enzyme bond. This step, involving TS6 with a barrier of 20.9 ± 1.2 kcal·mol^{-1}, is stabilized by the oxyanion hole residues Thr219 and Gly220, as well as the hydrogen bond between Thr186 and the acyl oxygen atom (Figure 4.13b). The reaction yields a protonated aspartic acid as the product.[181]

Finally, an additional mechanistic step was explored for the restauration of the initial protonation state of the enzyme. A proton is transferred from the carboxylic group of the aspartic acid to the amino group of the N-nucleophile. The reaction mechanism, QM part and CVs, as well as the obtained free energy profile and the CVs evolution are given in Figure 4.14, while the values of the most relevant distances are given in the Table 4.6. This regeneration step has a quite low barrier of 5.2 ± 0.8 kcal·mol^{-1}. The structure of the transition state (TS7) is given in Figure 4.15. After reaching TS7, the free energy decreases, resulting in an exergonic reaction with a free energy change of -7.2 ± 0.6 kcal·mol^{-1} relative to the Michaelis Complex (see Figure 4.14c).

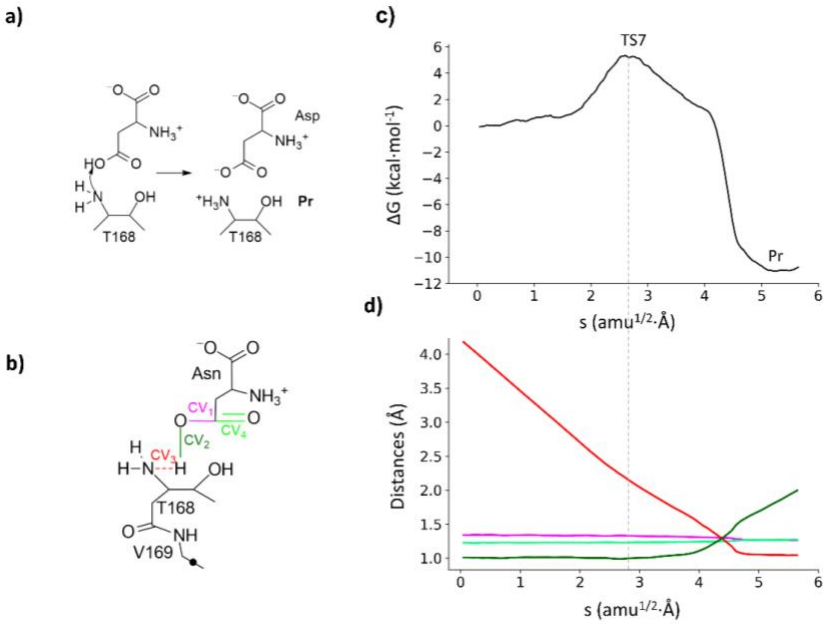


Figure 4.14. (a) Illustration of the mechanism of the regeneration of the hASNase3. (b) Visualization of the quantum mechanical (QM) region and collective variables (CVs) employed to determine the minimum free energy path (MFEP), with the link atom depicted as a black dot. (c) Free energy landscape along the path collective variable (s), computed at the B3LYP-D3/6-31+G(d)/MM level of theory. (d) Evolution of the chosen CVs along the MFEP, with color codes matching panel (b). Dashed light-grey lines indicate transition state positions.

Table 4.6. Free Energies (kcal·mol^{-1}) of stationary structures associated with regeneration of the hASNase3 and key distance along the MFEP (Å).

	R	TS7	Pr
Free energy	0.0	5.2	-11.2
CV1 (AsnCγ-Ow)	1.34	1.33	1.27
CV2 (Hw-Ow)	0.99	1.02	1.90
CV3 (Hw-Thr168Nδ)	4.16	2.28	1.02
CV4 (AsnCγ-AsnOδ)	1.25	1.27	1.27

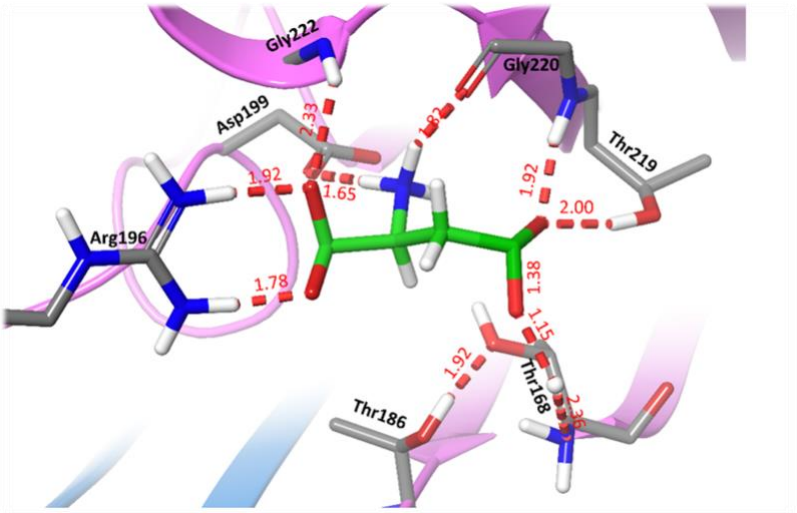


Figure 4.15. Structural representations of transition state (TS7) involved in the regeneration of the hASNase3. All distances provided in Å.

4.1.2.4 Full catalytic cycle of the hASNase3

The full catalytic cycle of hASNase3 and its B3LYPD3/6-31+G(d)/MM free energy profile are summarized in Figure 4.16.

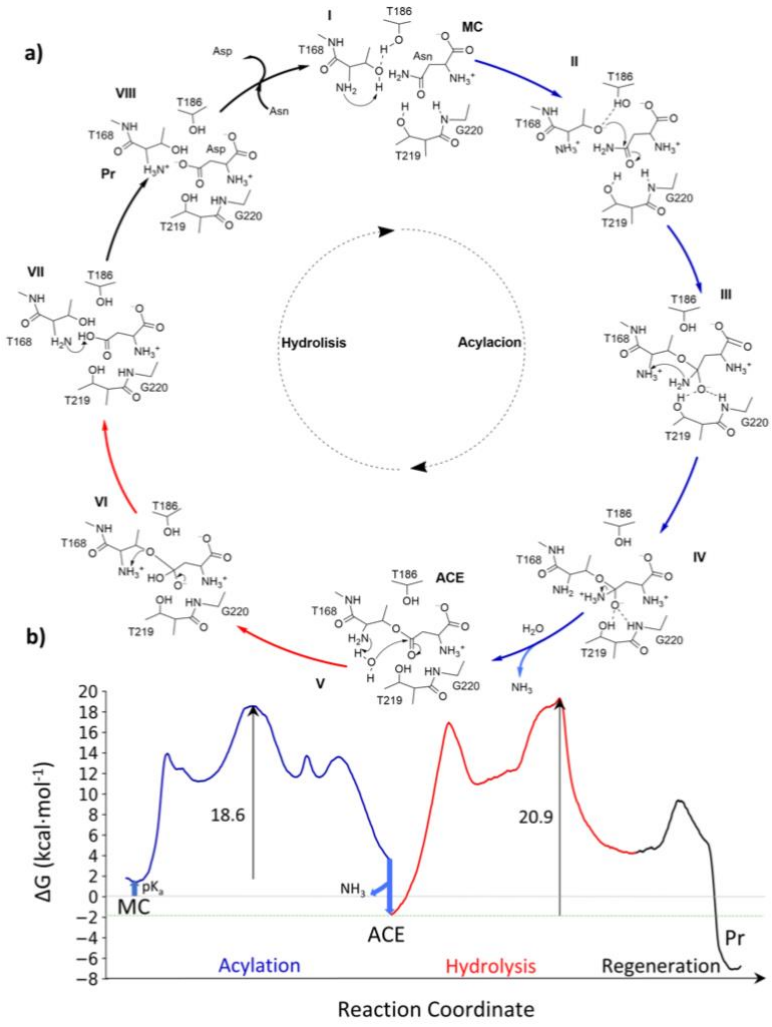


Figure 4.16. Computationally derived hydrolysis mechanism of asparagine catalyzed by hASNase3. (a) Diagram illustrating the enzymatic cycle. (b) Free energy landscape of the reaction, with the acylation phase shown in blue, deacylation in red, and enzyme regeneration in black.

This final profile combines the three individual free energy pathways, accounting also for the deprotonation of the nucleophile and ammonia release. The rate-limiting step, involving the cleavage of the acyl-enzyme bond, has a calculated free energy barrier of 20.9 kcal·mol^{-1}, aligning well with the experimental value

of 17.5 kcal·mol^{-1}, in particular considering that several steps contributed to this value.[180] Overall, the cycle is exergonic, with a net free energy change of -7.2 kcal·mol^{-1}. Additional support of the proposed mechanism is also provided by the structural validation. Relatively low RMSD of 0.97 Å is measured between the covalent acyl-enzyme complex and the X-ray structure (PDB code: 4O0H [180]) (see Figure 4.17). The obtained free energy profile also explains the possibility to crystallize the covalent intermediate. The release of ammonia renders acyl-enzyme formation irreversible while deacylation is the rate limiting step, leading to the accumulation of a stable intermediate.

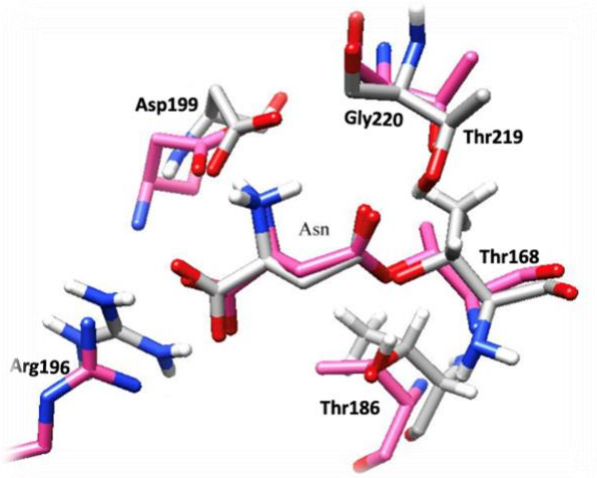


Figure 4.17. Comparison of active site structures: X-ray structure (PDB code: 4O0H) of the covalent acyl-enzyme (pink) versus product structure from ASM simulations (grey).

4.1.3 Short summary of hASNase3 results

Active oligomeric state: The results of classical MD simulations identified key interactions between the substrate and active site residues in the Michaelis complex. The simulations demonstrate that the dimeric form of hASNase3 is essential for catalytic activity. While the active site residues originate from a single protomer, inter-protomer interactions are critical for maintaining substrate positioning and stabilizing the active site. The sodium-binding loop was found to be stable during the simulation and the analysis revealed its crucial role in

stabilizing the C-terminal (Gly167) residue after cleavage, which, in turn, supports the N-terminal nucleophile during the catalytic reaction.

Protonation state of the N-terminal residue: Results from the Michaelis complex molecular dynamics identified Thr168 as the potential nucleophile. For its hydroxyl group to be activated, its own amino group has been proposed to subtract a proton. At pH 7.5, the N-terminal group of Thr168 is predominantly protonated, but with a pK_a value in the enzyme lower than in solution mainly due to a reduced solvation shell. The determined energy cost for deprotonation is small (1.38 kcal·mol^{-1}), suggesting that the amino group could indeed act as the base in charge of activating the hydroxyl group of the very same residue.

Reaction mechanism: The reaction mechanism was explored using multiscale QM/MM simulations and the Adaptive String Method. The most promising proposals were recalculated at the B3LYP-D3/6-31+G(d)/MM level of theory. The full reaction cycle consists of 3 main stages: acyl-enzyme formation, hydrolysis and hASNase3 regeneration, with seven transition states involved along the catalytic cycle. The first stage starts with a deprotonated N-terminal group in Thr168 abstracting a proton from the hydroxyl group of the same residue. The reaction then proceeds with the nucleophilic attack of the Oγ atom of Thr168to the Cγ of the substrate with the subsequent release of the ammonia. The proposed mechanism for the formation of the acyl-enzyme complex also explains its potential for crystallization, as this structure appears as a stable intermediate in the reaction free energy profile. Additionally, the acyl-enzyme complex structure obtained computationally aligns well with the X-ray data. In the second stage, a water molecule gets activated by the amino group of the N-terminal nucleophile to attack the Cγ of the substrate, reaching in this way the rate-determining step of the overall process, with the free energy barrier is 20.9 kcal·mol^{-1}. Finally, the protonation state of the enzyme is regenerated by means of a proton transfer from the aspartic acid to the amino group of Thr168, followed by a drop in the free energy, resulting in an exergonic process with a reaction free energy of -7.2 kcal·mol^{-1}. This proposal also explains the lack of reactivity in the mutants Thr219Val, Thr219Ala, and Thr186Val. Thr186 was found to stabilize the negative charge on the activated nucleophile, while Thr219 serves as an

oxyanion hole, stabilizing the negative charge buildup on the oxygen atom of the substrate.

4.1.4 Technical Details

4.1.4.1 System Preparation

The Michaelis complex structure was prepared using the PDB structure 4O0H, corresponding to the homodimeric form of the enzyme (each monomer consisting of 309 amino acid residues) in the acyl-enzyme state.[180] Since the loops comprising residues 153−167 were unsolved, their structure was modeled with AlphaFold2.[153] In order to simulate the cleaved active protein state, needed for the ASNase activity, the bond between residues 167 and 168 (as well as the corresponding bond in the other monomer) was broken. Given that the starting PDB structure corresponds to the acyl-enzyme state, the covalent bond between the enzyme and the substrate was also broken to obtain the Michaelis complex. The conformation of the substrate within the Michaelis complex was chosen optimizing its interactions with the active site residues observed in the X-ray structure. All water molecules and sodium ions present in the X-ray structure were kept in the model. Protonation states of titratable residues at pH 7.5 were determined using PROPKA3.0.[184] The N-terminal group of Thr168 was initially modeled as protonated, with its pK_a evaluated as described below.

4.1.4.2 Molecular Dynamics Simulations

All classical MD simulations were conducted using the GPU-optimized Amber18 pmemd software.[185, 186] Parameters for the substrate, free asparagine (Asn) in its zwitterionic form, were adopted from Horn et al.,[187] while standard amino acids were modeled with the ff14SB force field.[140] The system was solvated in a TIP3P water box,[141] ensuring protein-inhibitor atoms were at least 12 Å from the simulation box edges, prepared using AmberTools18 tleap.[188] To neutralize the total charge of the system Na^+ were added.

Minimization was performed in multiple cycles, each comprising 50,000 steps. The first 10,000 steps used the steepest descent (SD) method, followed by the

conjugate gradient (CG) method, until the root-mean-square gradient was reduced to approximately 10^{-4} kcal·mol^{-1}Å^{-1}. Minimized structures were then gradually heated to 310 K using Langevin dynamics with a collision frequency of 5.0 ps^{-1} and a linear heating ramp that increased the temperature from 0 to 310 K. During the heating process, periodic boundary conditions with isotropic position scaling were applied and a time step of 1 fs was used.

The equilibration stage was done in the NPT ensemble during which a mild parabolic restraint potential (20 kcal·mol^{-1}Å^{-1}) was applied to the protein backbone atoms, and, for systems containing a substrate, to selected substrate-active site residues distances. Once heated, the restraints were removed at a rate of approximately 1 kcal·mol^{-1}Å^{-1} ns^{-1}. The time step was increased to 2 fs after applying the SHAKE algorithm to constrain bonds involving hydrogen atoms.[189]

Once all restraints were removed, the equilibrated structures entered the production phase, consisting of 1 μs simulations in the NVT ensemble using the Amber18 GPU version of pmemd. The simulation parameters remained consistent with those used during equilibration. To enhance sampling, three independent 1 μs replicas were performed for each system. Electrostatic interactions were calculated via particle-mesh Ewald[190], with a 10 Å cutoff for non-electrostatic interactions.

4.1.4.3 Thermodynamics Integration

pK_a calculations

The thermodynamic Integration (TI) method was used to estimate the free energy change associated to the deprotonation of the N-terminal group of Thr168. Briefly, TI was performed by numerically integrating the derivative of free energy with respect to a coupling parameter (λ), which drives the alchemical transformation of the N-terminal group from unprotonated to protonated. Free energy calculations were carried out in Amber18 using the dual topology approach, following a published protocol for GPU-based simulations.[191] The

temperature was maintained at 310 K using Langevin dynamics with a collision frequency of 5.0 ps^{-1}. The pressure was held constant at 1.01325 bar using a Monte Carlo barostat with a pressure relaxation time of 2.0 ps. The disappearing atom during the alchemical transformation (the proton) was placed in the soft-core region for both van der Waals and electrostatic interactions. All simulations were performed with pmemd.cuda on a GPU [186], with a time step of 1 fs. As previously described in the Methods section 3.2.2, the transformation from the initial to the final step was done in several steps. To ensure reliable averaging, alchemical transformations were conducted using 9 different λ values (λ = 0.01592, 0.08198, 0.19331, 0.33787, 0.5, 0.66213, 0.80669, 0.91802, 0.98408), with five separate replicas run of 5 ns at each λ value. The final values were integrated using Gaussian quadrature method and a custom python script. The first 1 ns of each simulation was considered equilibration for the corresponding λ window and excluded from the post-analysis of $dU/d\lambda$. The free energy changes from five replicas were averaged arithmetically. The average free energy change for the alchemical transformation in water was subtracted from the same calculation in the protein environment, and this difference ($\Delta\Delta G$) was used in equation (4.1). The error was calculated as the standard error of the mean by dividing the standard deviation by the square root of the number of replicas.

Free energy calculation of the NH_3 leaving the active site

The free energy change associated with NH_3 leaving the enzymatic active site ($-\Delta G_{bind}$) can be calculated by combining the different terms present in the thermodynamic cycle shown in Figure 4.18:

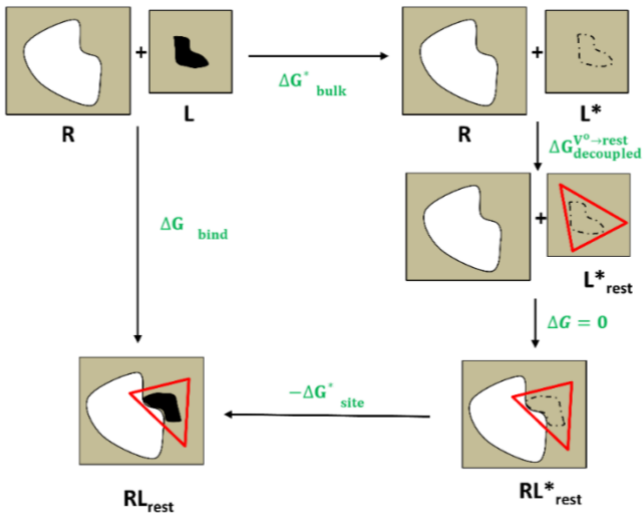


Figure 4.18. Thermodynamic cycle for calculating the free energy change of NH_3 exiting the active site. R represents the unbound protein, L denotes the unbound NH_3 in water, L* refers to the decoupled NH_3, and RL is the protein-NH_3 complex. Adapted from [89].

$$-\Delta G_{bind} = \Delta G^*_{bulk} + \Delta G^{V^o \rightarrow rest}_{decoupled} - \Delta G^*_{site} \tag{4.2}$$

The first (ΔG^*_{bulk}) and last (ΔG^*_{site}) terms represent the free energies for NH_3 disappearing in bulk water and in the active site, respectively. These magnitudes were evaluated using TI in five independent replicas. For this purpose, a softcore potential was applied to NH_3, enabling SHAKE constraints, while Hamiltonian replica exchange ensured better sampling during 20 ns simulations per window. The alchemical transformation was performed in 12 steps. Simulations were run in the NVT ensemble to maintain constant volume for effective replica exchange. To retain NH_3 in the active site during alchemical transformations, a restraining potential was applied to the distance between the nitrogen of NH_3 and the substrate Cγ. The free energy associated to restraining the decoupled NH_3 ($\Delta G^{V^o \rightarrow rest}_{decoupled}$) was calculated as follows:

$$\Delta G_{decoupled}^{V^o \rightarrow rest} = -RTln\left(\frac{Q}{V^o}\right) \quad (4.3)$$

where Q is given as:

$$Q = \int_0^\infty 4\pi r^2 \, e^{-\beta U_{rest}(r)} dr \quad (4.4)$$

The force constant of the applied restraining potential of the form $U_{rest} = k(r - r_0)^2$ was 50 kcal·mol^{-1}·Å^{-2}, where *r* represents the AsnCγ-AsnNδ distance. The cutoff of the restraining potential (r_0) was set to 3.5 Å, matching the distance used in the final node of the string calculation.

4.1.4.4 QM/MM Simulations and Adaptive String Method

Hybrid quantum mechanics/molecular mechanics (QM/MM) simulations were employed to investigate the free energy profiles associated with the chemical process. Initial exploratory calculations of various mechanistic proposals were conducted at the DFTB3/MM level.[123] The choice of the QM region choices depends on the mechanism tested and are given in the corresponding sections. The best mechanisms, in terms of activation free energies, were recalculated with the QM subsystem described using the B3LYP [114, 115] functional with D3 [192] dispersion corrections (B3LYP-D3) and the 6-31+G(d) basis set. The selection of the B3LYP functional was based on previous studies of ASNases type 1 and 2,[32, 40, 76, 77] where it provided energies consistent with experimental data.

ASM, developed by our research group, was used to explore the free energy profile of the reaction mechanism.[147, 193] Calculations were carried out using Amber24 coupled with Gaussian16 [194] for density functional theory computations. As described in the Methods section of this thesis, in ASM, a series of replicas (string nodes) are defined along a hypothesized reaction pathway that connects the reactant and product states, represented within a space of collective variables (CVs). Node positions are evolved along the free energy gradient, while maintaining equidistance to ensure convergence to the Minimum Free Energy

Path (MFEP). To improve convergence, Hamiltonian exchange between neighboring nodes is attempted every 50 steps. The CVs used to define the mechanism included bond distances associated with bond-breaking and bond-forming events and are described for each reaction above. Convergence of the string was assessed using the root-mean-square deviation (RMSD) of the CVs, with a convergence criterion of RMSD ~0.1 $amu^{1/2} \cdot Å$ maintained for at least 2 ps. After convergence, a path coordinate (denoted as *s*) was defined to measure the system's position along the MFEP. This coordinate was subsequently used as the reaction coordinate in 10 ps US calculations. [90] The free energy profile was estimated using the WHAM.[195] Force constants for US calculations were determined dynamically to ensure a uniform probability density along the reaction coordinate. Simulations were run in the NVT ensemble at a temperature of 310 K, employing a Langevin thermostat with a collision frequency of 2.0 ps^{-1}. A time step of 1 fs was used, with a cutoff of 15 Å for DFT/MM interactions. During ASM simulations, transferred hydrogen atoms were assigned an increased mass of 2 a.m.u. to enhance the proton transfer sampling.

4.2 Guinea Pig and Human Asparaginase type 1 (gpASNase1 and hASNase1)

Guinea pig ASNase type 1 (gpASNase1) and human ASNase type 1 (hASNase1) enzymes belong to the class 1 ASNases. Both enzymes are homotetramers formed from two dimers of intimate dimers and spanning four active sites. In this section, residue names accompanied by a single prime notation (') indicate residues originating from the other protomer of the same intimate dimer. A double prime (") indicates residues from the adjacent (non-intimate dimer protomer), and a triple prime ('") designates residues from the protomer opposite to the active site (the furthest away from the active site into consideration).

These enzymes share high sequence similarity (~85%) and are therefore thought also to share same catalytic mechanism.[28, 29] In addition, these enzymes are thought to go through similar rearrangements upon substrate binding. Specifically, the loop from the adjacent protomer of the intimate dimer undergoes a conformational change after substrate binding, acting as a lid that sterically closes the active site.[28] Additionally, there have also been proposals that suggest that this conformational change also play a crucial role in catalysis.[32] However, to our knowledge, theoretical investigations of the reaction mechanism of gpASNas1 have been limited [32, 76, 77] and the dynamics of the flexible loop and its implication on catalysis has not been explored. In addition, hASNase1 has not yet been crystallized, which can partly explain the lack of the theoretical studies on this system.[79]

Apart from these minor changes, hASNase1 is also thought to suffer larger scale conformational changes, closing up its quaternary structure upon the substrate binding.[28] These large conformational rearrangements in hASNase1 has been speculated to be the reason for its weaker binding affinity for the substrate compared to gpASNase1, which does not undergo similar large scale rearrangements.[28] Interestingly, despite the high structural similarity between the two enzymes, hASNase1 presents allosteric regulation by the substrate, while this behavior has been detected in gpASNase1. Additionally, gpASNase1 has

been found to significantly lack glutaminase activity, an important feature for ALL treatment.[28]

Given all previously mentioned favorable properties and its mammalian origin, gpASNase1 has garnered interest as a therapeutic enzyme. Therefore, recent efforts using directed evolution and DNA shuffling with hASNase1 have produced humanized variants with promising kinetic profiles that have recently been patented.[29] However, the structural and mechanistic basis for the properties of these chimeras remains unexplained.

This section discusses the results presented in a preprint by Andjelkovic et al.[196] It begins with a detailed examination of active site interactions in gpASNase1, followed by an exploration of binding selectivity, highlighting the molecular origins for its preference toward asparagine over glutamine. Next, conformational changes of the flexible loop are discussed. We then discuss the reaction mechanism, identifying key residues and structural motifs that contribute to the stabilization of the transition states. We use this analysis to rationalize the properties of humanized chimeras and to assess the potential implications for therapeutic applications. The section concludes with a summary, followed by the Technical Details, which outline the methodologies and computational approaches used in the present section.

4.2.1 Active site interactions in the gpASNase1

To investigate the properties of gpASNase1, MD simulations were performed on both the apo enzyme and the Michaelis complex with asparagine. A previous study of the gpASNase1 simulated one of the dimers, adding restraints to keep a catalytic conformation in the active site.[32] However, in this study, simulations were carried out for the full tetramer, revealing that, in the absence of any external restraints, both the enzyme and substrate remained stable across all four active sites throughout 1 μs simulation and three independent replicas, as demonstrated by RMSD time profiles (see Figure 4.19) and visual inspection.

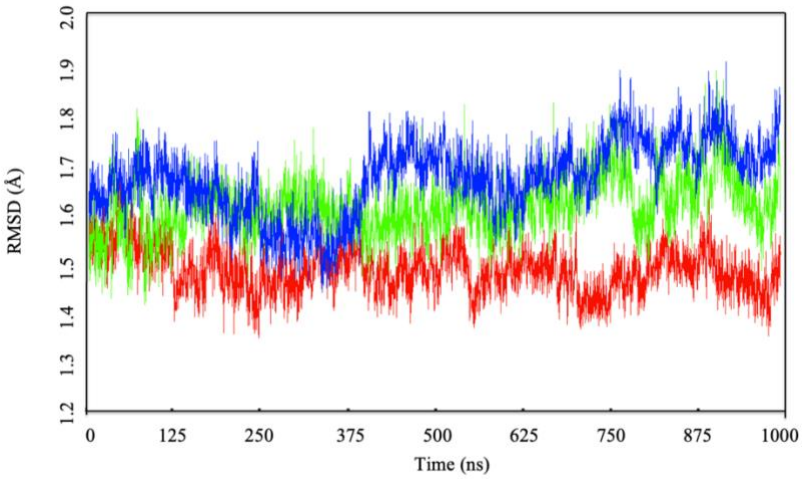


Figure 4.19. Root Mean Square Deviation (RMSD) during the MD simulations of the Michaelis complex in the dimeric form of hASNase3 with substrate bound in the active site.

One of the ongoing debates in the research of ASNase types 1 and 2 revolves around the protonation state of a lysine residue located in the active site (K188 in gpASNase1, see Figure 4.20). Certain mechanistic models propose that Lys188 (or its equivalent Lys in other ASNases) remains deprotonated, allowing it to serve as a general base that activates the nucleophile.[77] Alternatively, other proposals suggest that this lysine remains mostly protonated and plays a role in proton transfer to the leaving amino group.[76] For this reason, free energy calculations were performed to unveil the probability of these two protonation states. The free energy values were obtained as averages of 5 replicas each following the alchemical transformation performed in the holo protein environment and in aqueous solution (similar to the one given in Figure 4.7). Further details are given in the Technical Details section of this chapter (see section 4.1.4).

These results indicate that deprotonation of Lys188 within the Michaelis complex of gpASNase1 involves a significant free energy penalty, 11.8 ± 0.1 kcal·mol^{-1} at pH 7.5 and 310 K (see Table **4.7**). As a result, we modeled Lys188 as protonated in our simulations of the Michaelis complex.

Table 4.7. Free energies of the alchemical transformation of the protonation states of the lysine residue in holo form of gpASNase1 relative to water and corresponding pK_a shift and pK_a value. Free energy costs of threonine deprotonation at pH 7.5 and T = 310 K and the calculated probabilities of its protonated and deprotonated states are given under the same conditions.

Form	gpASNase1
$\Delta\Delta G$ (kcal·mol^{-1})	7.6 ± 1.0
$pK_{prot} - pK_{aq}$	5.3 ± 1.0
pK_{prot}	15.8 ± 1.0
$\Delta G(deprotonation)_{pH=7.5,T=310\ K}$ (kcal·mol^{-1})	11.8 ± 0.1
$P(protonated)_{pH=7.5,T=310\ K}$	0.99
$P(deprotonated)_{pH=7.5,T=310\ K}$	0.01

Throughout the simulations, the substrate maintained its position within the active site with its binding orientation closely matching that observed in the X-ray crystal structures of the product and Michaelis complexes (PDB codes: 4R8L [28] and 5DNC [33]). The interactions between the enzyme and substrate and the probability distributions for key distances are provided in Figure 4.20.

As represented in Figure 4.20, positively charged amino group of the substrate establishes strong interactions with the carboxylate side chains of Asp84 and Asp117. On the other side, the negatively charged α-carboxylate group of the substrate engages in hydrogen bonding with the backbone amine group of Ser85, its side chain hydroxyl group, and the backbone amine group of Asp117. Furthermore, the carbonyl oxygen atom of the substrate forms hydrogen bonds with the backbone amine groups of Thr116 and Thr19, contributing to the formation of an oxyanion hole that stabilizes the negative charge developed on the substrate's carbonyl oxygen atom during the reaction. These interactions collectively orient the NH_2 group of the substrate in a way that promotes the departure of the leaving group (NH_3) during the reaction. Notably, the carbonyl carbon (Cγ) of the substrate is near the hydroxyl oxygen of Thr19 (Oγ), with an average distance of 3.46 ± 0.44 Å, suggesting that Thr19Oγ may act as the

nucleophile in the subsequent attack on the carbonyl carbon atom to form the ACE complex.

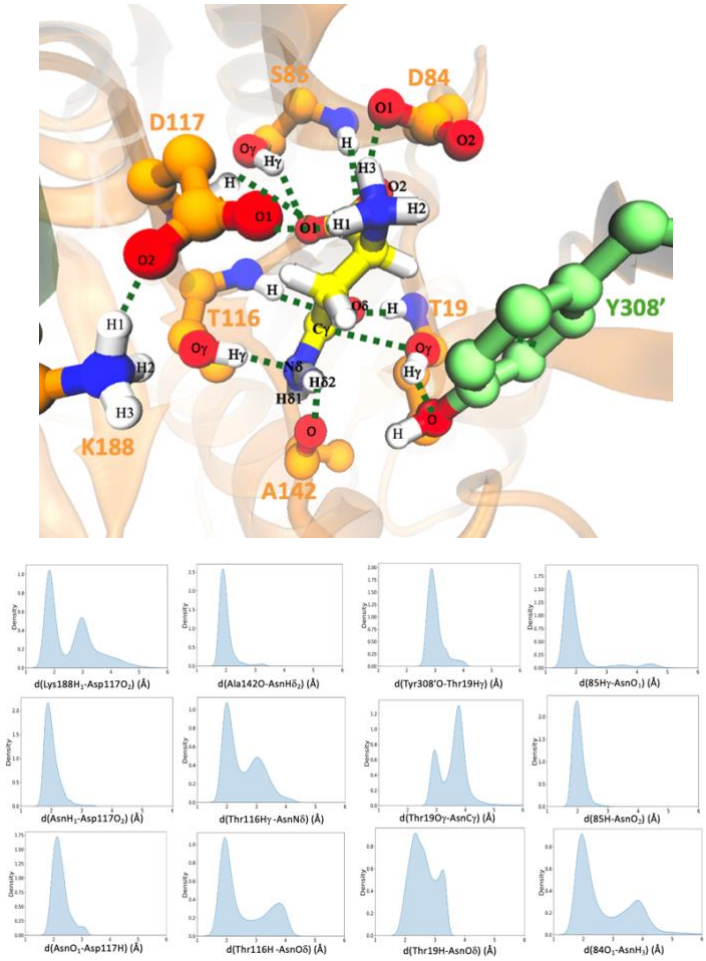


Figure 4.20. Representation of the active site of gpASNase1 with the Asn substrate (yellow balls and stick representation) from MD simulations. Distributions of the important distances in (Å) obtained over three replicas of 1 μs of classical molecular dynamic simulation run on the Michaelis complex.

4.2.2 Binding selectivity of gpASNase1

As mention in the introduction section of this doctoral thesis (see section 1.3.1), some recent research, highlighted the lack of glutaminase activity of gpASNase1,

attributed to the poor binding affinities for this residue.[28] To investigate why gpASNase1 selectively binds Asn over Gln, MD simulations of the Michaelis complex with Gln were also run. Using alchemical transformation and thermodynamic integration (TI) (details given in the Technical Details section), the relative binding free energy of gpASNase1 with Asn was calculated to be 6.3 ± 1.3 kcal·mol^{-1} more favorable than with Gln, consistent with the experimental findings that gpASNase1 lacks glutaminase activity.

To identify the residues responsible for the distinguishing Gln and Asn, substrates' interaction energies (E_{int}), sum of the electrostatic and van der Waals components, were computed. The calculations were done using the frames extracted from the 1 μs MD simulations. Further details are given in the Technical Details section. Results are represented in Figure 4.21.

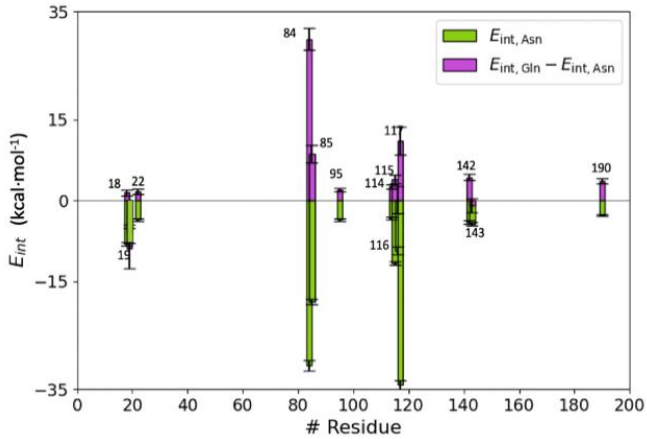


Figure 4.21. Average interaction energy contributions of individual residues to Asn binding ($E_{int,Asn}$, represented by green bars) and the difference in interaction energy between Gln and Asn ($E_{int,Gln} - E_{int,Asn}$, shown as purple bars). Only residues with $|E_{int,Asn}|$ greater than 2 kcal·mol^{-1} are included for clarity. Energy values are in kcal·mol^{-1}, and error bars indicate standard deviations of the mean.

As depicted in Figure 4.21, several key residues were identified as responsible for the binding affinity of the enzyme by Asn. The most prominent contributions are those of the residues that interact with the charged carboxylate and amino groups of the zwitterionic substrate, including Asp84, Ser85, Thr116, and

Asp117. Additionally, significant contributions are reached by Asp190 and Ala142 to substrate binding. Ala142 forms a critical hydrogen bond with Asn, positioning it for the nucleophilic attack by Thr19. Furthermore, residues Thr19 and Thr116 contribute to the oxyanion hole that stabilizes the substrate. Most residues involved in Asn binding exhibit stronger interactions with Asn than with Gln, with the most pronounced differences observed for Asp84, Ser85, and Thr117 (refer to the purple bars in Figure 4.21). To accommodate the larger Gln substrate, the side chain of Ser84 must be reoriented pointing away from the active site (see Figure 4.22a). Notably, Thr19 shows a stronger interaction with Gln than with Asn, as indicated by the negative purple histogram bar in Figure 4.21.

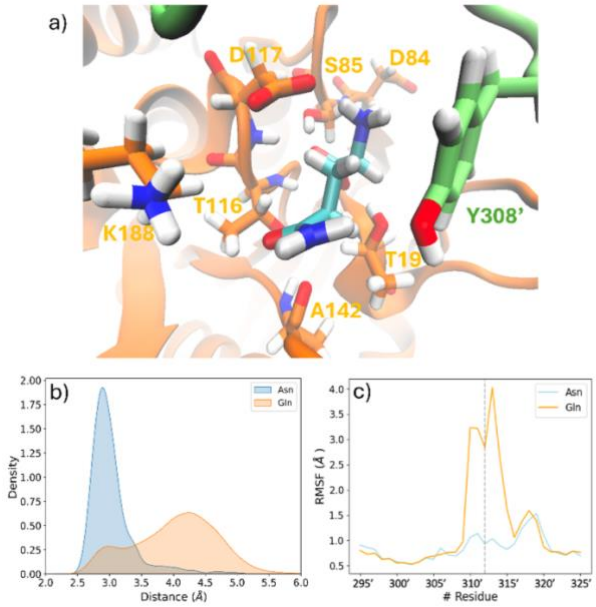


Figure 4.22. a) Visualization of the binding pose of Gln within the gpASNase1 active site, highlighting key residues with significantly altered interaction energy contributions compared to Asn; b) Distance probability distributions between Thr19 and Tyr308' during 1 μs simulations with Asn (blue) and Gln (orange) bound in the active site; c) Root-mean-square fluctuations (RMSF) of Cα atoms in the Tyr-loop residues for simulations with Asn (blue) and Gln (orange) present in the active site.

Regarding the binding pose of Gln in the active site, the hydroxyl groups of Thr19 and Thr116 are positioned toward the carboxylate group of this residue, as shown

in Figure 4.22a. This configuration indicates that Gln adopts a non-reactive binding pose. First, the hydroxyl group of Thr19 does not form a hydrogen bond with Tyr308', an interaction critical for activating the nucleophilic attack observed with Asn (Figure 4.22b). The absence of the Thr19-Tyr308' interaction impacts not only the preorganization of the active site for the nucleophilic attack but also the stability of the Tyr-loop, the flexible loop containing Tyr308 and that closes the active site, which becomes more flexible with Gln as the substrate (Figure 4.22c). Second, the hydroxyl group of Thr116 is not aligned with the NH_2 leaving group of Gln (Figure 4.22a), a key interaction for its protonation.

4.2.3 Conformational change of the flexible loop

As mentioned in the introduction section of this thesis (see subchapter 1.2.2), the Tyr-loop, containing Tyr308', is believed to play a key role in activating Thr19 for the nucleophilic attack on the substrate in type 1 and 2 ASNases. Namely, this loop transitions between an open state, allowing substrate binding, and a closed state, reorganizing the active site for catalysis with the formation of a crucial Tyr308'-Thr19 interaction. To identify the structural features driving this conformational change, we analyzed classical MD simulations of the enzyme in both apo and holo forms for the open and closed states. Figure 4.23a illustrates these states alongside five collective variables (CVs) - two torsional angles and three distances - used in ASM simulations of the loop conformational change. In the closed state (Figure 4.23a), Tyr308' forms a hydrogen bond with Thr19, while in the open state, its side chain rotates to interact with the carbonyl group of Pro274. These two distances, along with the intraloop hydrogen bond between Ala309' and Ala313' observed in the open state, were selected to guide the conformational change in the simulations.

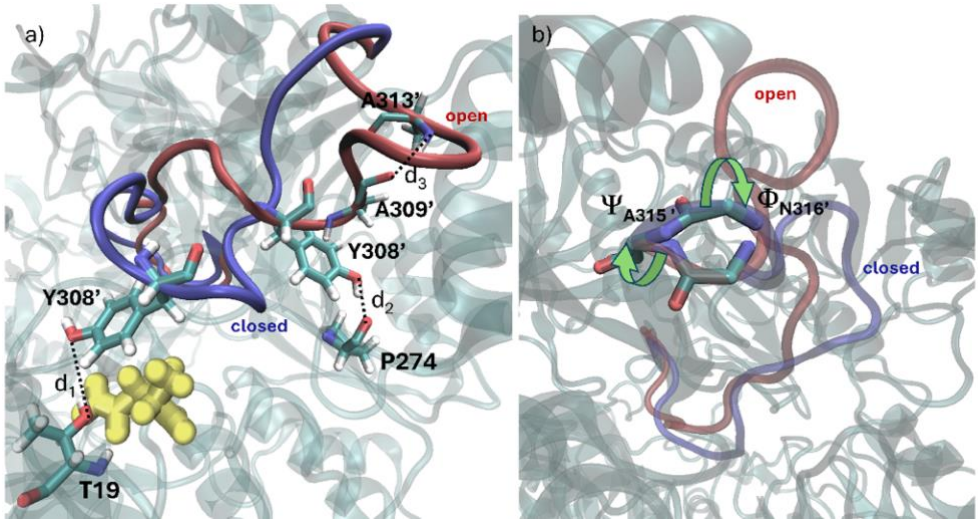


Figure 4.23. Open (red) and closed (blue) conformations of the Tyr-loop. a) Key residues (Thr19, Pro274, Tyr308', Ala309', and Ala313') define distances d_1, d_2, and d_3, which characterize loop positioning in the two states (d_1: distance between Tyr308'Oγ and Thr19Oγ; d_2: distance between Tyr308'Oγ and Pro274Oγ; d_3: distance between Ala309'O and Ala313'N, capturing the formation of a helix-like turn in the open state). The substrate is depicted as yellow ball-and-stick models. b) Backbone torsional angles around the Ala315'-Asn316' peptide bond that describe the loop conformational transition between the open and closed states.

In addition to these distances, backbone torsions also play a significant role in the transition between the open and closed conformations. Probability distributions for all Φ and Ψ dihedral angles of the loop backbone (residues 300–316) and the three distances defined in Figure 4.23 were analyzed from 1 μs MD simulations of the open and closed states in both the apo and holo forms of gpASNase1 (see Figure **4.24** and Figure 4.25).

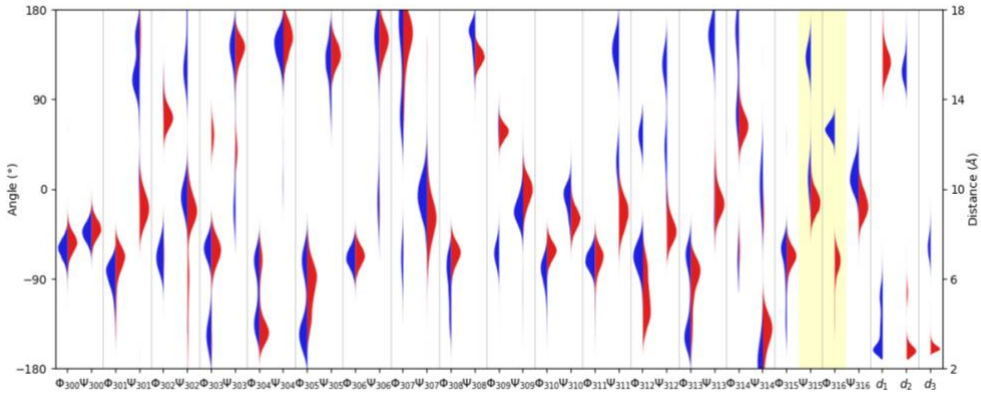


Figure 4.24. Violin plots showing the distributions of Φ and Ψ (in ◦) dihedral angles for the open (red) and closed loop states, along with the two distances derived from MD simulations of the apo form of gpASNase1. The distances d_1 (Tyr308'Og to Thr19Og) and d_2 (Tyr308'Og to Pro260Og) are included. The yellow-patched background highlights the distributions used to distinguish between the open and closed loop states.

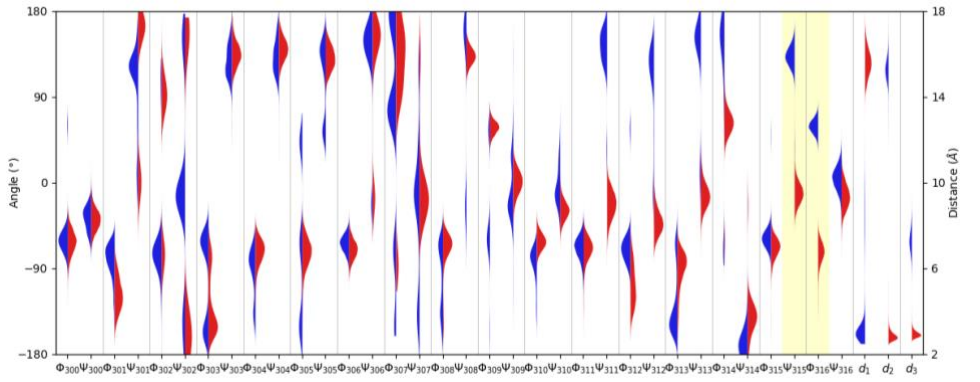


Figure 4.25. Violin plots showing the distributions of Φ and Ψ (in ◦) dihedral angles for the open (red) and closed loop states, along with the two distances derived from MD simulations of the holo form of gpASNase1. The distances d_1 (Tyr308'Og to Thr19Og) and d_2 (Tyr308'Og to Pro260Og) are included. The yellow-patched background highlights the distributions used to distinguish between the open and closed loop states.

As depicted in Figure **4.24** and Figure 4.25, the Ψ315' torsional angle, and to a lesser extent the Φ316' angle, exhibit distinct, non-overlapping distributions

between the two states in both forms, apo and holo.These two backbone torsions govern the rotation of the Ala315'-Asn316' peptide group, which display a different orientation with respect to the backbone in the open and closed states. A similar mechanism involving the rotation of a single peptide group controlling the open-to-closed transition was observed for the WPD-loop of the PTP1B enzyme.[197] Figure 4.26 displays the probability distributions of these two torsional angles. Given that these two torsions display the only distribution with no overlap between the open and closed states, they were chosen as the CVs which govern the conformational change, alongside the previously defined distances.

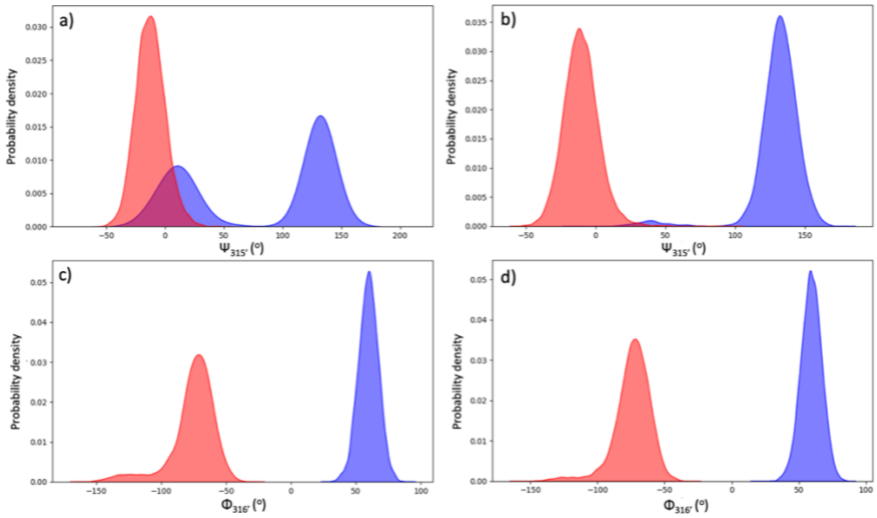


Figure 4.26. Distribution of the probability densities for the dihedral angles Ψ315' and Φ316 (in ◦) in the open (red) and closed loop configurations, derived from MD simulations of the apo form of gpASNase1 (panels a and c) and the holo form of gpASNase1 (panels b and d).

Using the selected set of five CVs, MFEPs were obtained using the ASM. The free energy profiles along these paths are shown in Figure 4.27.

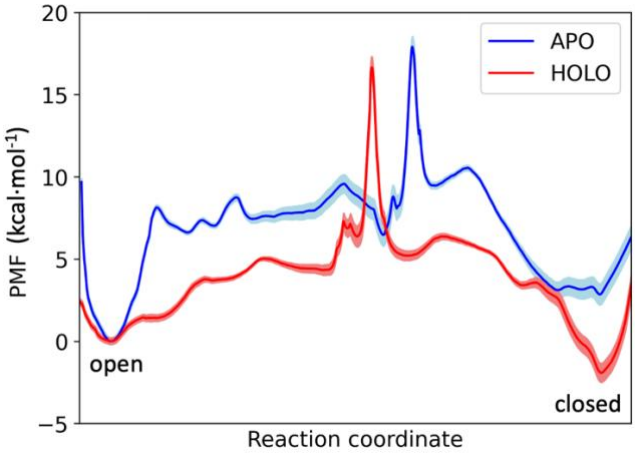


Figure 4.27. Free energy profiles of Tyr-loop transition from closed to open in the apo (blue) and holo (red) forms of gpASNase1. The shaded area represents the uncertainty in the data, calculated using bootstrapping.

Figure **4.28** and Figure 4.29 provide individual free energy profiles and the corresponding evolution of CVs, for the apo and holo forms, respectively. To further validate the adequacy of the chosen CVs in capturing the conformational transition of the loop, the structures obtained from the ASM simulations were compared with the X-ray structures. The root-mean-square deviation (RMSD) of the Cα atoms of the loop, relative to the X-ray data, were calculated at each string node for both the open and closed states, and is presented in Figure **4.28**b and Figure 4.29b. These results confirm that the ASM simulations effectively drive the loop from one conformation to the other. The analysis of the evolution of the CVs in Figure **4.28**c and Figure 4.29c reveal that the loop transition in both apo and holo states occurs in two distinct stages.

The first stage involves the breaking of the hydrogen bond between Tyr308' and Thr19 as the system moves from the closed to the open state. The second stage, marked by a steep change in the free energy profile, is primarily influenced by the rotations in the Ψ315' and Φ316' dihedral angles. The barrier is dominated by the rotation of the peptide group between residues 315'–316'. In the final open state, the stability is enhanced by the formation of a hydrogen bond between

Tyr308' and Pro274, as well as the creation of a small helical turn (indicated by the change in the distance between Ala309' and Ala313', d_3).

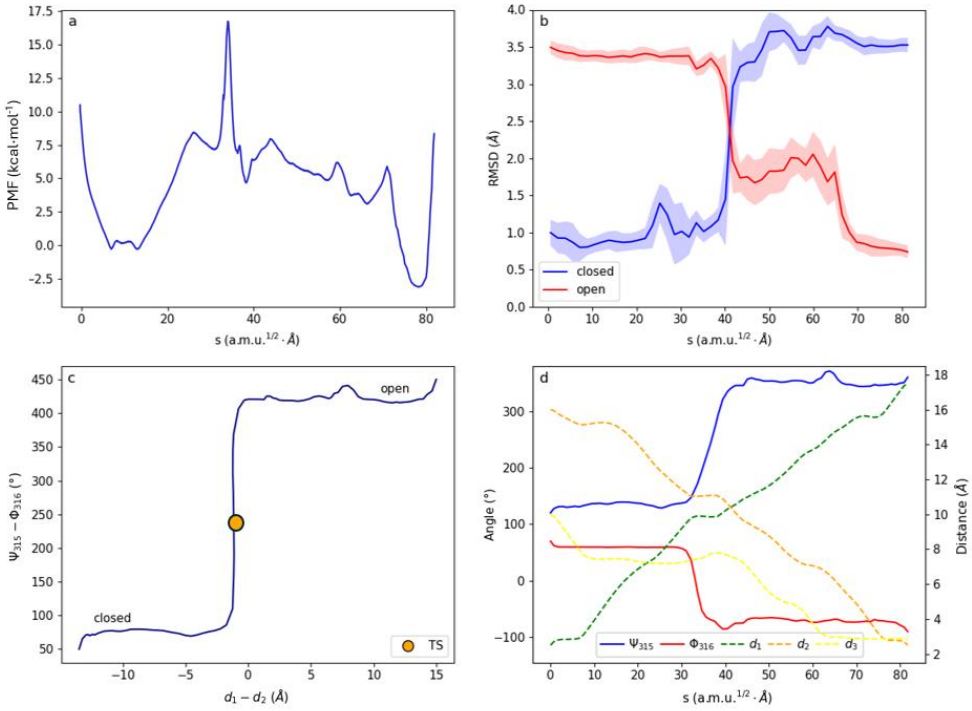


Figure 4.28. Free energy profile for the Tyr-loop transition from closed to open in apo gpASNase1. a) Free energy curve for the transition from closed (left) to open (right) states along the pathCV. b) RMSD of the Ca atoms in the Tyr-loop for umbrella sampling snapshots along the path-CV compared to the X-ray structures of the closed (blue) and open (red) states. The shaded area represents statistical uncertainty (95% confidence interval). c) Projection of the minimum free energy path (MFEP) along the combined distances and dihedral angles used as collective variables, with the yellow dot marking the Transition State. d) Evolution of the individual collective variables (distances on the right axis, dihedrals on the left) along the MFEP. The CVs include the Ψ315' and Φ316' dihedral angles, distances d_1 (Tyr308'Og-Thr19Og), d_2 (Tyr308'Og-Pro274Og), and d_3 (Ala309'C-Ala313'N).

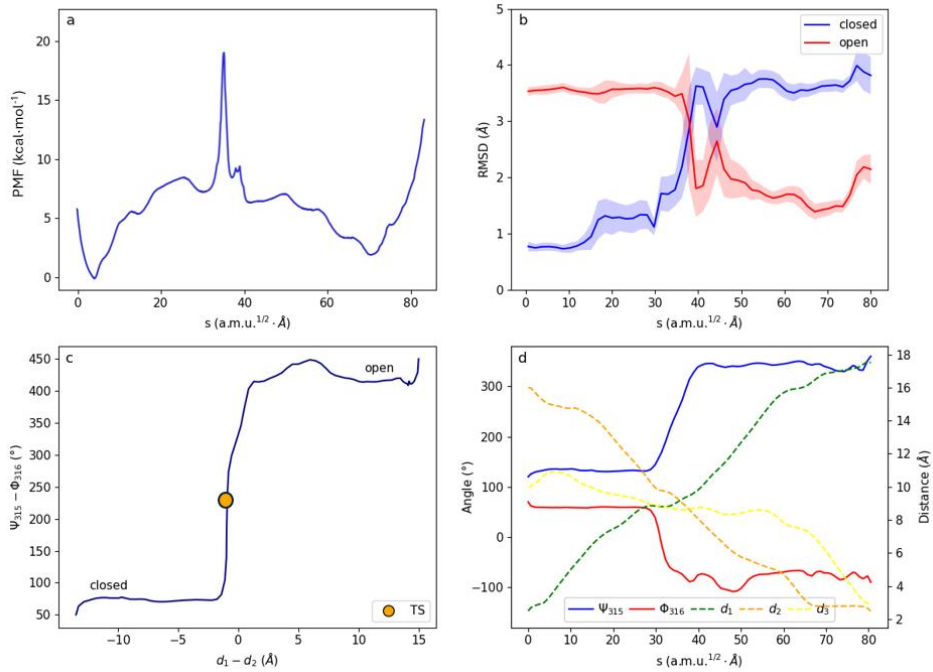


Figure 4.29. Free energy profile for the Tyr-loop transition from closed to open in holo gpASNase1. a) Free energy curve for the transition from closed (left) to open (right) states along the pathCV. b) RMSD of the Ca atoms in the Tyr-loop for umbrella sampling snapshots along the path-CV compared to the X-ray structures of the closed (blue) and open (red) states. The shaded area represents statistical uncertainty (95% confidence interval). c) Projection of the minimum free energy path (MFEP) along the combined distances and dihedral angles used as collective variables, with the yellow dot marking the Transition State. d) Evolution of the individual collective variables (distances on the right axis, dihedrals on the left) along the MFEP. The CVs include the Ψ315' and Φ316' dihedral angles, distances d_1 (Tyr308'Og-Thr19Og), d_2 (Tyr308'Og-Pro274Og), and d_3 (Ala309'C-Ala313'N).

This transition closely mirrors the one observed in the PTP1B system [197] and could be an indication of a general mechanism for loop transitions. As illustrated in Figure 4.27, both apo and holo forms exhibit a similar free energy barrier for Tyr-loop closure (17.9 kcal·mol^{-1} for the apo form and 16.7 kcal·mol^{-1} for the holo form). In the absence of the substrate, the open state is 2.8 kcal·mol^{-1} more stable than the closed state, whereas in the holo form, the closed state is favored by 1.9 kcal·mol^{-1}. This aligns with the X-ray observations, where the apo form

predominantly adopts the open conformation, while the holo form is seen in the closed state.[28]

To identify the residues that stabilize the open and closed states in the holo enzyme, we calculated the interaction energies between the loop and the remaining residues of the protein, considering the loop as the ligand. Table **4.8** lists the interaction energies for residues with values exceeding 15 kcal·mol^{-1}.

Table 4.8. Interaction energies (in kcal·mol^{-1}) of residues with the Tyr-loop in both open and closed states. Only residues with contributions less than -15 kcal·mol^{-1} are shown. Colored residues indicate those involved in stabilizing the open (blue) and closed (red) loop states, as depicted in Figure 4.30.

Residue	Closed loop	Open loop	Residue	Closed loop	Open loop
Leu29	-16.5		Asp190''	-15.4	-19.2
Asp84	-38.8	-18.8	Glu195''	-22.4	-30.7
Asp87	-18.2	-17.4	Asp211''	-22.4	-21.8
Asp117	-26.1	-26.0	Asp217''	-17.2	-16.5
Asp152	-23.1	-24.2	Asp190'	-19.5	-23.4
Glu155		-16.4	Glu266'	-19.5	-17.7
Asp190	-95.3	-84.1	Asn272'	-20.4	
Glu195	-43.	-66.3	Gln298'	-59.7	-60.5
Glu266	-18.4	-18.4	Val318'	-81.3	-80.5
Asp322	-15.2	-17.3	Met323'	-34.3	-26.3
Glu326	-21.1	-21.7	Ala327'	-19.3	-16.0
Asp152''		-16.1	Leu354'	-15.1	
Glu155''	-16.5	-17.6	Met358'	-17.2	

Figure 4.30 shows those residues that contribute most to the stabilization of the open (red) and closed (blue) states, as determined from the differences in the interaction energies provided in Table **4.8**. Residues Leu29, Asp272', Leu354', and Met358' are primarily involved in stabilizing the closed loop state, while Glu155 and Asp152'' (from an adjacent monomer) favor the open state. These results imply that mutations in these residues could shift the conformational

equilibrium of the loop and potentially impact the catalytic activity of gpASNase1.

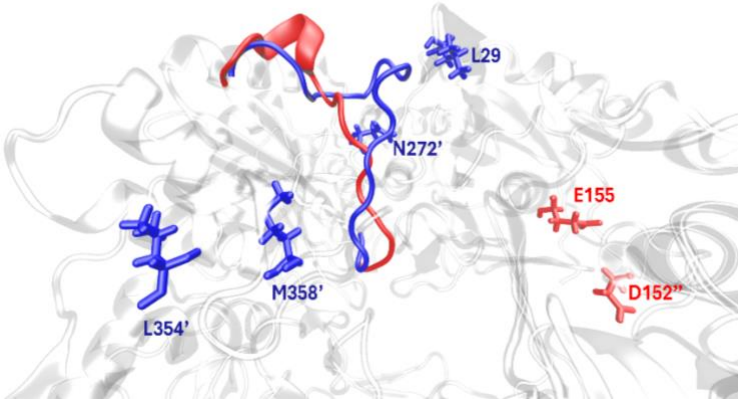


Figure 4.30. Residues with interaction energies greater than 15 kcal·mol^{-1} that stabilize only the open Tyr-loop of gpASNase1 (red) and those stabilizing only the closed Tyr-loop (blue).

4.2.4 Reaction Mechanism in gpASNase1 and hASNase1

The conversion of asparagine to aspartate, is proposed to occurs in two stages: the formation of the acyl-enzyme complex and its subsequent hydrolysis (as illustrated in Figure 4.31)[32].

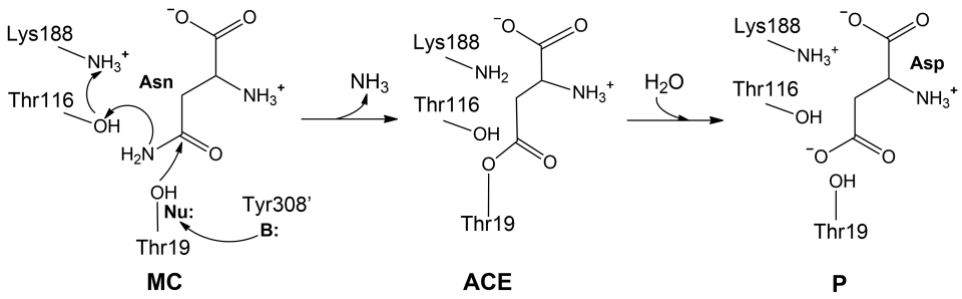


Figure 4.31. Asn hydrolysis mechanism in gpASNase1. Tyr308' (B:) serves as a base, activating the nucleophile, Thr19 (Nu:) to initiate the attack on the substrate. The Lys188/Thr116 pair protonates the leaving group, releasing ammonia and resulting in the formation of a covalent acyl-enzyme complex. In the second phase, the complex undergoes hydrolysis, yielding the final product (P).

Initially, we performed QM/MM simulations using the ASM approach at the DFTB3/MM level of theory, followed by a recalculation of the most promising mechanism at the B3LYP-D3/MM level. Our proposed mechanism is consistent with a previous model suggested by Sánchez and coworkers, based on their QM/MM optimizations.[32] Computational details are given below (see section 4.2.9.7).

The first phase of the reaction catalyzed by gpASNase1 involves the formation of an acyl-enzyme (ACE) complex between Asn and Thr19, accompanied by the release of ammonia. Figure 4.32 outlines the reaction mechanism for this stage (panel a) and the CVs employed to map the free energy landscape and the definition of the QM subsystem (panel b). The free energy profile obtained via the string method at the B3LYP-D3/MM level is shown in Figure 4.32c, with the corresponding CVs evolution depicted in Figure 4.32d.

The acyl-enzyme formation mechanism involves four transition states (TS1–TS4). Initially, Tyr308' is activated by a proton transfer to Asp117, facilitated by a chain of water molecules (see the evolution of CV1 to CV8 in Figure 34d). This leads to TS1, with an activation free energy of 11.7 ± 0.4 kcal·mol^{-1}. After reaching the intermediate I1, the deprotonated Tyr308' activates the nucleophile, Thr19, triggering the nucleophilic attack on the substrate, reaching TS2. A representative TS2 structure, the rate-limiting TS for the formation of the acyl-enzyme complex, is given in Figure 4.33.

As seen in Figure 4.33, TS2 is stabilized by interactions with the oxyanion hole residues spanned by residues Thr19 and Thr116, which provide hydrogen bonds to stabilize the negative charge build upon the carbonyl oxygen atom of the substrate. The activation free energy for TS2 (18.7 ± 0.7 kcal·mol^{-1}) is in close agreement with the experimentally derived value from k_{cat} (16.1 kcal·mol^{-1}).[28] Zero-point energy contributions were not included, which would likely lower the calculated value for the activation free energy.

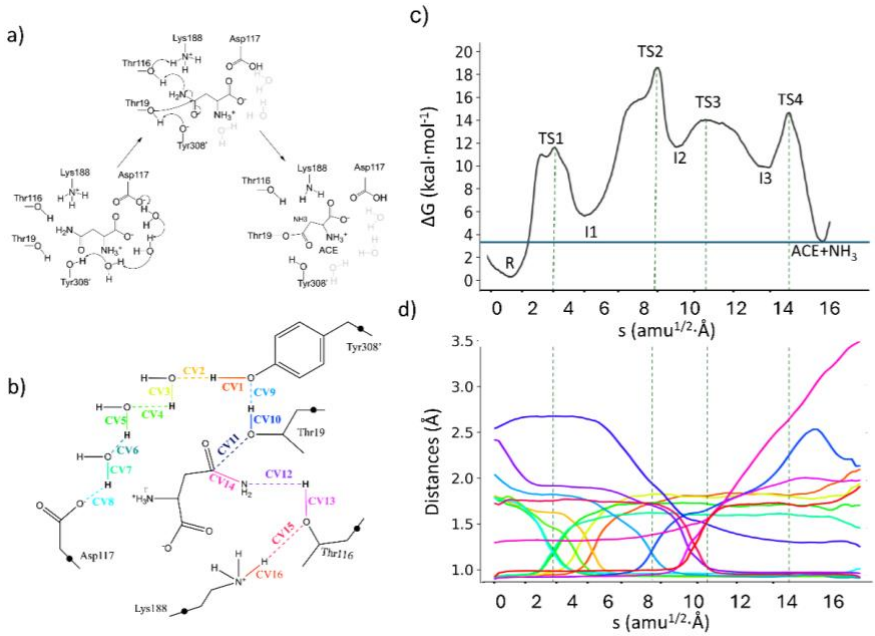


Figure 4.32. (a) Illustration of the mechanism of the acyl-enzyme (ACE) formation in gpASNase1. (b) Visualization of the quantum mechanical (QM) region and CVs employed to determine the minimum free energy path (MFEP), with the link atom depicted as a black dot. (c) Free energy landscape along the path collective variable (s), computed at the B3LYP-D3/6-31+G(d)/MM level of theory. (d) Evolution of the chosen CVs along the MFEP, with color codes matching panel (b). Dashed light-grey lines indicate transition state positions.

TS3 is formed by a proton transfer from Thr116 to the amino group of the substrate, assisted by Lys188 (see the evolution of CV15 and CV16 in Figure 4.32d). Finally, the Cγ-Nδ bond-breaking gives rise to TS4, with a free energy of 14.9 ± 0.7 kcal·mol^{-1}. Afterwards, the free energy drops monotonically, to reach the acyl-enzyme complex with ammonia still present in the active site. The free energy of this state is 3.9 ± 0.3 kcal·mol^{-1} above the Michaelis complex. In order to complete the acylation stage of the gpASNase1 mechanism, the free energy associated to ammonia release was calculated, similarly to what was done in the hASNase3 case (see Technical Details section for details). The free energy change for ammonia release was calculated to be -4.2 ± 0.7 kcal·mol^{-1}. Then, the overall reaction free energy for the acylation stage results to be -0.3 ± 1.1 kcal·mol^{-1}.

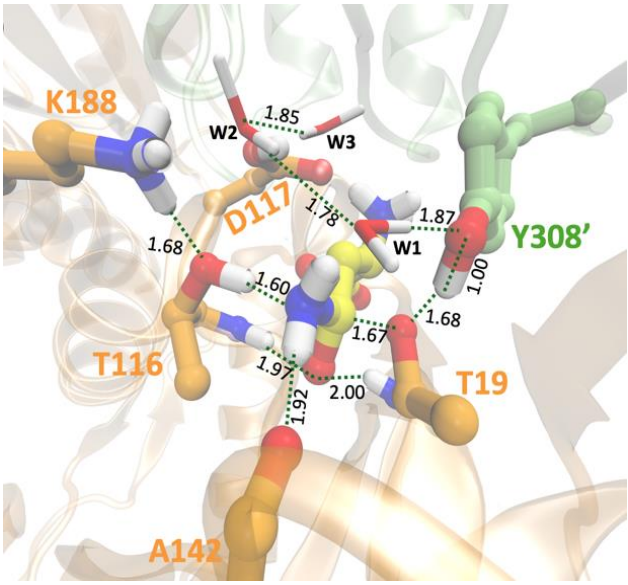


Figure 4.33. Structure of transition state TS2 involved in the acyl-enzyme complex formation in gpASNase1. All distances are provided in Å.

Given that hASNase1 shares the same active site residues, the same mechanistic proposal was explored. Even though the collective variables followed the same evolution (Figure 4.34b), a significantly higher energy barrier was observed in this case (26.5 kcal·mol^{-1}, see Figure 4.34a). The QM region, the choice of CVs and the initial guess for the ASM were identical as in the gpASNase1mechanism. This result is likely due to the fact that the initial model of the hASNase1 used for Michaelis complex preparation was predicted by AlphaFold2 [198], because a X-ray structure is not available. We suspect that the predicted structure corresponded to the apo form, where the enzyme is not in a reactive state. In the apo state, hASNase1 adopts an open donut conformation state, and it is only upon substrate binding that it transitions into the reactive holo state, characterized by a closed donut shape. To test this hypothesis, the prepared model of hASNase1 was superimposed onto the X-ray structures of EcASNase1 in both the holo (PDB: 2P2N) and apo (PDB: 2P2D) states. The RMSD of hASNase1 relative to the holo structure was 3.35 Å, while its RMSD to the apo structure was 1.83 Å. These values suggest that the system was likely still in the open state. Most likely this conformational change occurs on a timescale beyond the duration of the MD

simulations conducted in this doctoral thesis. Then the reaction profile presented was most probably obtained in a non-reactive state, resulting in the observed higher free energy barrier.

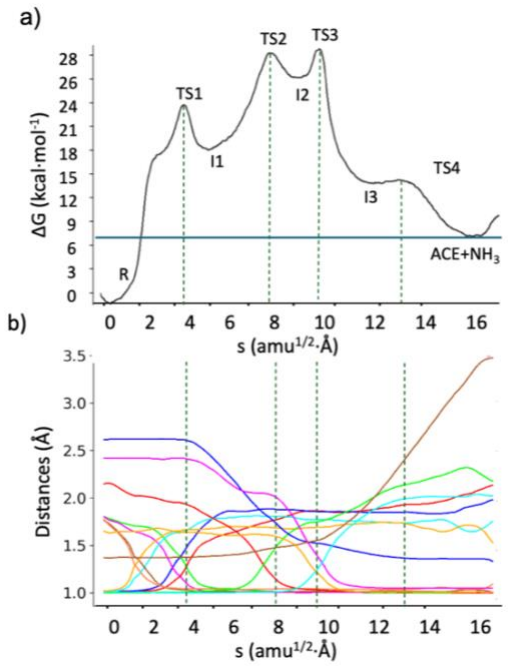


Figure 4.34. (a) Free energy landscape along the path collective variable (s), computed at the DFTB3/MM level of theory. (b) Evolution of the chosen CVs along the MFEP. Dashed light-grey lines indicate transition state positions.

The mechanistic proposal for the second stage in gpASNase1, the hydrolysis of the acyl-enzyme complex, together with the definition of the QM region and the CVs employed in the exploration are given in Figure 4.35a and Figure 4.35b. The overall free energy profile and evolution of the CVs along the mechanism are given in the Figure 4.35c and d, respectively.

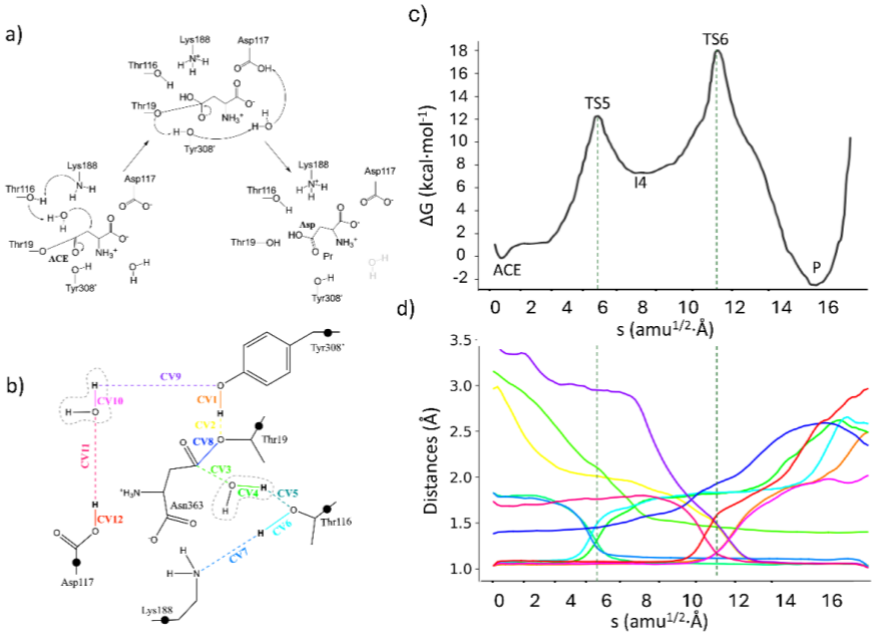


Figure 4.35. (a) Illustration of the mechanism of the acyl-enzyme (ACE) hydrolysis in gpASNase1. (b) Visualization of the quantum mechanical (QM) region and CVs employed to determine the minimum free energy path (MFEP), with the link atom depicted as a black dot. (c) Free energy profile along the path collective variable (s), computed at the B3LYP-D3/6-31+G(d)/MM level of theory. (d) Evolution of the chosen CVs along the MFEP, with color codes matching panel (b). Dashed light-grey lines indicate transition state positions.

The hydrolyses starts with the deprotonation of Thr116 by Lys188 (see evolution of CV5, CV6 and CV7 in Figure 4.35d). Once deprotonated, Thr116 activates the water for the nucleophilic attack, leading to TS5 with an activation free energy of 12.2 ± 0.7 kcal·mol^{-1}. Similarly, to TS2, TS5 is stabilized by the oxyanion hole formed by the main chain NH groups of Thr19 and Thr116 residues. After the tetrahedral intermediate (I4), a proton transfer from Tyr308' to Thr19 breaks the acyl-enzyme bond, giving rise to TS6 and leading to the formation of aspartic acid. This transition state has a free energy of 17.9 ± 0.7 kcal·mol^{-1} relative to the acyl-enzyme complex. A representative structure of TS6, the rate-limiting one in the hydrolysis stage, is represented in the Figure 4.36.

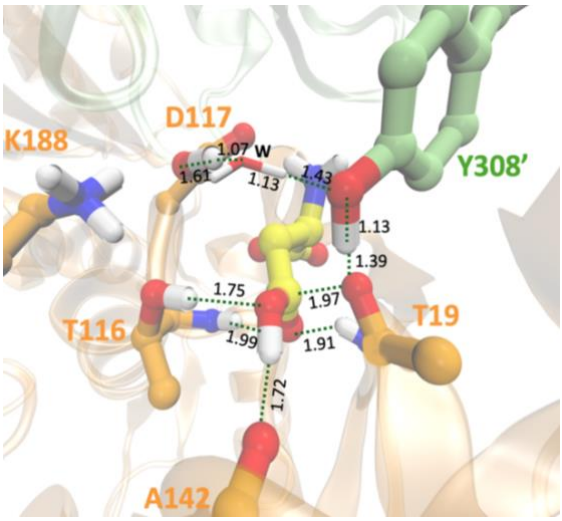


Figure 4.36. Structural representations of transition state TS6 involved in the acyl-enzyme complex hydrolysis in gpASNase1. All distances are provided in Å.

After TS6, the product formed is aspartic acid, protonated at the side chain carboxyl group, with the proton easily transferred to the solvent upon product release. The free energy of this product relative to the ACE complex is -2.1 kcal·mol^{-1}, and the overall enzymatic process, including ammonia release, is exergonic, with a reaction free energy of -2.4 ± 1.2 kcal·mol^{-1} relative to the Michaelis complex.

4.2.5 Electric Field Analysis

During the rate-limiting step (TS2), a negative charge is built upon the carbonyl oxygen of the substrate, making its stabilization crucial for facilitating the reaction. While the contributions of residues forming the oxyanion hole (Thr116 and Thr19) have been previously highlighted, other long-range electrostatic interactions from individual or groups of residues might also play a vital role in the catalysis. To investigate this, we computed the electric field generated by the surrounding environment in the midpoint of the carbonyl bond of Asn and its projection along the bond direction over a 1 μs trajectory of the acyl-enzyme state. As shown in Figure 4.37, a positive electric field indicates an electrostatic contribution that aids in the development of the negative charge on the carbonyl

oxygen, thereby stabilizing TS2. We then examined the contributions of individual residues and structural motifs by summing the effects of their respective residues. Figure 4.37b presents the contributions from residues with a projected electric field larger than 2 MV·cm^{-1}, while Figure 4.37c illustrates the per-motif contributions for the tetrameric structure of gpASNase1.

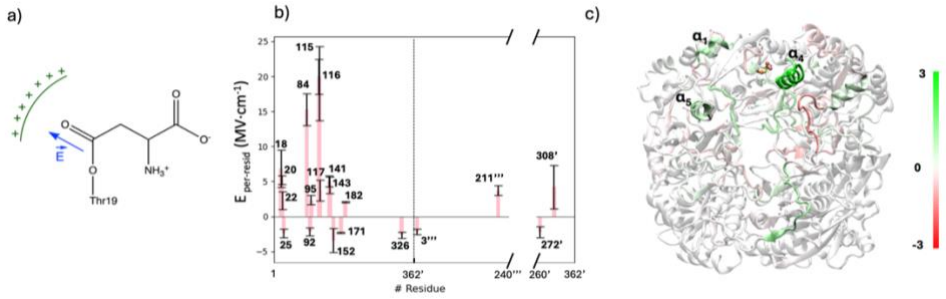


Figure 4.37. Analysis of the electric field distribution along the C=O bond of the substrate in the acyl-enzyme complex. a) Diagram illustrating the electric field (blue arrow) that facilitates the transfer of the negative charge from the substrate's carbonyl group to the oxygen atom. b) Projected electric field values per residue, showing only residues contributing more than 2 MV·cm^{-1} shown. Residues from the adjacent protomer (chain D) are marked with a prime, while residues from chains B and C are indicated with double and triple primes, respectively. Standard errors are represented as the standard deviation of the mean. c) Contribution of each motif to the projected electric field for the full tetrameric structure of gpASNase1 (Color scale units in MV·cm^{-1}).

As expected, residues located closer to the active site have a greater impact on the electric field along the carbonyl bond of the substrate (see Figure 4.37b). Among those with a positive contribution that help stabilizing the transition state, are Gly18, Leu20, Met22, Asp84, Gly115, Thr116, Asp117, Gly141, Gln143, Arg182, Asp211'', and Tyr308'. Many of these residues also play a role in substrate binding and in differentiating between Asn and Gln as substrates (see Figure 4.21). As shown in Figure 4.37b, negative contributions to the electric field, which difficult the development of negative charge on the substrate's carbonyl oxygen, are generally much smaller and limited to a small number of residues: Lys25, Asp92, Asp152, Glu171, Glu326, Arg3'', and Asn272'. Interestingly, Figure 39c reveals that two enzymatic motifs, α_1 and α_4, also

contribute to electrostatic stabilization of the transition states during the reaction in gpASNase1. The dipole moments of these α-helices generate a positive electric field along the carbonyl bond of the substrate. These long-range electrostatic effects, stemming from large structural motifs, should also be considered when selecting an appropriate scaffold for ASNase design.

4.2.6 Rationalization of the properties of the humanized chimeras

As outlined in the Introduction chapter of this thesis, one strategy to address current limitations in ALL treatment involves creating ASNase chimeras with improved immunogenic properties. One such approach is the development of a humanized version of gpASNase1, as proposed by previous studies.[29] These researchers used DNA shuffling to generate two chimeras, 63-hC and 65-hC, that share high sequence similarity with the human enzyme (hASNase1), while retaining the kinetic and binding characteristics of gpASNase1. The kinetic parameters of WT enzymes and the two most promising chimeras are given in the Table **4.9**.

Table 4.9. Kinetics parameters of the native gpASNase1, hASNase1 and chimeras designed using DNA shuffling method. The values were taken from [29].

Variant	K_M (μM)	k_{cat} (s^{-1})
gpASNase1	50	17
hASNase1	3500	41
63hN-hC	47	32
65hN-hC	74	40

Figure 4.38 summarizes the findings of the previous sections on hASNase1 and gpASNase1, showing the sequence alignment for the N-terminal domain of gpASNase1 and hASNase1 and the two chimeras (63-N and 65-N). Figure 4.38 also includes the results of multiple sequence alignments (MSAs) with similar proteins. To create MSAs, four rounds of HHblits searches[199] were carried out against the UniRef30 database (accessed on July 15, 2024) using E-value thresholds of $1 \cdot 10^{-50}$, $1 \cdot 10^{-30}$, $1 \cdot 10^{-10}$, and $1 \cdot 10^{-4}$. For each position in the sequence alignment, we calculated the occurrence frequency of each amino acid and identified the most conserved amino acid at each site. We then filtered each

position based on the conservation level of the most frequent amino acid and selected the top 50% of the most conserved sites.

Figure 4.38. Alignment of the N-terminal domain sequences of gpASNase1, the 63-N and 65-N chimeras, and hASNase1. The MSA results presented here focus on the top 45-55% of highly conserved residues in gpASNase1, marked with an asterisk. Residues identified as crucial for binding, loop closure, or electrostatic catalysis are highlighted in bold red. Residues that align with MSA results are shaded in light blue, while the α_1 and α_4 structural motifs are represented by yellow and green bars, respectively.

We have identified 72 key residues and two structural motifs (α_1 and α_4) as crucial for binding selectivity, reactivity, and for stabilizing the closed form of the

flexible Tyr-loop in gpASNase1. These residues are highlighted in red bold letters in Figure 4.38. Out of these 72 residues, the two chimeras, 63-N and 65-N already incorporate 65 and 63 residues, respectively. 30 of these 72 residues were preserved in the MSA (indicated by blue squares in Figure 4.38). Some of the mutated residues in the chimeras, such as Leu20, Ile96, Pro136, Ser191 and Asn318, were also preserved in the MSA. Out of these, 27 residues critical for gpASNase1 activity were present in both chimeras, including Ser7, His10, Lys25, Val36 and others. Finally, four more residues important for gpASNase1 function were mutated in at least one of the chimeras (Lys48, Arg147, Asn151 and Ala153).

The two chimeras, 63-hC and 65-hC, differ at positions 48, 50, 147, 148, 151, 153, 265, and 275 (Figure 4.38). In 65-hC, these residues are maintained as in hASNase1, while in 63-hC, they are mutated to match the sequence in gpASNase1. Our findings suggest that most of these residues form stronger interactions with the substrate in gpASNase1 than in hASNase1 (Figure 4.39), which correlates with the lower K_M value observed in 63-hC than in 65-hC, indicating better substrate binding. Additionally, residues at positions 59 (Asp), 68' and 68''' (Arg) in the chimeras display weaker interactions with the substrate compared to the corresponding residues in gpASNase1 (His). These residues are distant from the active site, but their charged nature creates long-range coulombic repulsive interactions, weakening binding of the substrate. Mutating these positions to the corresponding residues in gpASNase1 (Asp59His and Arg68His) could improve substrate binding.

Our results also indicate that mutations Arg52Gln, Glu58Asp, Glu59His, and Ile72Val could stabilize the closed form of the Tyr-loop (Figure 4.39). Interestingly, Val72 is conserved in the MSA (Figure 4.38) but has not been mutated in either chimera. Regarding the enzyme selectivity for Asn over Gln, no significant changes in selectivity are expected between the chimeras, as residues crucial for this property (Gly18, Met22, Asp84, Ser85, Arg95, His114, Gly115, Asp117, Ala142, and Asp190) are already incorporated into both chimeras.

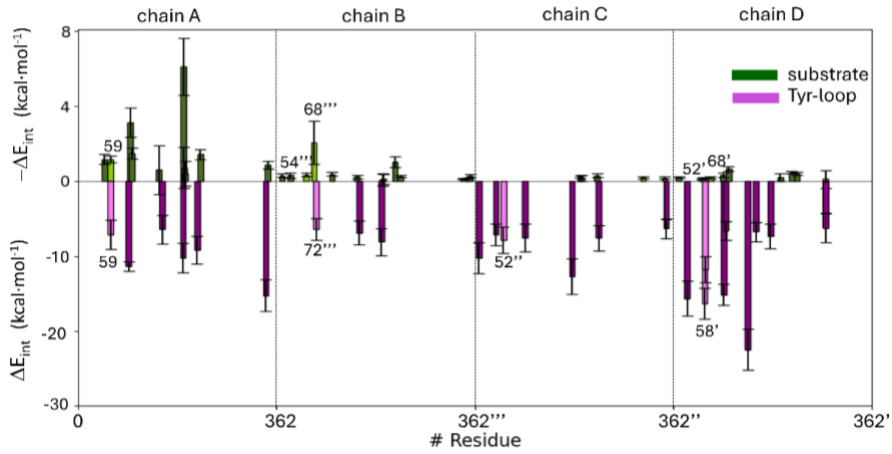


Figure 4.39. Comparison of interaction energy differences between hASNase1 and gpASNase1. Green histograms show the difference in average interaction energies of the substrate with each residue in hASNase1 versus gpASNase1 (-($E_{Asn\text{-}h,i}$ - $E_{Asn\text{-}gp,i}$)), with negative values indicating a preferential interaction with the guinea pig enzyme. Magenta histograms depict the difference in interaction energies of the closed Tyr-loop with each residue in hASNase1 versus gpASNase1 ($E_{int,\ loop\text{-}gp} - E_{int,\ loop\text{-}h}$), with positive values indicating a preferential interaction in the guinea pig version. Only residues with $E_{int,Asn}$ below -0.1 kcal·mol^{-1} and Eint,loop below -10 kcal·mol^{-1} are included. Residues from the adjacent protomer are labeled with a prime, while those from chains B and C are labeled with double and triple primes, respectively. Darker histograms represent residues incorporated in the chimeras. Energy values are in kcal·mol^{-1}, with error bars showing standard deviations.

Finally, most residues identified as important for electrostatic stabilization of the rate-limiting transition state (18, 20, 22, 84, 95, 115, 116, 117, 141, 143, 182, 211, and 308) are conserved between gpASNase1 and hASNase1, which may explain their similar kinetic properties.

4.2.7 Immunogenicity of gpASNase1 structural motifs

Two structural motifs, α_1 and α_4, were also shown to contribute to the electrostatic stabilization of the rate-limiting transition state. Remarkably, the α_4 motif is found to be a well-preserved motif in the MSA (see Figure 4.38). Considering that the goal of ASNases is the clinical application in the treatment of ALL, these motifs were tested as whether they present a significant immunogenic response

(details are given in the Technical Details section). The predicted allergenic peptides from gpASNase1 for the HLA-DRB1_0701 allele are shown in Figure 4.40. Strong immunogenic epitopes are the motifs whose binding percentile rank was above 10. As shown in Figure 4.40, neither of the two structural motifs (α_1 and α_4) are predicted to exhibit strong binding to the HLADRB1_0701 allele, which is linked to an elevated risk of hypersensitivity reactions and allergies following bacterial ASNase treatment.

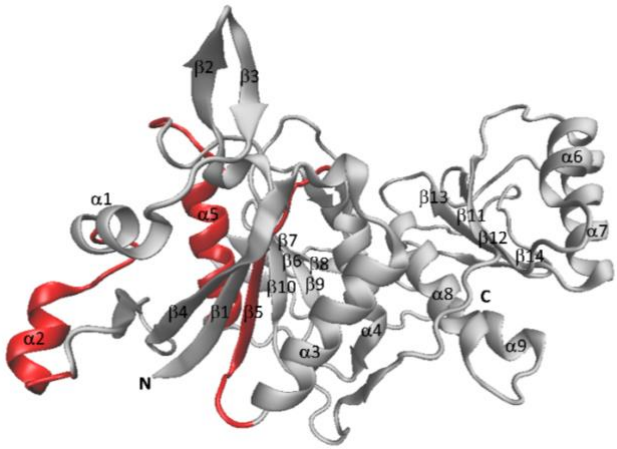


Figure 4.40. Structural motifs of gpASNase1 monomer predicted to have strong binding to HLA-DRB1*07:01 allele are shown in red, while grey areas indicate motifs with low to no binding.

4.2.8 Short summary of gpASNase1 and hASNase1 results

Active site interactions and reaction mechanism: Analysis of the MD simulations of the Michaelis complex with asparagine present in the active site identified critical interactions for the catalytic activity. We found that, at neutral pH, Lys188 is protonated and cannot easily act as a catalytic base. Instead, Thr19 can be activated to act as the nucleophile through proton transfer to Tyr308', which in turn can be deprotonated by means of a water-mediated proton transfer to Asp117. The close proximity of the substrate's Cγ atom to the hydroxyl oxygen atom of Thr19 supports its role as the nucleophile, essential for forming the acyl-enzyme complex. The rate-limiting step corresponds to the nucleophilic attack of

Thr19 on the substrate, coupled to its deprotonation by Tyr308', and presents an activation free energy of 18.6 kcal·mol^{-1}, in good agreement with the experimentally derived value. Electric field analysis revealed that residues Gly18, Asp84, Thr116, and Tyr308', along with the α_1 and α_4 helices, play a crucial role in stabilizing the negative charge on the substrate's carbonyl oxygen during the rate-limiting transition state, enhancing enzymatic catalytic efficiency.

Flexible loop dynamics: We investigated the conformational dynamics of the Tyr-loop, which contains the catalytic Tyr308' crucial for nucleophile activation during catalysis. The results show that loop closure is favored upon substrate binding, with the closed form being more stable in the holo enzyme and the open form in the apo enzyme. Additionally, per-residue interaction energy analysis identified key residues involved in stabilizing both the open and closed loop conformations.

Gln/Asn Selectivity: Thermodynamic integration calculations estimated the relative binding free energy preference for asparagine over glutamine, aligning with the experimental observation of the lack of glutaminase activity in gpASNase1. The analysis of the interaction energy per-residue contributions revealed that key active-site residues, including Asp84, Ser85, Thr116, and Asp117, play a significant role in asparagine binding, interacting with the charged carboxylate and amino groups of the substrate. Additionally, Asp190 and Ala142 contribute to interactions with the non-zwitterionic portion of asparagine, further favoring its binding over glutamine.

Rationalization of humanized chimeras: Considering all the results obtained, along with the comparison of the per-residue contributions to the binding free energies of hASNase1 and gpASNase1, around 40% of the mutations in the chimeras were explained. Importantly, six additional mutations (Arg52Gln, Arg54Gln, Glu58Asp, Asp59His, Arg68His, and Ile72Val) that could improve chimera properties were identified. The two structural motifs, α_1 and α_4 helix, preserved in the MSA and crucial for transition state stabilization, were also predicted not to cause significant allergenic reactions.

4.2.9 Technical Details

4.2.9.1 System Preparation

The Michaelis complex structure was prepared using the PDB structure 5DNC [33], which represents the N-terminal domain homotetrameric form. The N-terminal domain was used given that experimental evidences indicate that truncation of the C-terminal domain of gpASNase1 does not affect its catalytic activity or K_M value. To model a reactive complex, the mutation of the catalytically inactive Thr19Ala mutant was reverted in all four monomers. The conformation of Thr19 was adjusted so that its hydroxyl group aligned with the substrate, similar to the orientation observed in the structure with PDB code 4R8L [28], which corresponds to the wild-type enzyme with the product (aspartate) in the active site. Residues 1–7 and 361–362, unresolved in the 5DNC structure, were modeled using AlphaFold2.[153] All water molecules and sodium ions from the X-ray structure were retained, and missing hydrogen atoms were added using the Protein Preparation Wizard tool in Maestro.[200] Protonation states of titratable residues at pH 7.4 were calculated using PROPKA3.[184] Lys188 was modeled in the protonated state, based on pK_a calculations, although the possibility of a deprotonated state was also explored using free energy methods. The substrate (Asn or Gln) was then placed in the active site of the gpASNase1, initially maximizing all the interactions with the active site residues observed in the X-ray structures. Additionally, the apo form of the gpASNase1 was modeled starting from the 4R8K structure. [28]

The initial hASNase1 structure was obtained with AlphaFold2.[153] The sequence used for structure prediction corresponds to the N-terminal domain of the UniProt entry Q86U10. All the other steps (solvatation, simulation box dimensions, protonation states of the residues, etc.) were performed using the same protocol as in the gpASNase1 model.

4.2.9.2 Molecular Dynamics Simulations

Molecular dynamics simulations were performed using the Amber22 GPU version of pmemd [186]. Substrate parameters for free Asn and Gln in their

zwitterionic forms were obtained from Horn et al.,[187] while protein residues were described using the ff14SB force field. For the acyl-enzyme complex, parameters were generated using AmberTools [188] , with atomic charges calculated via the restrained electrostatic potential (RESP) method at the HF/6-31G* level.[201] The systems were solvated in a TIP3P water box [141] with a 12 Å padding from protein-substrate atoms to the simulation box boundaries. Sodium ions were added to neutralize the total charge.

Simulations began with energy minimization, comprising 1000 steps (20 steps with the steepest descent method followed by the conjugate gradient method). The structures were then gradually heated from 100 K to 310 K using Langevin dynamics with a collision frequency of 1.0 ps^{-1} and a 1 fs time step, applying restraints on the protein heavy atoms (100 $kcal \cdot mol^{-1} \cdot Å^{-2}$). After heating, the systems underwent relaxation at constant pressure for 1 ns, followed by an additional 1 ns where the restraint force on heavy atoms was reduced to 10 $kcal \cdot mol^{-1} \cdot Å^{-2}$. Afterwards, a minimization was performed, restraining only the protein backbone, allowing side chains to readapt. After this, short equilibration simulations at constant pressure were run, with the restraint force on backbone atoms progressively lowered until all restraints were removed. In all the simulations electrostatic interactions were calculated using particle-mesh Ewald,[190] and a 10 Å cutoff was applied to non-electrostatic interactions.

Production simulations of 1 μs were carried out in the NVT ensemble, using periodic boundary conditions and maintaining a temperature of 310 K with a Langevin thermostat. The time step was increased to 2 fs using the SHAKE algorithm to constrain hydrogen-involving bonds.[189] Three 1 μs replicas were run for both holo and apo forms of the gpASNase1 and three 1 μs replicas were run of the holo hASNase1 system, each initiated with different velocities.

4.2.9.3 Electric Field Analysis

To analyze the effects of the electric field in the molecular dynamics simulations, a modified version of the TUPÃ software was used.[98] Since the mechanistic proposal involves the accumulation of negative charge on the carbonyl oxygen

atom of the substrate, the midpoint of the C=O bond was chosen as the probe for calculating the electric field and its projection. This projection was determined using a unit vector defined from Oδ to the Cγ atoms of the substrate. Electric field analysis was conducted on snapshots from the MD simulations of the acyl-enzyme state, using a solvent cutoff radius of 15 Å. Additionally, TUPÃ per-residue decomposition option was used to provide detailed insights into the contributions of each residue and structural motifs to the electric field. [35]

4.2.9.4 MMGBSA Calculations

To evaluate the per-residue contributions to interaction energies, we employed the Molecular Mechanics Generalized Born Surface Area (MMGBSA) approach using the MMPBSA.py [202] script available in AmberTools22.[188] Complex, receptor, and ligand topology files were prepared using the Ante-MMGBSA.py utility, with GB solvation radii defined as mbondi2 and a salt concentration of 0.1 M. Per-residue interaction energies, including electrostatic and van der Waals contributions, were calculated. Unlike the default decomposition scheme in MMPBSA.py, which equally distributes interaction energy between the ligand and receptor residues, the full contribution was attributed to each residue.[203] The calculations were performed on uncorrelated frames extracted at 100 ps intervals from a 1 µs trajectory of the gpASNase1 and hASNase1 complexes with its substrate. The statistical error for the interaction energy contribution of each residue was determined as the standard deviation of 10 averaged values.

4.2.9.5 Thermodynamics Integration

To determine the pK_a of Lys188, Thermodynamic Integration (TI) was used to estimate the free energy change during the alchemical transformation between the neutral and protonated states, following a computational setup similar to the one previously described for the Thr168 in the section 4.1.4.3. of this chapter. Similarly, the calculation of the free energy change for the ammonia leaving the active site was performed using the protocol described in the same section.

The relative binding free energy between asparagine and glutamine was determined through an alchemical transformation between their bound and unbound states. Both substrates, Asn and Gln, were handled using the softcore potential to ensure a smooth transition. The number of λ windows was increased to 20 to improve the overlap between adjacent states. Transformations were performed in water and within the active site of gpASNase1, and the corresponding free energy change was determined. The relative binding free energy (ΔΔG) was obtained by subtracting the free energy change in water from that obtained in the active site, as outlined in the following equation:

$$\Delta\Delta G_{bind} = \Delta G_{bind,Asn} - \Delta G_{bind,Gln} = \Delta G_{bulk} - \Delta G_{site} \quad (4.5)$$

To achieve better estimations, the final values were calculated as an average value of 5 replicas.

In the case of the determination of the pK_a of Lys188, the Thermodynamic cycle given in Figure 4.41 is followed. The corresponding free energy values are given in Table 4.10.

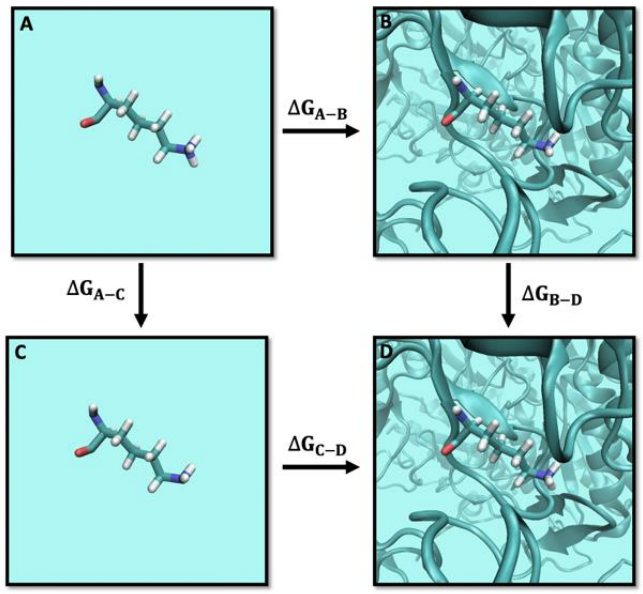


Figure 4.41. Thermodynamic cycle depicting the change in protonation state of the Lyus188 in water and protein environments. States A and C represent protonated and deprotonated Lys in water, respectively, while states B and D correspond to protonated and deprotonated Lys188 within the gpASNase1 enzyme. Alchemical transformations were carried out along the vertical axes from A to C and from B to D.3/11/2025 10:31:00 PM

Table 4.10. Free energy changes computed from five independent replicas (rep) of alchemical transformations between the protonated and unprotonated states of lysine. The calculations were carried out in aqueous solution (aq), as well as in holo protein environment. The indices of the free energy changes ΔG are explained in the Technical Details section. Both the average values and the associated standard deviations (std) are given in kcal·mol^{-1}.

	replica	ΔG_{A-C}		replica	ΔG_{B-D}
aq	1	-4.48	holo protein	1	2.23
	2	-4.48		2	3.15
	3	-4.46		3	3.16
	4	-4.46		4	3.30
	5	-4.47		5	3.74
mean		-4.48	mean		3.12
std		0.02	std		0.99

4.2.9.6 Free Energy Calculation of Loop Conformational Changes

To investigate the free energy profiles of the flexible loop conformational change between open and closed states in both the apo and holo forms, we used the Adaptive String Method with the CVs described above (see section 4.2.4 Reaction Mechanism in gpASNase1 and hASNase1). The starting points for the nodes of the string simulations were were extracted every 10 ns from 1 μs MD simulations for each state (open and closed). These simulations followed the previously described (see section 4.2.9.2) setup including a light restraint (10 $kcal \cdot mol^{-1} \cdot Å^{-2}$) to keep the substrate in the active site during the open-holo gpASNase1 simulations. RMSD of the CVs were followed along the dynamics of the string and once converged (RMSD ~0.1 $amu^{1/2} \cdot Å$) the path coordinate (denoted as *s*) was defined to measure the system's position along the MFEP. US calculations were run along this coordinate [90], and the free energy profile was estimated using the Weighted Histogram Analysis Method (WHAM).[195] Technical details are given in the Table 4.11.

Table 4.11. Details of ASM calculations for the conformational change of the Tyr-loop in gpASNase1.

	apo	holo
String nodes	88	88
Number of CVs	5	5
REX period (fs)	500	500
String friction	1000	1000
Force friction	50	50
Preparation (ps)	10	10
String optimization (ns)	30	30
Umbrella Sampling (ns)	10	10
Timestep (fs)	2	2

4.2.9.7 QMMM Simulations and Adaptive String Method

Preliminary calculations utilized GFN2-xTB/MM [124] and DFTB3/MM levels [123] for initial exploration of the free energy landscape. For the acyl-enzyme formation step, three water molecules were included in the QM region (Figure

4.32b), and two water molecules were set in the QM region for the acyl-enzyme hydrolysis step (Figure 4.35b). The best mechanistic candidate was recalculated describing the QM subsystem using the B3LYP functional [114, 115] with D3 dispersion [192] corrections and the 6-31+G(d) basis set. These QM level was chosen for its proven accuracy in previous studies on type 1 and 2 ASNases.[76] These QM/MM calculations were carried out using Amber24 coupled with Gaussian16 [194] for density functional theory computations.

To investigate various mechanistic possibilities, free energy profiles were determined using the ASM, starting from different initial guesses for each mechanism. CVs were defined as the distances of bonds being formed or broken during the reaction. Sampling efficiency was enhanced by assigning a mass of 2 a.m.u. to all transferring hydrogen atoms and using a 1 fs time step. Technical details of the ASM calculations run to explore the reaction mechanisms are given in Table 4.12.

Table 4.12. Details of ASM calculations for the reaction mechanism in gpASNase1 and hASNase1.

	Acyl-enzyme formation	Hydrolysis of the acyl-enzyme
String nodes	96	96
Number of CVs	15	12
REX period (fs)	50	50
Preparation (ps)	0	0
String optimization (ps)	4	4
Umbrella Sampling (ps)	10	10
Timestep (fs)	1	1
Cutoff for DFT/MM interactions (Å)	15	15

4.2.9.8 Prediction of Epitopes in T-Cells and Determination of Epitopes Density

The gpASNase1 segments that could potentially bind to Major Histocompatibility Complex class II (MHC II) molecules were predicted using the NetMHCIIPan 4.1 EL tool.[204] This tool utilizes a neural network model

that forecasts MHC binding affinities based on an amino acid sequence, trained on a large dataset of peptide-MHC class II binding information covering thousands of human MHC molecules. We focused on the HLADRB1_0701 allele due to its known association with an increased risk of hypersensitivity reactions and allergies following bacterial ASNase treatment.[205]

4.3 *De novo* design of the soluble Epoxide Hydrolase (sEH)

This chapter presents part of the results obtained during my research stay at the Baker Lab, at the Institute for Protein Design within the Department of Biochemistry at the University of Washington (Seattle, Washington, USA). These results are the outcome of a collaborative effort of several researchers and are not exclusively mine. I began working on this project to learn the pipeline for the *de novo* protein design and to better understand the techniques developed at the Institute for Protein Design. A significant portion of this research was made possible thanks to the mentorship and support of the PhD student Anna Lauko and the postdoctoral researcher Samuel J. Pollock at the Institute for Protein Design. A key part of this work was directly guided by Prof. David Baker, who introduced me to the master student, Stann Van Baaren, from the University of Paris-Saclay. Stann explored various sets of *de novo* epoxide hydrolase designs before I arrived, without observing any enzymatic activity in any of his designs. As Stann's research stay was coming to an end, Prof. Baker suggested that I could gain valuable experience from this project and push it further. He envisioned that, as Stann did, I could eventually pass it on to the next student before my departure. When I joined the project, I first attempted to explain the lack of the activity in the previous designs using simulation tools and then provided some theoretical foundation and ideas to be incorporated into new designs. In this chapter, only some of the designs that integrated my proposals, which Stann and I worked on together, will be presented.

4.3.1 Introduction

Epoxide hydrolases (EHs) are a family of hydrolase enzymes that catalyze the conversion of epoxides into diols. Epoxides can arise through endogenous biochemical pathways or via the oxidation of xenobiotics, such as drug molecules, by cytochrome P450 enzymes.[206–208] Their hydrolysis therefore plays a crucial role in drug metabolism, as epoxides are highly reactive and can form harmful adducts with proteins and DNA, leading to cellular damage. The *de novo* design of EHs could provide efficient, stable, and substrate-specific enzymes for therapeutic, industrial, and synthetic applications.[209] Engineered

EHs can improve drug metabolism, detoxify harmful epoxides, and facilitate green chemistry, while also enhancing the preparation of enantiopure diols and epoxides, key synthons in organic synthesis.

There are several EHs types [210]. Soluble epoxide hydrolases (sEHs) have been the most widely studied type due to their role in metabolizing bioactive lipids and their implication in various diseases.[210] sEHs belong to the α/β-hydrolase family, given that the residues of the conserved catalytic triad (nucleophile-histidine-acid) are positioned within this α/β-motif. The structure selected as starting point in this work was the *Solanum tuberosum* sEH (see Figure 4.42).

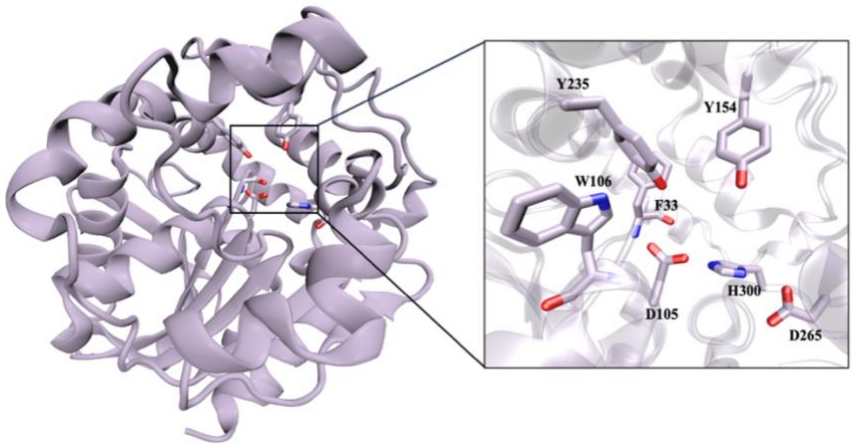


Figure 4.42. The structure of the *Solanum tuberosum* soluble epoxide hydrolase (PDB code: 2CJP [211]). On the right, a close-up view of the active site residues is presented.

In this enzyme the catalytic nucleophile (Asp105) is located on a loop between a β-strand and an adjacent α-helix. The catalytic histidine (His300) and aspartate (Asp265) are both located on the loops connecting β-strands and α-helices. This triad is located in the so called "αβ-fold domain". Two Tyr residues, Tyr154 and Tyr235, form the oxyanion hole and are located on the two adjacent α helices, the so called "lid domain".

The catalytic mechanism in sEH is well-characterized and consists of two-steps (see Figure 4.43): acyl-enzyme formation (step 1) and hydrolysis (step 2), resembling the catalytic mechanisms of asparaginases presented before.

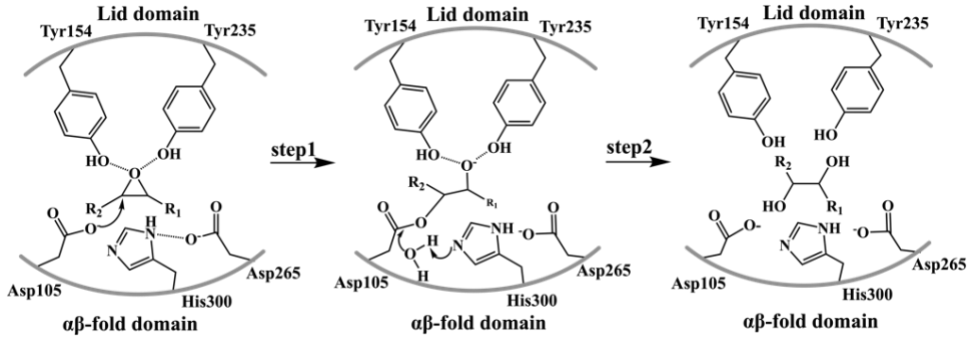


Figure 4.43. Proposed reaction mechanism for sEH. Adapted from [212].

The conserved Asp105 acts as the catalytic nucleophile, forming an ester intermediate with the substrate in the first step of the reaction (step 1 in Figure 4.43). The negative charge built on the oxygen atom of the epoxide ring gets stabilized by the oxyanion hole spanned by the two tyrosine residues mentioned above. Additionally, these two tyrosine residues that form hydrogen bonds with the oxygen of the epoxide ring also enhance the electrophilicity of the substrate and position the substrate optimally for the nucleophilic attack by the aspartate.[213] The other two key residues, a histidine (His300) and an aspartate (Asp265), create a charge relay system that activates a water molecule via proton abstraction, enabling the hydrolysis of the ester intermediate in the second step.[213] In addition, the backbone amino groups of nearby residues Phe33 and Trp106 (see Figure 4.44) have also been reported to form an oxyanion hole, stabilizing the negative charge on the carboxylate oxygen of Asp105 (not the nucleophilic one). The backbone carbonyl group of Phe33 has also been reported to be crucial for catalysis as it can aid in orienting the nucleophilic water during the hydrolysis, by forming the hydrogen bond with the water molecule, positioning it between His300 and Phe33.[214] It is also interesting mentioning that in the crystal structure itself (PDB code: 2CJP [211]), a crystallographic water molecule is positioned between these two residues (see Figure 4.44).

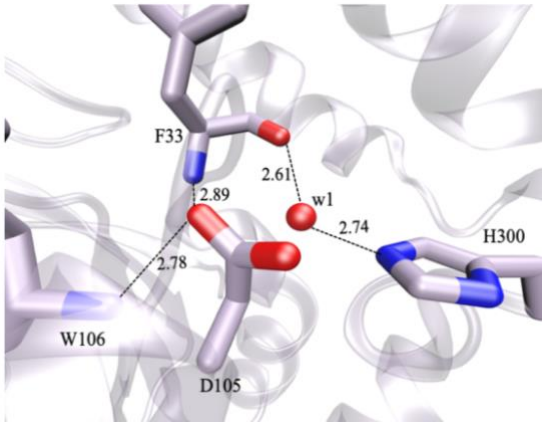


Figure 4.44. Hydrogen bonds formed between the nucleophilic Asp105 and backbone amino groups of Trp106 and Phe33 and positioning of the water molecule (w1) in between H300 and Phe33 in the X-ray structure of sEH (PDB code: 2CJP [211]). All distances are given in Å.

4.3.2 Computational design

The PDB structure 2CJP [211] provides a high-resolution model of a sEH, offering detailed information of its active site, including the positioning of key catalytic residues and the geometry of the acyl-enzyme intermediate. Before I arrived to the Institute for Protein Design, several options for the active site were selected to be scaffolded into *de novo* design. First a minimal active site containing only the triad was attempted (see Figure 4.45a), then the triad and two tyrosine residues (Figure 4.45b) and lastly the catalytic triad, the two tyrosine residues and the two backbone amino groups (Figure 4.45c). During the sequence design, these amino groups were allowed to take any identity as long as the amino groups were respecting the distance constraints assigned to keep them close to Asp105. Unfortunately, despite carefully following the design pipeline, none of the chosen variants exhibited any traceable enzymatic activity.

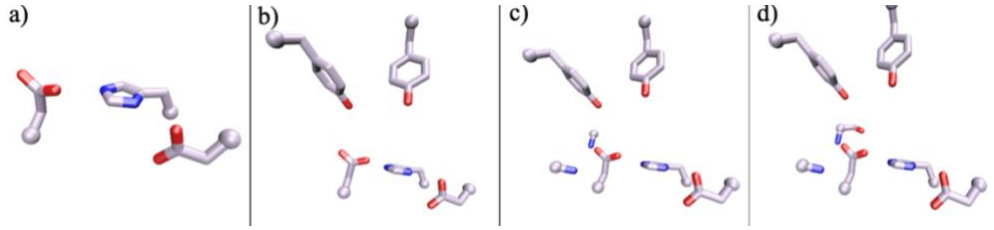


Figure 4.45. sEH active site templates to be scaffolded into *de novo* EH.

In a previous theoretical study, the QM model of sEH included the carbonyl backbone of Phe33, as the authors suggested that it could play a role orienting the nucleophilic water molecule and also stabilizing the tetrahedral intermediate during hydrolysis.[214] Based on this, I proposed incorporating this carbonyl group into the initial template. Without this group, the design might function primarily as a "substrate binder", leading to the accumulation of the acyl-enzyme intermediate without an efficient turnover because of the lack of a hydrolytic water in the active site. As a result, we developed a new template (Figure 4.45d) as a starting point for further optimization.

Given that the hydrolysis of an epoxide by sEH produces a diol with spectral properties similar to the parent epoxide, specialized substrates are often needed in order to follow the enzymatic activity.[215] α-Cyanoesters address this issue by undergoing O-deacylation, releasing a cyanohydrin intermediate that rapidly decomposes into the highly fluorescent 6-methoxy-2-naphthaldehyde. These substrates are highly sensitive, hydrolytically stable, and exhibit significant UV and fluorescence shifts upon hydrolysis. Therefore, phome ([cyano-(6-methoxynaphthalen-2-yl)methyl] 2-(3-phenyloxiran-2-yl)acetate) was used as a substrate, as its product releases 6-methoxy-2-naphthaldehyde (Figure 4.46) can easily be tracked fluorometrically.

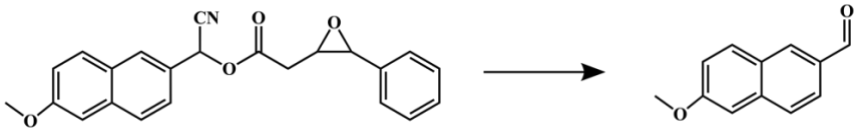


Figure 4.46. The phome substrate (([cyano-(6-methoxynaphthalen-2-yl)methyl] 2-(3-phenyloxiran-2-yl)acetate)) is hydrolyzed by sEH to give the fluorescent aldehyde 6-methoxy-2-naphthaldehyde (on the right).

We used tip-Atom CA version of the RFdiffusion software to generate protein backbones (of 160-220 aa). All the loops were selected for backbone remodeling using RF joint inpainting [174] to enhance the thermal stability of the designed scaffold at high temperatures. This approach generates multiple structures per each backbone generated with RFdifussion. For example, two structures with loops shorter than 15 residues are created, three structures are created in the case loops have between 15 and 20 residues, four with loops between 20 and 25 residues, and five with loops longer than 25 residues. We subsequently filtered all backbones based on the gyration radius (around 34 Å) and the solvent-accessible surface area fraction (< 0.25), as indicators of the overall compactness and stability of the protein structure. These values were empirically derived from natural enzymes of similar size and function with the purpose of filtering out too expanded proteins (which might indicate instability) ir too tightly packed ones (which could restrict active site flexibility). Additionally, we excluded backbone designs containing long alpha-helices (> 32 amino acids) and excessive loop regions, as these features could lead to unstable structures.

We then employed LigandMPNN [179] to generate sequences for the previously designed and filtered backbones. To preserve the catalytic geometry during the sequence design cycles, we incorporated Rosetta constraints on distances between amino acid residues and angles. The sequence design process involved three iterative cycles of LigandMPNN and Rosetta FastRelax [216]. Namely, once the sequence has been generated, its structure has been predicted and relaxed using Rosetta. [216] Then, such a relaxed backbone is being used as an input for the sequence design in the next steep and so on. Technical details are given in the corresponding section of this chapter.

The designed sequences were folded using AlphaFold2 [153] and the resulting structures were evaluated based on pLDDT scores and RMSD relative to the original ProteinMPNN-generated structures. Further filtering was performed using ChemNet [217]to assess structural accuracy and consistency. Given a protein backbone with a bound substrate, ChemNet randomizes the positions of the side chains within a given radius from the substrate. It then regenerates new coordinates for these groups, producing an ensemble of possible side-chain conformations and small-molecule docking poses through multiple iterative trajectories. For more details, consult the Technical Details section of this chapter.

MD simulations of the acyl-enzyme intermediate of sEH were run using Amber24 [133]. Additional details on the MD simulations are given in the Technical Details section. In the acyl-enzyme intermediate a covalent bond is formed between the substrate and the enzymatic residue Asp105 (Figure 4.47). This acyl-enzyme state resembles the transition state (TS), as the epoxide group is already opened and a negative charge has been developed on the oxygen atom (see Figure 4.47). By identifying designs whose ChemNet predictions aligned with the acyl-enzyme behavior observed in MD simulations, we could more effectively select candidates that stabilize the transition state, increasing the likelihood of catalytic activity. To achieve this, we analyzed the average distances between the active site residues and the substrate from the MD simulations (see distributions in the Figure 4.47) and selected designs whose ChemNet trajectories frequently adopted similar configurations.

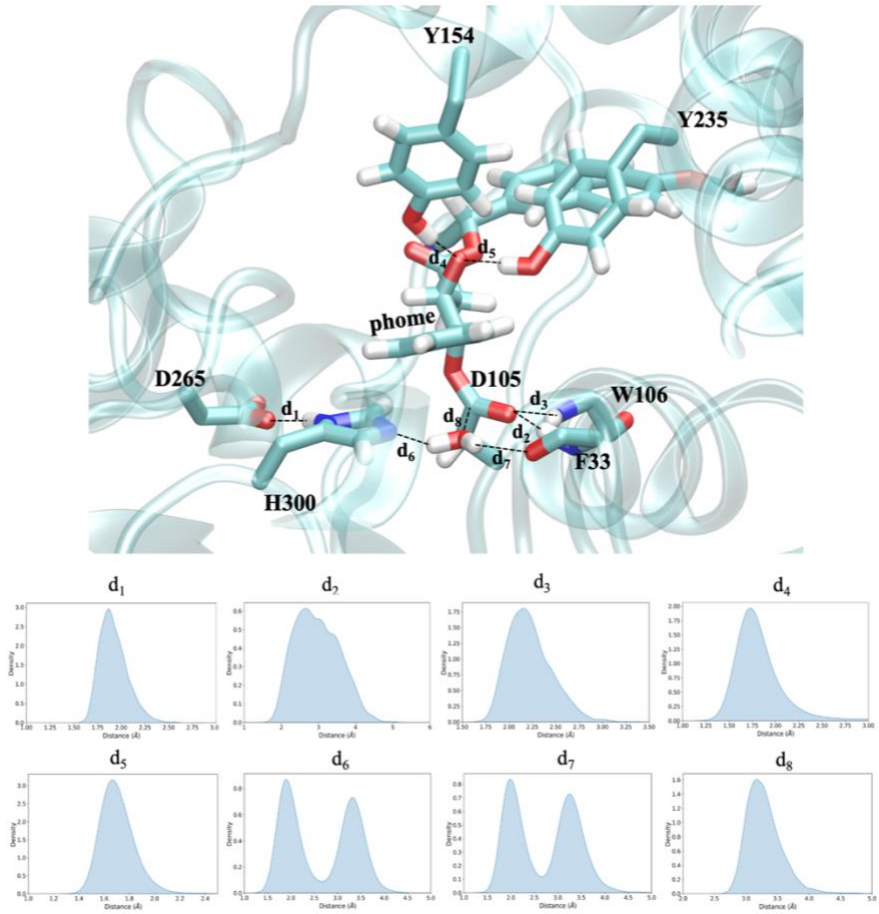


Figure 4.47. The structure of the acyl-enzyme formed between the Asp105 residue of sEH and phome from the classical MD simulation. Panels display the probability distribution of important distances (in Å) obtained from the 1 μs simulation.

Following this approach, we designed, expressed and tested 48 designs (detailed results of all the designs are not included in this thesis). Out of these, six variants (see Figure 4.48) demonstrated noticeable enzymatic activity and they will be discussed here. Their sequences are provided in the Technical Details section.

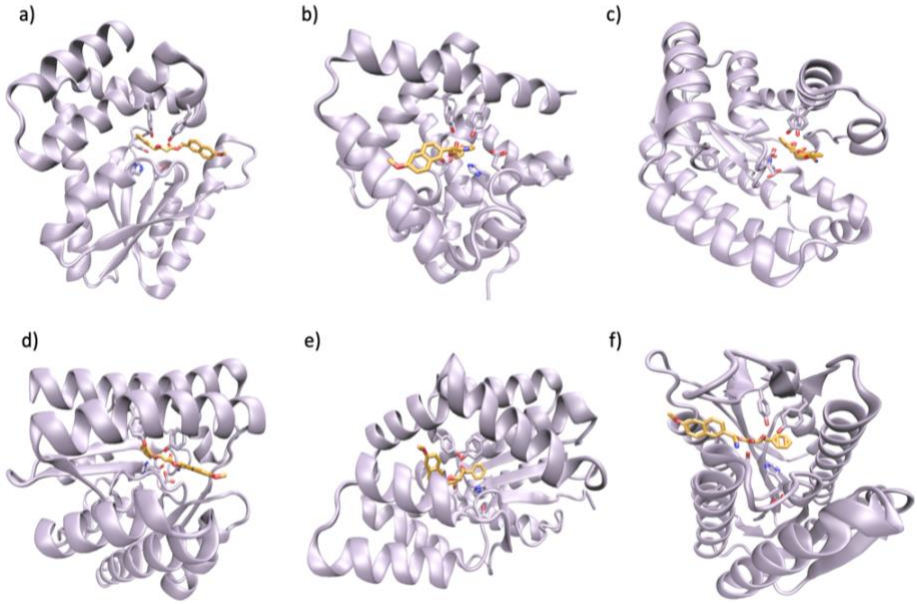


Figure 4.48. The structures of the active designs and active site residues as predicted by AlphaFold2. The backbone is given in the violet color, while the phome substrate is presented in orange stick model. a) Design A5; b) Design A11; c) Design B8; d) Design C5; e) Design C8; f) Design C11.

Interestingly, apart from the fact that backbone structures were able to accommodate the initial scaffold motif, it is noticeable that the catalytic triad residues were almost always consistently localized within the αβ-fold domain, mirroring the arrangement observed in the native backbone. Additionally, as shown in Figure 4.48, the active sites were centrally located within the protein structure and remained shielded from direct surface exposure. A well-defined substrate access funnel was present in all cases presented, allowing the bulky substrate (orange color representation in Figure 4.48) to effectively penetrate the active site and position itself optimally for interaction with the catalytic residues.

All active designs eluted as a single sharp peak on size-exclusion chromatography (SEC), indicating proper folding and monodispersity (data not shown). The concentration of the protein in each eluted sample was calculated by measuring absorbance at 280 nm with a NanoDrop spectrophotometer

(Thermo Fisher Scientific). Molecular weights and extinction coefficients calculated from the sequences using ProtParam tool are given in the Technical Details section.

The kinetic activity was monitored using a plate reader, measuring the fluorescence emission of each purified protein sample incubated with the fluorogenic substrate. The reaction was performed in 50 µL volumes using 96-well half-area plates and was continuously recorded at 30°C for 1 hour. More details are given in the Technical Details section.

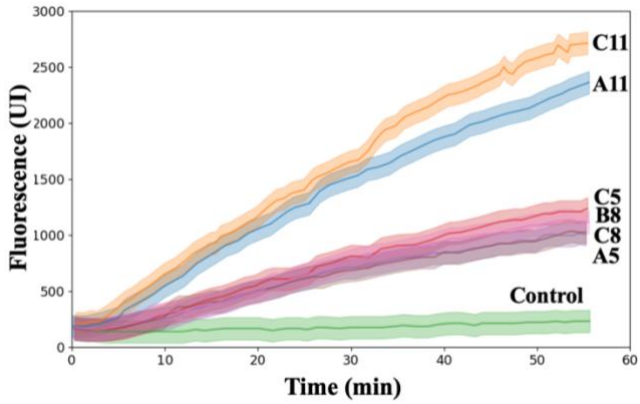


Figure 4.49. Michaelis Menten plots obtained by measuring the product fluorescence with $[S]_0$= 50 µM and [E]= 10 µM at pH = 7.4 and T=30ºC.

The Michaelis-Menten plots presented in Figure 4.49 did not reach a plateau, what it is likely since the substrate concentration did not reach a high enough concentration to saturate the enzyme. However, the six designs exhibit significantly higher activity compared to the control, remaining consistently more active even beyond the error margins, as indicated by the shaded regions. This suggests that the engineered designs show promising catalytic efficiency.

Initial velocities were used to determine catalytic efficiencies k_{cat}/K_M. Values obtained for each design are presented in Table 4.13.

Table 4.13. Catalytic efficiencies (k_{cat}/K_M) of the *de novo* variants of sEH.

Variant	$(k_{cat}/K_M) \cdot 10^3$ (M^{-1} s^{-1})
A5	0.50 ± 0.02
A11	1.22 ± 0.02
B8	0.59 ± 0.01
C5	0.63 ± 0.01
C8	0.50 ± 0.01
C11	1.44 ± 0.02

The native sEH has previously been shown to exhibits very distinct kinetic parameters for different substrates. [218] As a reference, the human sEH has been measured to have a specific activity with the phome substrate of 714 ± 23 nmol min^{-1} per milligram of protein and k_{cat}/K_M values of 2.6 μM^{-1} s^{-1}[215]. Even though the native enzyme displays significantly higher catalytic activity, the designs have promising potential for further optimization rounds. Although the k_{cat}/K_M values of our designs were lower than other *de novo* enzymes, such as the *de novo* serine hydrolase that reached 2.2×10^5 M^{-1} s^{-1} [166], it must be taken into account that in this case the resampling of backbones underwent several rounds and the filtering process was more extensive. The presented designs are thus good candidates for future improvements.

4.3.3 Summary

MD simulations proved to be an invaluable asset in guiding the *de novo* design of enzymes. By modeling enzyme-substrate interactions at an atomic level, MD simulations allowed for a deeper understanding of the interactions between the active site residues and the substrate, leading to more informed design decisions.

Following the standard de novo design pipeline developed in the Baker Lab (RFdiffusion, loops inpainting, LigandMPNN), aided with the analysis of the classical MD simulations, we obtained six designs of functional sEHs that exhibited k_{cat}/K_M comparable to native proteins.

4.3.4 Technical Details

4.3.4.1 Backbone generation

A new version of RFdifussion, refered to CA RFdiffusion (CA as in alpha-carbon), was used to generate backbones to scaffold motifs.[175] Given the Disclosure Policy, the code is the property of the Institute for Protein Design and therefore cannot be shared here. It has been developed as part of the work by Lauko et al. [166] and will be published once the manuscript of that paper has been accepted. Basically, in comparison to the all atom RFdifussion (AA RFdifussion), CA RFdiffusion is designed to generate protein backbone structures using a two-step process: generating first only alpha-carbon traces and reconstructing full backbones in the second, refinement, step. Additionally, instead of using static 3D coordinates for motif information, it encodes motifs through inter-residue pairwise distances and orientations, enabling flexible motif scaffolding without predefined rigid body transforms. This approach, developed independently alongside similar methods, enhances the ability to incorporate multiple discontinuous motifs while taking advantage of the pretrained RF all-atom network for structure prediction. In addition, the Tip-Atom program was used, which is specifically designed to recenter the generated backbone around a targeted active site or motif after the initial denoising steps. This ensures that the finally generated backbone is optimally scaffolded around the active site, with the motif properly buried rather than exposed on the surface.

In the input files for the CA RFdifussion, protein lengths between 160-220 amino acids were specified. The X-ray structure of the *Solanum tuberosum* Epoxide hydrolase (PDB code: 2CJP [211]) was used as a reference to provide the active site residues to be preserved in the scaffold. Generated backbones were then filtered based on gyration radius, visual discontinuities and some other structural considerations to ensure the quality of the generated models.

4.3.4.2 Sequence generation

In order to design sequences for the RFdiffusion-generated backbones, three cycles of LigandMPNN [179] and Rosetta FastRelax [216] were run. Catalytic

residues were fixed, and Rosetta enzyme constraints [219, 220] were applied during relaxation steps to preserve catalytic geometry. These constraints were defined for each hydrogen bond interaction based on the initial motif geometry, with tolerances of 0.1 Å for distances and 5° for angles and dihedrals.

4.3.4.3 AlphaFold and ChemNet

After sequence design, the designs were filtered based on their ability to recapitulate the catalytic geometry of the motif after FastRelax [216] and on the shape complementarity of the binding site with the substrate. The sequences that passed these filters were input into AlphaFold2 [153] for single-sequence structure prediction, using model 4 with three recycles. Designs were then filtered with a global Cα RMSD < 1.5 Å, pLDDT > 85, and catalytic residue Cα RMSD < 1.0 Å, similar to the pipelines followed before.[166, 168, 169] Designs that met these AlphaFold2 criteria were further analyzed using ChemNet.[217] ChemNet is a denoising neural network trained on X-ray and EM structures, designed to restore correct atom positions from partially corrupted input structures, provided that all chemical information about the system is known in advance. ChemNet predictions were performed for a spatial crop of 600 atoms nearest to the active site. The inputs included the protein backbone coordinates within the crop and the amino acid sequence, with side chain coordinates randomly initialized around the corresponding C-alpha atoms. For proteins lacking a crystal structure, the AF2 model was used as the input.

4.3.4.4 Variants sequences

Sequence design A5: MKIAIIDPTGGGLELINAVGDKLGIKAEVIPVR
GEYTSPEAYRAAIVEDAIRAADKAIEEGYAIVLDRSVTDQRLLDYVRGR
GYTREYPLFEDGHDGSYHTLYVHPSVPPEEQEKFRELAERVRGALERV
RAIDPETFDEARRRARELYHTKGGWALYDDPRNRTVYFGHIYAAVAV
ARGDPRAADYVAETAKAV

Sequence design A11: MLDEKQREAIKRAGVDYYVKETRAFLEAR

GYDADKIAGFIYGIVEKAKKGVSPLYVGLANHIALVGIIPSVDPFGTGIQ
VLAASHYISYEELVDFTVDLFDHRQAELEADPSIYSFDYWFEKLKEEFPE
LAGLGKEDFLEVAEELLKKDIDLAGKYKSLEEAEKKSQEIRDWYSKLTP
EERRAFAVEYAGKYYELRKKLES

Sequence design B8: KKINIAYYHGGFDPLAGGVGLLIDFQERIGNK
GTIFSLHLPIDPEIEEEYKKEAKAKGIKIEFFRELEPYYAEWGKLYKADP
NNPKALYKAIEELAPKVAELAIEYLEKNKDKIDVLIAIPDIILDYLRRIDP
EYVEKNIAAGPPVPVPDRNARKVGEYAKKLGIKVIEIPEEDIPKLDYDKL
NKLDKEAFKKEYLGQYEPLFDILEEAILN

Sequence design C5: KKKLTVDLLLDGGGAIGGGIELLYELATGERIPFE
EKLQRFLREFLEGTLEERLELYYSFAKEVLSKLEEKGIGFRLSVVDLPDG
SVTKIVRKYFGPMERVTTVHVSERWAEKYEEIKDTYLRELVELGWASG
PSITLFGDADTVRFLDAVGVDGSPLGALPEGTGLHLFYFVNDRGRSFVL
TIPDPFAWRGHDVSDPRYIEAAERLGRAIADRLGLDVDFPWRDLLPKAA
AAFRKLAAGEPLNEEEEEALVEFYRALYEEFLRRARAL

Sequence design C8: MRTVRAAGHAPYDAGLVLASDSREEARFYYELSR
FVYLNLLKWLAATGKVDVSAPIIDIGGTGLDQLYDDALAKLDAGATWE
EIFDFIRKNRKPGVLPEENPAVFDKLLELVEKAGIDEEKKKEFWEKYFSP
ENIEFYYNTGKELFKDRQDTELWLSVGGIPGYGSLLAALGAVEAGANL
RLVIIVAPEKAFEALPPEAQELILGYGEELRARGVRVEFFFITDEESAKRV
LKEVREVIEKWLES

Sequence design C11: GVAILLLPHLADTAMPYALRRLIAVLRDAGYEVI
GVITVDPMAIIPPSRPVFKERLKEFEGVLKVLKEAGIPVKKLYLVYLDEE
SKPSAEAMKEVAEKYGVKDVRIIPFSEYRQAVDDALAELDKLYKEVAD
KGAILVGLGGGAPYADTPGRVAEALREASELRWDLIIRSFQLAKEVGLE
TYLVAYAITGDFDIVDPETGEVLGRYTDAGNERLRERLAATGADKIFYA
EAPVDENNDVFGAARRAADGAIGQFADYLA

4.3.4.5 Molecular dynamics simulation of the sEH acyl-enzyme

Classical MD simulations were conducted using the Amber24 pmemd software.[133] Parameters for the substrate (phome) covalently linked to the Asp105 were obtained using the non-standard residue parameterization procedure implemented in Amber with the Antechamber program [221] from the AmberTools package [188], while the atomic charges were obtained using the Restrained Electrostatic Potential (RESP) method [201] at the HF/6-31G* level. Standard, amino acids were modeled with the ff14SB force field.[140] The system was solvated in a TIP3P water box [141] ensuring protein-inhibitor atoms were at least 12 Å from the simulation box edges, prepared using AmberTools tleap.[188] To neutralize the total charge of the system Na^+ were added.

Minimization was performed in cycles of 10,000 steps, beginning with 5,000 steps of steepest descent (SD) followed by conjugate gradient (CG) until the root-mean-square gradient reached ~10^{-4} kcal·mol^{-1}Å^{-1}. The minimized structures were then heated to 310 K using Langevin dynamics (collision frequency: 2.0 ps^{-1}) with a linear temperature ramp from 0 to 310 K and a 1 fs time step under periodic boundary conditions with isotropic position scaling.

A mild parabolic restraint (20 kcal·mol^{-1}Å^{-1}) was applied to the protein backbone during heating, and it was gradually released during equilibration (at a rate of ~1 kcal·mol^{-1}Å^{-1} ns^{-1}) and finally a 5 ns equilibration simulation was run without any restrains. The equilibrated configurations underwent 1 μs production simulations in the NVT with the time step increased to 2 fs using the SHAKE algorithm. [189] Electrostatic interactions were calculated via particle-mesh Ewald, [190] with a 10 Å cutoff for non-electrostatic interactions.

4.3.4.6 Protein expression and purification

Linear DNA fragments (eBlocks; Integrated DNA Technologies) encoding designs were cloned into Lm627 vector (Aggene 191551) flanked by the MSG residues at the N-terminal and by a SNAC [222] and his-tag at the C-terminal (His-tag) through Golden Gate (GG) cloning. Plasmids were subsequently

inserted into E. coli BL21 (DE3) via heat-shock and the transformation reactions were used to inoculate starter cultures in 1 mL of super optimal broth (SOC) media and kanamycin. After shaking overnight at 37°C, the dense cultures were diluted 100-fold into 50 mL of autoinduction medium with 50 μg/mL kanamycin.

These cultures were incubated at 37°C while shaking for 4 h, and then the temperature was lowered to 18ºC. The cultures were then kept incubating overnight at 18°C. Next day, the cells were harvested by centrifugation for 10 min at 4000g at the room temperature. The pellets were resuspended in lysis buffer (20 mM Tris, 300 mM NaCl, 40 mM Imidazole, 1 mg/ml lysozyme, 0.1 mg/ml DNAseI, pH 8.0) and disrupted by sonication on ice using a Qsonica Q500 instrument for 5 min (pulses off/on every 10 seconds of the 80% amplitude). Soluble fractions were centrifugated for 45 min at 14,000g at 4ºC. The supernatant was purified by affinity chromatography passing it through HisTrap Ni^{2+}-NTA resin (Qiagen) column equilibrated with a wash buffer containing 20 mM Tris, 300 mM NaCl, 40 mM Imidazole. The column was then washed with twice the amount of wash buffer before eluted with 20 mM Tris-HCl, 500 mM Imidazole, pH 8.0.

Eluted protein samples were purified by size-exclusion chromatography using ÄKTA pure system with a Superdex 75 Increase 10/300 GL column with 50 mM NaCl, 0.1 M phosphate buffer pH 7.45 as the running buffer. The eluted samples were collected and stored at 4 °C.

The protein concentration was assessed by measuring absorbance at 280 nm with a NanoDrop spectrophotometer (Thermo Fisher Scientific). Molecular weights and extinction coefficients were derived from amino acid sequences using the ProtParam tool (see Table 4.14).

Table 4.14. Extinction coefficient and molecular weight (M_w) calculated using ProtParam tool.

Variant	Extinction coef. ($M^{-1} \cdot cm^{-1}$)	M_w ($g \cdot mol^{-1}$)
A5	30370	23957
A11	37360	25177
B8	31860	26411
C5	42400	32384
C8	47900	29623
C11	30370	30512

4.3.4.7 Kinetic activity essay

The kinetic activity was tested by incubating purified protein samples with fluorogenic substrates and measuring fluorescence.[215] Phome ([cyano-(6-methoxynaphthalen-2-yl)methyl] 2-(3-phenyloxiran-2-yl)acetate) (Sigma Aldrich) was used as the substrate. Kinetic screens were performed in 50 µL reaction volumes using 96-well half-area plates (Corning 3650). The generation of the fluorogenic product from Phome hydrolysis was monitored continuously on the Synergy Neo2 plate reader (BioTek, Inc) at 30°C for 1 hour, with excitation at 330 nm and emission at 465 nm.

Protein and substrate solutions were prepared in sodium phosphate buffer (50 mM NaCl, pH 7.4) containing 10% DMSO. Each well of the microtiter plate contained 25 µL of enzyme solution, and reactions were initiated by adding 25 µL of a 200 µM substrate solution.

4.4 MPNN redesign of the gpASNase1 native backbone

This chapter presents the findings obtained during the second part of my research stay at the Baker Lab, at the Institute for Protein Design within the Department of Biochemistry at the University of Washington (Seattle, Washington, USA). The work was conducted in collaboration with dr. Susana Vázquez-Torres, dr. Ljubica Mihaljević, Kiera H. Sumida, dr. Arvind S. Pillai and Prof. David Baker, whose contributions and instructions were essential. The results outlined in this chapter reflect the collective efforts of the team and provide insights gained through our collaborative investigations in which I was the responsible person.

4.4.1 Short introduction and context

The MPNN-based redesign of the native gpASNase1 backbone was applied to explore whether this approach could enhance the expressibility, stability and function of ASNases for ALL treatment. Both expressibility and stability have been shown to improve after redesign in previous studies, such as the work by Sumida et al., on myoglobin and tobacco etch virus (TEV) hydrolase.[168]

The gpASNase1 was chosen for the MPNN redesign because it is a mammalian protein, minimizing the risk of immunogenic responses, while also exhibiting high enzymatic activity and a favorable micromolar K_M.[28] Moreover, this enzyme has served as the starting point for a recent attempt for humanization through DNA shuffling.[29] However, one of the concerning finding was that only about 200 mg of gpASNase1 was expressed from 6 L of culture, despite very meticulous expression and purification procedures.[28] Additionally, the same authors observed that the activity of the protein frozen overnight at -80 °C was lower than that of fresh protein, highlighting some potential stability issues.[29] Given these limitations, we aimed to determine whether MPNN could improve the expressibility and stability of gpASNase1 while maintaining the kinetical properties of this enzyme. Finally, the redesign of the native gpASNase1 was particularly interesting due to the relatively large homotetrameric structure of gpASNase1 compared to the proteins typically redesigned with MPNN, making it an interesting test case for the applicability of the method to complex

oligomeric systems. Therefore, we wanted also to investigate if mutations introduced by MPNN could alter this oligomeric state and whether they would be as successful as those previously observed in the much smaller TEV and myoglobin cases.[168]

4.4.2 Results

ProteinMPNN design methodology was applied to the N-terminal domain (first 362 amino acid residues) of the native gpASNase (gN). The structure was taken from PDB code 4R8K [28]) and was relaxed during 1 μs classical MD simulation (see 4.2.9.1 System Preparation section). The C-terminal domain of the gpASNase1 (gC) was swapped to that of hASNase1 (hC, with UniProt code: G3V1Y8 [223]) in all the cases (including the native one), as done by Lavie *at al.* in the preparation of the humanized chimeras [29]. These authors showed that the gN-hC version exhibits kinetical properties, within the margin of error, comparable to those of gN-gC (full native gpASNase1).[29] Furthermore, their results also suggested that the C-terminal truncated version of gpASNase1 exhibited the same kinetic properties as the full-length protein, but significantly more unstable, concluding that the C-terminal region likely contributes to protein stability through protein-protein interactions.[29] Therefore in all the cases presented here, the C-terminal domain was not changed and was taken from hASNase1 and only the N-terminal domain of gpASNase1 undergone the MPNN redesign process. In the following text, the term "native system" will be used to refer to either the system with both a native N-terminal and C-terminal domains, or to the system with native N-terminal domain from gpASNase1 and native C-terminal domain derived from hASNase1 (gN-hC).

During the design process active site residues (residues with C_α atoms within 6 Å of the substrate) were preserved to ensure the conservation of the catalytic properties in the redesigns (see blue color in Figure **4.50**a and Figure **4.50**b). Additionally, the design space was also selected to preserve those amino acids that were found to be highly conserved across multiple sequence alignments (MSAs), selecting the top 50% and 70% of the most conserved sites. Therefore, two distinct design sets were created. The first set (m1) had as designable all the

residues within the N-terminal domain of gpASNase1 except the active site residues and the top 50% of the most conserved residues (light green color on the Figure **4.50**a). The second set of designs (m2) had protected all the residues in the active site residues and the top 70% of the most conserved residues within the MSA (light green color on the Figure **4.50**b). The full list of the protected residues is given in the Technical Details section.

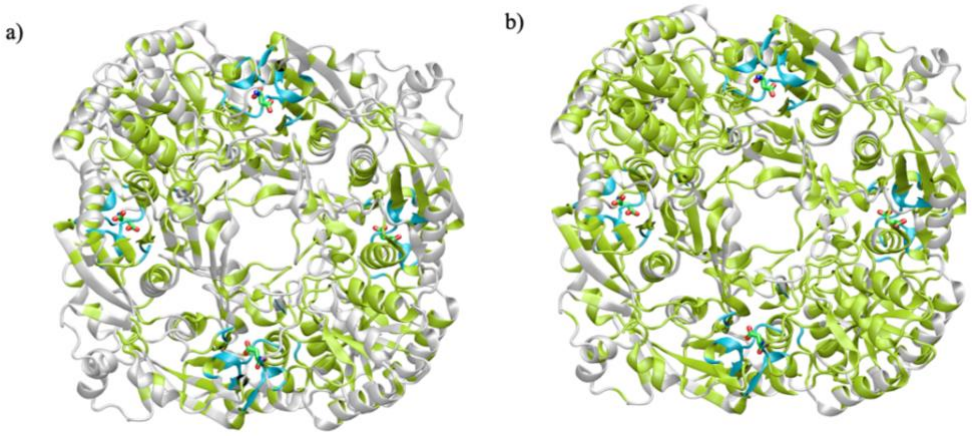


Figure 4.50. Positions adjacent to the ligand were kept fixed during sequence design (shown in blue). The top a) 50% residues preserved in MSA (design set m1) shown in green color and b) 70 % residues preserved in MSA (design set m2) represented in green color). Non-conserved regions (in white) underwent backbone remodeling.

Using ProteinMPNN, we generated 96 novel sequences and analyzed their potential to replicate the gpASNase1 backbone conformation via AlphaFold multimer predictions.[198] AlphaFold Multimer was used with 20 refinement cycles and three different models for each structure prediction. Both pLDDT and RMSD of C_α atoms with respect to the native structure were tracked for all designs. In the Figure **4.51**, a plot of pLDDT versus RMSD is shown for all 96 designs. The blue dots represent the m1 design set, while the orange dots represent the m2 design set.

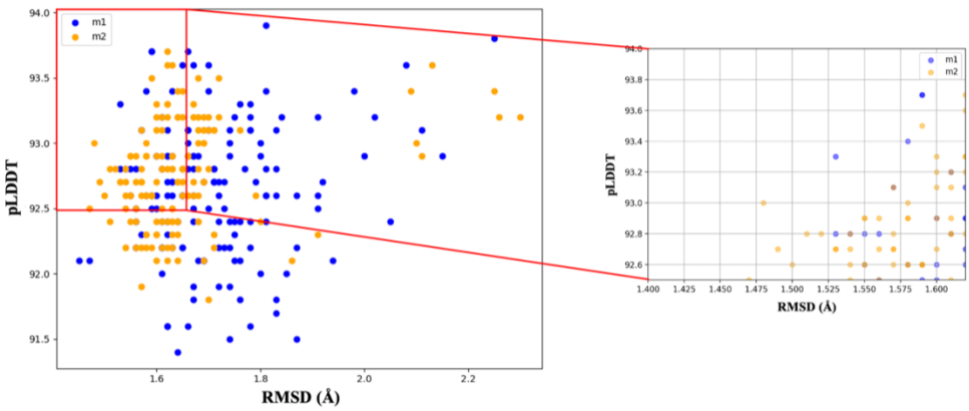


Figure 4.51. The pLDDT vs RMSD of C_α with respect to the native structure for the 96 designs generated with ProteinMPNN. An insight of the best candidates based in these scores is shown on the right. Orange dots represent the design set m1 and blue dots represent design set m2.

Eleven designs exhibited very strong structural alignment with the native structure (pLDDT > 93.0 and Cα RMSD < 1.5 Å) (right upper corner of Figure **4.51**). Amino acid sequences of these designs are given in the Technical Details section. These designs showed excellent correspondence within the ligand binding region and were therefore chosen for experimental assessment.

Genes encoding each protein were synthesized as E-Blocks and assembled into plasmids using Golden Gate (GG) cloning. The resulting plasmids were transformed into E. coli BL21 (DE3) competent cells, which were then used to inoculate starter cultures. These cultures were diluted 100-fold into 2YT medium containing the appropriate antibiotic and incubated at 37°C with shaking until the optical density (OD) at 600 nm reached 0.6–0.8, at which point protein expression was induced. The cultures continued incubating overnight, after which the cells were harvested by centrifugation. The resulting pellets were resuspended and lysed via sonication. The soluble fraction was separated by centrifugation, and the supernatant was purified using affinity chromatography with a HisTrap Ni^{2+}-NTA resin column. Further details are provided in the

Technical Details section. The eluted protein samples were filtered and subjected to size-exclusion chromatography (SEC), where they eluted as single peaks (Figure 4.52a).

In Figure 4.52a, SEC peaks of the native enzyme together with A8 and A11 designs are shown, while on the right the expression yield of each design is plotted versus the sequence similarity with the native enzyme. The yield was calculated as the concentration of the protein in each eluted sample using nanodrop at 220 nm and the theoretically obtained molecular weight and extinction coefficients.

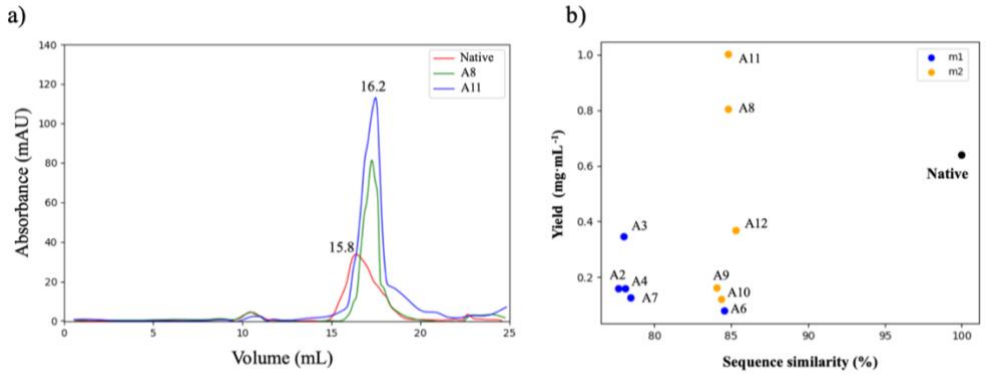


Figure 4.52. a) SEC traces of the designed ASNase variants (blue lines) and the native gpASNase1-hC enzyme (red line) and b) Expression yield (mg L^{-1}) of the designed variants and the native gpASNase1 with respect to the sequence similarity with the native gpASNase1. Sequence similarity has been calculated only over the N-terminal domain.

Theoretically predicted molecular weights and extinction coefficients of these designs along with the native sequence, obtained from the amino acid sequences using the ProtParam tool [224], are given in the Technical Details section. As seen from Figure 4.52a, both the native sequence and the designs were expressed successfully and eluted as a single peak on SEC. For simplicity reasons, SEC peaks of the other designs are not shown here. When it comes to expressability, decreased evolutionary constraints correlate with higher soluble expression levels, as previously noticed by Sumida *et al.* [168] The expression yield of two variants (A8 and A11) was measured to be almost twice as higher as that of the native enzyme (see Figure 4.52b). However, the SEC peak positions of these

designs were noticed to be shifted compared to the native protein (see Figure 4.52a). It was, however, unclear whether this shift definitively indicated a difference in molecular mass.

To clarify this, we conducted a mass photometry (MP) experiment on the two designs that exhibited higher expression levels than the native protein (designs A8 and A11). Results are presented in Figure 4.53.

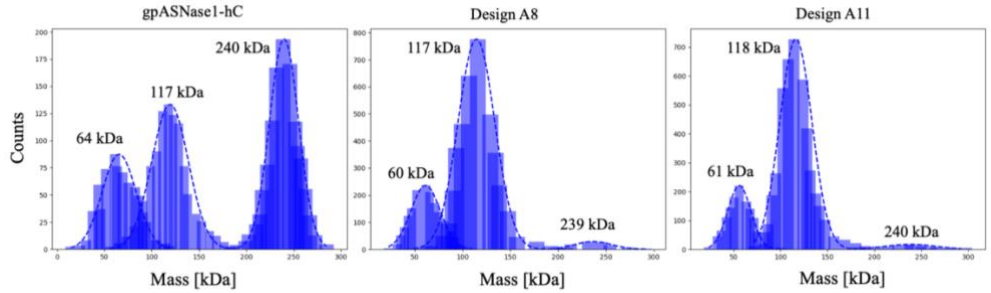


Figure 4.53. Mass photometry spectra of the native enzyme and the two most expressible variants A8 and A11.

As shown in Figure 4.53, the native enzyme primarily exhibits a peak at ~240 kDa, consistent with a tetrameric assembly. In contrast, the designed variants show peaks around ~117 kDa, indicating that they predominantly exist as dimers rather than forming tetramers as the native protein does. While it was unclear if the tetrameric form is needed to have an active form of the enzyme, we questioned whether the designs might still be capable of forming tetramers at higher concentrations, given that mass photometry requires sample dilution to prevent overcrowding in the field of view. We run a native gel on the native enzyme and the two most expressible designs (A8 and A11) across a range of concentrations (1500 nM, 1000 nM, 500 nM, 200 nM, 100 nM, 10 nM, and 5 nM). More details are given in the Technical Details section. The native gel results for the two designs and the native enzyme are presented in Figure 4.54.

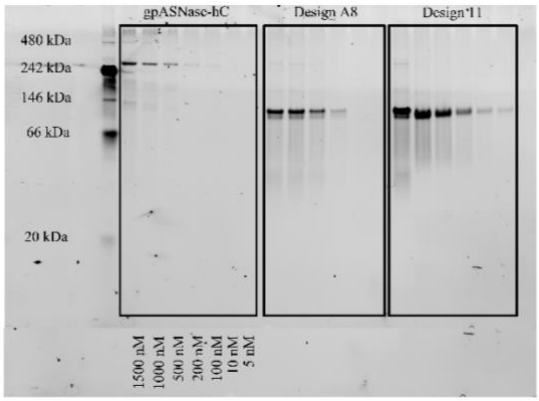


Figure 4.54. Native gel of the native gpASNase1-hC and two designed variants across a range of concentrations. For each sample, the first lane corresponds to the highest concentration, with subsequent lanes representing decreasing concentrations, as explicitly labeled for the native gpASNase1-hC. The tested concentrations were 1500 nM, 1000 nM, 500 nM, 200 nM, 100 nM, 10 nM, and 5 nM. A molecular weight ladder is included on the left to facilitate molecular weight estimation.

As expected, at higher concentrations, the bands are most prominent in the first lanes for each variant (see Figure 4.54). As the sample concentration decreases, the bands gradually become less visible, eventually disappearing at the lowest concentrations due to the detection limits of the gel. Most importantly, across all concentrations tested, up to 1500 nM, the bands corresponding to the two designs consistently migrate to a position slightly below the 146 kDa marker, likely around 120 kDa. This strongly suggests that the designed variants predominantly exist as dimers rather than tetramers, even at high concentrations. We also performed gel tests for the rest of the designs (data not presented here) all of which yielded results like the A8 and A11. In contrast, the native enzyme bands consistently migrate to a position corresponding to approximately 240 kDa, even at the lowest concentration tested until reaching to the detection limit, confirming that it remains in its tetrameric state, serving as a solid control for the experiment.

Although it was clear at this point that the designed variants existed as dimers, we sought to further characterize their stability and assess their denaturation profiles, with the expectation that they might exhibit greater stability than the native enzyme. Given that the active sites are located within each intimate dimer,

the designs could still retain enzymatic activity despite their dimeric state. To evaluate their thermal stability, we conducted circular dichroism (CD) spectroscopy as described in the Technical Details section. The CD spectra were recorded from 200 to 260 nm across a temperature range of 25 to 95 °C, with a time step of 10 °C. The results are presented in Figure 4.55.

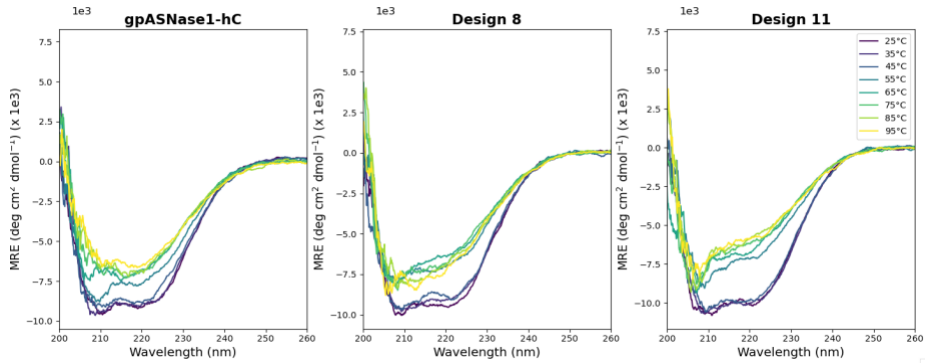


Figure 4.55. CD spectroscopy signal of native gpASNase1 and two design variants A8 and A11 over a temperature gradient from 25ºC to 95ºC. CD signal reported in molar residue ellipticity (MRE).

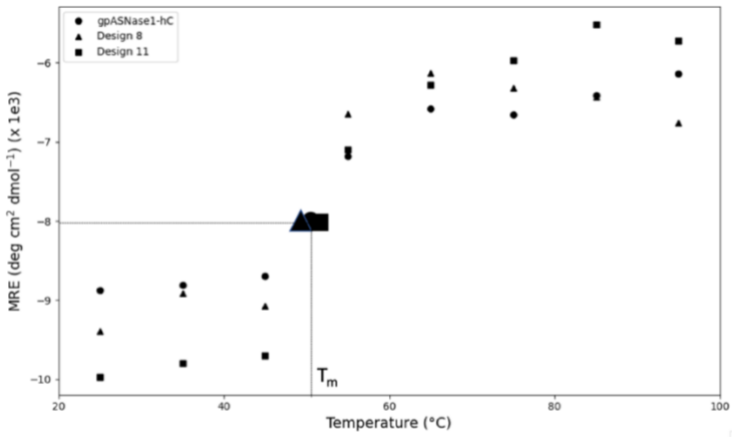


Figure 4.56. CD melting temperature plots of the two designed variants (A8 and A11) compared to native gpASNase1-hC. Signal reported in molar residue ellipticity (MRE).

By taking the molar residue ellipticity (MRE) at 222 nm from Figure 4.55, melting curves were constructed for the two designs and the native probe (see Figure 4.56). The melting temperatures (T_m) were determined as the inflection point of the sigmoidal Boltzmann curve using Prism10 Software.[225] The native gpASNase1-hC had a T_m of 52.6 °C, while the two designs exhibited the following melting temperatures: T_m (A8) = 51.1°C and T_m (A11) = 52.6°C. These results indicate that the stability of the two designs is comparable to that of the native protein in spite of not forming tetramers.

Urea denaturation curves were obtained by gradually increasing the urea concentration, which leads to protein unfolding. These curves provide more information than melting curves, as they allow for a more controlled assessment of protein unfolding, enabling the analysis of unfolding intermediates, cooperativity, and thermodynamic parameters. Most importantly, urea denaturation curves allow the determination of the folding Gibbs free energy (ΔG_f) as a better measurement of the thermodynamic stability of the protein. Results are presented in Figure 4.57.

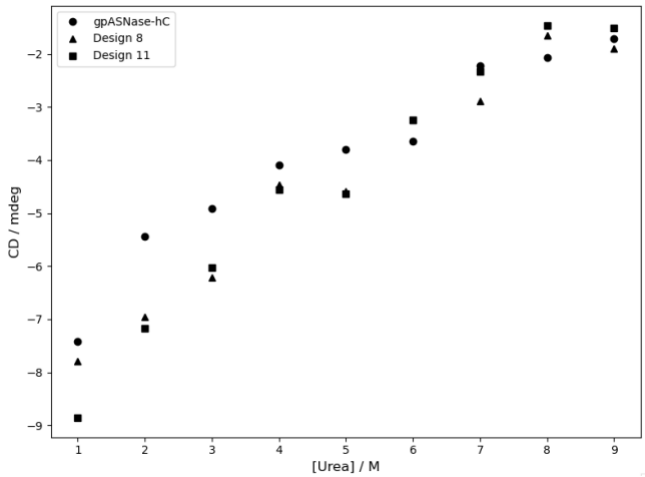


Figure 4.57. Urea denaturation curves of the native enzyme and the two most expressible designs. The CD signal is given in the mili-degrees.

From the data given in Figure 4.57, the free energy of folding (ΔG_f) was determined extrapolating the plot $RT\ln(f_D/f_N)$ to zeroth concentration of

urea.[226] f_D and f_N are the fractions of the protein in the denatured (unfolded) and native (folded) states, respectively. The Gibbs free energy of folding of the native enzyme was estimated to be -1.2 kcal/mol. The free energy of folding for two most expressible variants are ΔG_f (A8) = -1.0 kcal/mol, and ΔG_f (A11) = -1.3 kcal/mol. Then no significant differences in stability were observed, even if the native protein is present as a tetramer and the designs as dimers.

To determine the kinetic properties in the designs, we performed a kinetic assay using Nessler's reaction. In the ASNase reaction with L-Asn ammonia is being produced. This ammonia forms a complex with the Nessler's reagent with a dark orange color that allows it to be measured spectrophotometrically. We therefore assessed the activity by measuring the concentration of ammonia produced after adding L-Asn to each sample. The samples were incubated at 37°C for 1 hour. After this period trichloroacetic acid was added to terminate the reaction. Further technical details are provided in the Technical Details section. Results are shown in Figure 4.58, indicating the amount of the ammonia produced in 1 hour by the native enzyme and designed variants at 37ºC and pH=7.5.

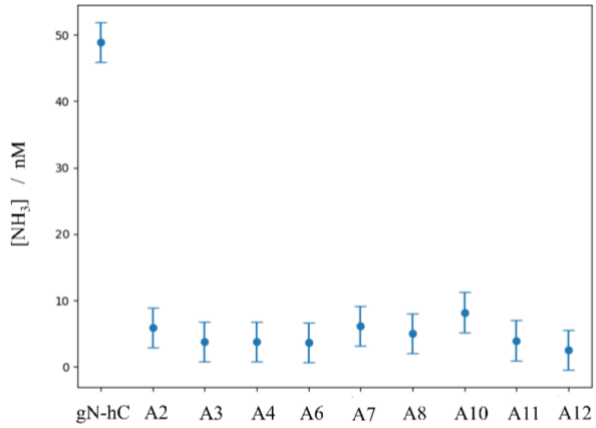


Figure 4.58. The amount of ammonia produced after 1h at 37ºC and pH=7.5 by the native enzyme and the designed variants. In all cases the initial concentration of substrate was $[L\text{-}Asn]_0$=50 nM, while the protein concentration was [E]=5 nM.

As observed, in contrast to the native enzyme, the activity of all the designed variants falls almost within the detection error. We hypothesized that the absence of ASNase activity is due the inability of the designs to form tetramers. This

raised the question of why the intimate dimer, despite having the active site positioned well within the intimate dimer, showed no significant reactivity. To solve this question, we ran MD simulations, which are discussed in the following subsection.

4.4.3 The importance of the Quaternary Structure of the gpASNase1

To investigate the stability of the designed protein A11, molecular dynamics simulations were run on the hypothetical tetrameric form, following the same protocol explained in the previous chapter (section System Preparation of the chapter 4.2.9.1). The initial tetrameric structure was obtained using AlphaFold2. In total, 1 μs MD simulation was run. We then computed the interaction energies between individual monomeric chains. The results of these calculations are presented in Figure 4.59.

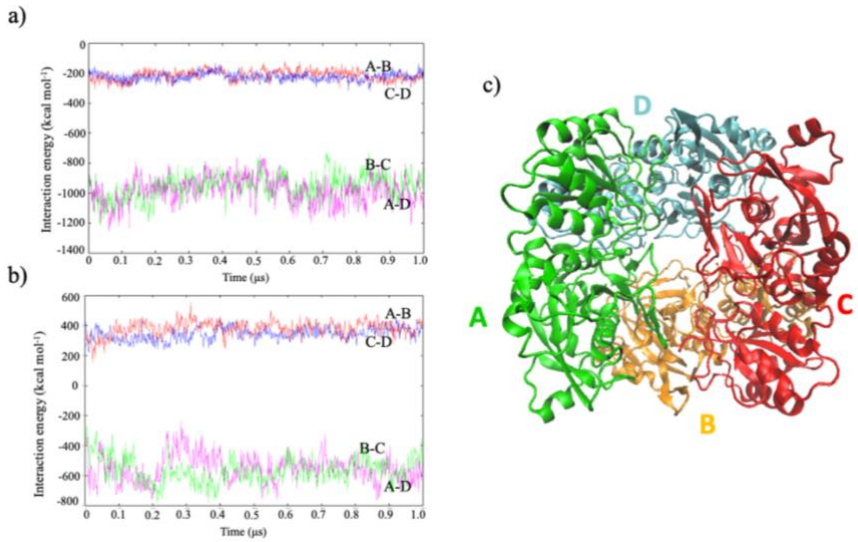


Figure 4.59. Interaction energies (Electrostatic and Van der Waals, in kcal mol^{-1}) between the protomers in a) Native gpASNase1 and b) A11 design. Blue and red color correspond to the interaction energies between two adjacent protomers that do not form intimate dimer and green and pink colors to the interaction energies between the protomers in the intimate dimer. c) The protomers assignation A, B, C and D of both the variant and gpASNase1.

As shown in Figure 4.59, the interaction energies between the two protomers forming the intimate dimer (B-C and A-D) and between the non-intimate dimer protomers (A-B and C-D) are less favorable for the design A11 compared to the native enzyme (compare Figure 4.59a and b). This suggests that, since the residues at the protein-protein interface were not explicitly preserved during MPNN redesign, some interface positions that have undergone mutations introduced unfavorable interactions, preventing tetramer formation. Furthermore, the radius of gyration of the A11 design also shows a continuous increase during the 1 μs MD simulation, suggesting a separation of the non-intimate dimers to form two dimers (see Figure 4.60). These results show that MD simulations can provide some insights into the oligomeric state of the protein.

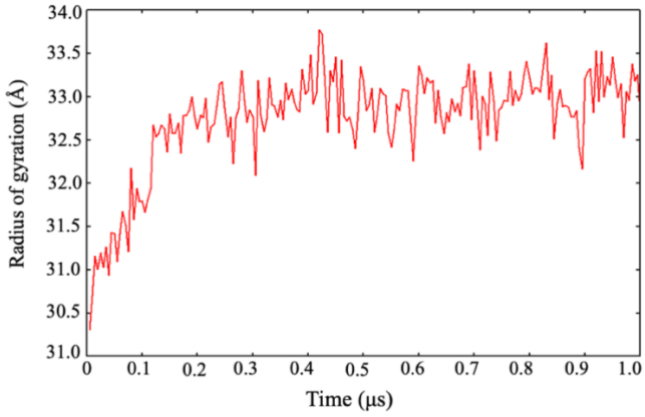


Figure 4.60. Time evolution of the radius of gyration during the MD simulation of the design A11.

For the most highly expressed variants, we analyzed the residues placed at the interfaces between protomers and examined how they were mutated by MPNN (see Table 4.15). Notably, some interface residues were mutated to residues with the opposite charge or with a significantly different size. For example, in design A8, the mutation at position 276 from Lys to Asp introduces a negative charge where a positive one existed, disrupting key electrostatic interactions crucial for stabilizing the tetramer. Similarly, the mutation at position 251 from Leu to Arg introduces a bulky, positively charged residue in the place of a smaller,

hydrophobic one, which result in steric clashes or disruption of the hydrophobic core, further destabilizing the structure. In design 11, the mutation at position 192 from Gln to Phe replaces a polar, hydrogen-bonding residue with a bulky, non-polar aromatic side chain, interfering essential interactions and introducing steric hindrance. These mutations, while aimed at optimizing certain properties, can inadvertently compromise the stability and correct assembly of the tetramer. A promising approach would therefore be to generate new designs while preserving also the residues found at the interfaces between protomers within the intimate and non-intimate dimers. This strategy is currently being explored in ongoing work. We aim to order genes that encode sequences of the previously mentioned variants A8 and A11, except for the interface mutations, which we plan to revert to the residues found in the native gpASNase1. Additionally, we plan to design a set of new variants from the very start, with the residues at both interfaces (the intimate dimer and the other dimer) protected and not allowed to undergo mutation.

Even though these results offer insight on the oligomeric form of the design variants, it still does not clarify why the dimeric form is not catalytically active, given that the active site is located deep within the intimate dimer (see Figure 4.59c). To address this, a classical MD simulation of the dimeric form of the gpASNase1 was run. The Michalis complex system was prepared as described in the previous chapter, section 4.2.9.1. The analysis of active-site distances from MD simulations of both the dimeric and tetrameric forms of gpASNase1 reveal some notable differences. Specifically, the dimer predominantly adopts a nonreactive conformation during the free molecular dynamic simulation in comparison to the tetramer. For instance, in the case of the MD simulation of the dimer, the Thr116Hγ–AsnNδ distance exhibits a bimodal distribution. This distance is associated to the proton transfer from Thr116 to the NH_2 leaving group of the substrate that must form ammonia after formation of the acyl enzyme complex. The observed increase in this distance could therefore lead to a higher reaction barrier and thereby contribute to the lack of catalytic activity in the dimeric form. This structural change observed in the active site of the dimer can be connected to the absence of the second dimer. Thr116 is kept in its catalytically active conformation thanks to a hydrogen bond interaction with

Lys188, which in turn is hydrogen bonded to Gln143, a residue belonging to a flexible loop (see Figure 4.61c).

Table 4.15. Residues at the interfaces of the intimate dimer and non-intimate dimer in the native gpASNase1-hC enzyme and in designs A8 and A11.

Intimate dimer interface			
Residue	Native gpASNase1	Design A8	Design A11
89	Thr	Thr	Thr
189	Val	Thr	Val
192	Gln	Glu	Phe
251	Leu	Arg	Gln
275	Ser	Ser	Ser
276	Lys	Asp	Asn
303	Ser	Lys	Ser
Non-intimate dimer interface			
41	Thr	Glu	Ser
65	Pro	Pro	Pro
67	Ser	Ala	Ala
148	Val	Leu	Val
155	Glu	Glu	Glu
192	Gln	Glu	Phe
193	Lys	Leu	Leu
210	Ala	Ala	Ala
211	Asp	Glu	Phe

The conformational change of Gln143 is linked to a shift of the entire loop further away from the active site in the dimeric form, as observed in Figure 4.61b. Specifically, in dimeric simulations, the absence of the adjacent protomer, which would otherwise impose steric constraints, results in an increased flexibility of the Gln143-containing loop. This increased mobility is evident from the RMSF analysis, which shows larger fluctuations in this region in the dimeric form compared to the tetrameric system (see Figure **4.62**a).

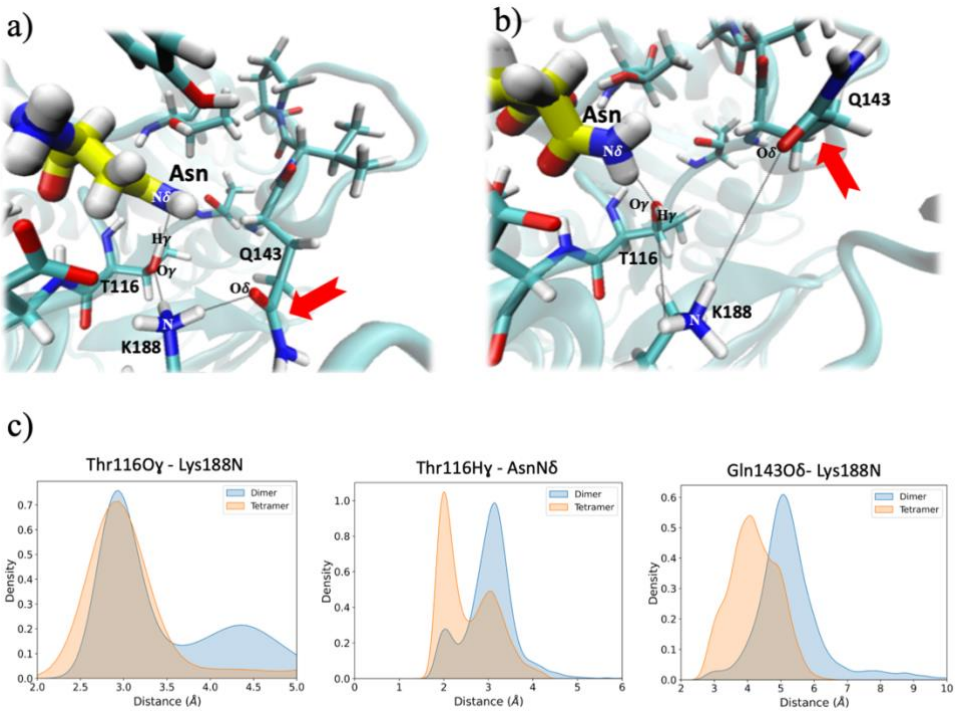


Figure 4.61. Structure of the active site of the gpASNase1 in the simulation of the tetramer (a) and dimer (b). In the Figure c distribution of the important distances marked on the a and b panels are given.

In the absence of the adjacent protomer (e.g., protomer C in Figure **4.62**b), the Gln143-containing loop shifts away from the active site, disrupting the catalytic arrangement. This displacement further favors the population of a nonreactive conformation (Figure 4.61b). These findings provide strong evidence that the dimeric form of gpASNase1 is catalytically inactive, despite the active site residues being fully contained within the intimate dimer. To the best of our knowledge, this is the first insight into the relationship between the oligomeric state of type 1 ASNase and its catalytic activity.

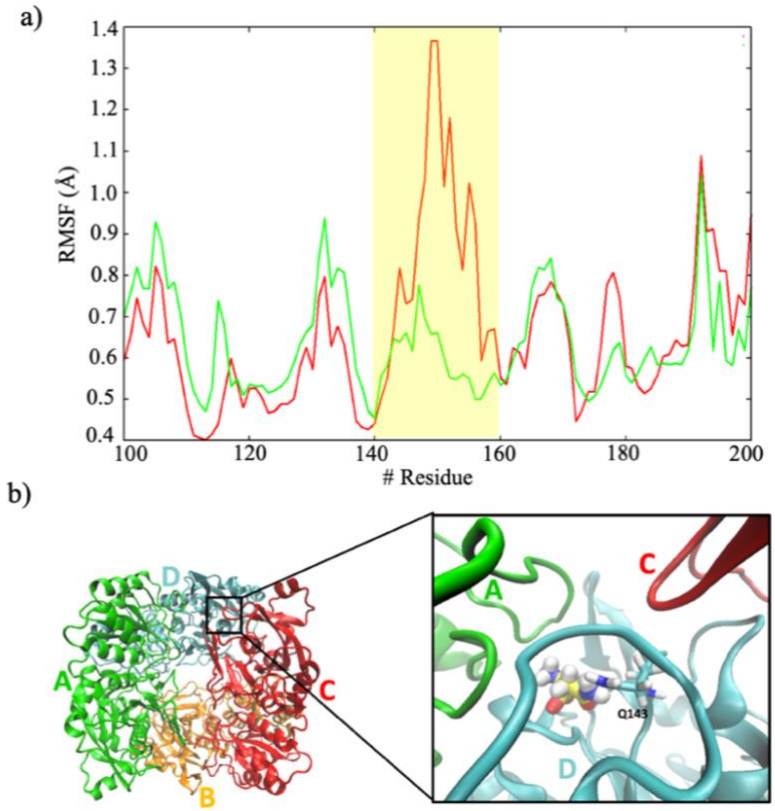


Figure 4.62. a) Root mean square fluctuation (RMSF) of C_α atoms of the protein residues obtained from the dimeric gpASNase1 (red line) and tetrameric (green line) molecular dynamics simulations. The the yellow rectangle highlighs the Gln143-containing loop region. b) Structural depiction of the steric limitations on the movement of Gln143 in the tetrameric gpASNase1 structure.

4.4.4 Short summary of the MPNN redesign of the gpASNase1 native backbone

The redesign of gpASNase1 provided insights into its active form and, more broadly, the active form of type 1 ASNases, even though no catalytically active variants have been successfully designed. Since the interface residues were not preserved, the redesigned variants formed dimeric structures instead of tetramers. MD simulations of dimeric gpASNase1 offered a theoretical basis for

understanding the experimental observation that these variants lack enzymatic activity. Analysis of these simulations revealed that the absence of the adjacent protomer causes the Gln143-containing loop to shift away from the active site, disrupting the catalytic arrangement and promoting a nonreactive conformation.

4.4.5 Technical Details

4.4.5.1 ProteinMPNN design technical details

ProteinMPNN design methodology was employed only on the N-terminal domain while the C-terminal domain of the gpASNase1 (gC) was swapped to that of hASNase1 (hC). The pipeline followed was the same as in Sumida et al.[168] During the design process active site residues (residues with C_α within 6 Å of the substrate) were preserved to ensure the conservation of the catalytic properties. These included residues 18, 19, 20, 22, 83, 84, 85, 86, 114, 115, 116, 117, 142, 143, 144, 188, 272, 308, 310 within all four chains A, B, C, D of the homotetrameric gpASNase1. Additionally, the design space is also selected to preserve those amino acids that were found to be highly conserved across multiple sequence alignments (MSAs). To create the MSA, four rounds of HHblits searches [199] were carried out against the UniRef30 database (accessed on July 15, 2024) using E-value thresholds of 1e-50, 1e-30, 1e-10, and 1e-4. For each position in the sequence alignment, the occurrence frequency of each amino acid was calculated and the most conserved amino acid at each site were identified. We then filtered each position based on the conservation level of the most frequent amino acid and selected approximately the top 50% and 70% of the most conserved sites.

Each set was run through MPNN at three different temperatures: 0.1, 0.2, and 0.3, with 16 sequences generated per temperature. Temperature variations were used to achieve greater diversity of the generated protein sequences, as higher temperatures can promote more exploration of sequence space, leading to greater variability in the designs.

The list of 50% of most preserved residues in MSA protected in the m1 design set is: 1, 2, 3, 5, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 27, 29, 31, 35, 39, 43,

44, 45, 61, 64, 71, 72, 78, 81, 82, 83, 84, 85, 86, 87, 88, 90, 93, 96, 97, 100, 104, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 150, 151, 152, 153, 156, 157, 160, 161, 164, 165, 169, 171, 172, 174, 176, 180, 182, 183, 184, 185, 188, 191, 194, 196, 197, 199, 200, 201, 203, 204, 205, 206, 209, 212, 228, 238, 241, 242, 243, 244, 245, 246, 247, 256, 261, 262, 264, 265, 266, 267, 268, 269, 271, 272, 273, 274, 279, 280, 283, 286, 290, 293, 294, 295, 297, 298, 299, 302, 304, 308, 310, 311, 312, 315, 316, 317, 318, 319, 320, 322, 323, 324, 326, 327, 328, 329, 330, 331, 332, 333, 335, 336, 337, 341, 350, 354, 355, 356, 357, 359.

The list of 70% of most preserved residues in MSA protected in the m1 design set is: 1, 2, 3, 4, 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 27, 28, 29, 31, 35, 38, 39, 43, 44, 45, 48, 49, 57, 61, 64, 65, 66, 70, 71, 72, 74, 78, 79, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 92, 93, 96, 97, 100, 103, 104, 105, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 149, 150, 151, 152, 153, 155, 156, 157, 158, 159, 160, 161, 163, 164, 165, 169, 170, 171, 172, 173, 174, 175, 176, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 191, 194, 195, 196, 197, 199, 200, 201, 202, 203, 204, 205, 206, 207, 209, 210, 212, 215, 219, 228, 230, 231, 234, 235, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 252, 254, 256, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 279, 280, 283, 284, 286, 289, 290, 291, 292, 293, 294, 296, 297, 298, 299, 300, 302, 304, 307, 308, 309, 310, 311, 312, 315, 316, 317, 318, 319, 320, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 341, 342, 344, 345, 346, 347, 349, 350, 352, 353, 354, 355, 356, 357, 358, 359.

Design A1 (native gpASNase1): MARASGSERHLLLIYTGGTLGMQS
KGGVLVPGPGLVTLLRTLPMFHDKEFAQAQGLPDHALALPPASHGPRV
LYTVLECQPLLDSSDMTIDDWIRIAKIIERHYEQYQGFVVIHGTDTMASG
ASMLSFMLENLHKPVILTGAQVPIRVLWNDARENLLGALLVAGQYIIPE
VCLFMNSQLFRGNRVTKVDSQKFEAFCSPNLSPLATVGADVTIAWDLV
RKVKWKDPLVVHSNMEHDVALLRLYPGIPASLVRAFLQPPLKGVVLET

FGSGNGPSKPDLLQELRAAAQRGLIMVNCSQCLRGSVTPGYATSLAGA
NIVSGLDMTSEAALAKLSYVLGLPELSLERRQELLAKDLRGEMTLPT

C-terminal sequence of all the designs (hC): SVEERRPSLQGNTLGG
GVSWLLSLSGSQEADALRNALVPSLACAAAHAGDVEALQALVELGSD
LGLVDFNGQTPLHAAARGGHTEAVTMLLQRGVDVNTRDTDGFSPLLL
AVRGRHPGVIGLLREAGASLSTQELEEAGTELCRQRQPGTWQWWPFYR
AWRVRLVPRPHAQKCCLVSNLKASCCSISHSFLP

Design A2 sequence: MARVSGEVRRLLLIYTGGTLGMARTGGLLSP
GGLRELLRELPMFYDKETAEELGLPEDELVLPRASAGPRVVYKVLEMQ
PLLDSSDMTIEDWKRIASLIAENYDKYQGFVVIHGTDTMASGASMLSF
MLENLNKPVILTGAQVPIWELWNDARSNLLGALLIAGQFNIPEVALFMN
NKLYRGNRVTKVDSTSFDAFESPNESPLATVGRDVSINWDLVKEVKED
KPLVVHTDMDSDVALLRLYPGIPAEQVRAFLQPPLKGVVLETFGSGNG
PTNPELLAALREAAERGLIMVNLSQCLRGSVSPGYTTSLSGANIVSGRD
MTSEAALAKLSYVLGLPGLSLEERLRLLAEPLRGERTLPS

Design A3 sequence: MARVSGEVRRLLLIYTGGTLGMARSGGLLSP
AGGLKELLRELPMFYDKERAEELGLPDDELVLPEAAAGPRVVYKVLEL
QPLLDSSDMDIEDWIRIAEVIAENYDKYQGFVVIHGTDTMASGASMLSF
MLENLDKPVILTGAQVPIWELWNDARENLLGALLIAGQFRIPEVALFMN
WRLYRGNRVTKVNSTSFDAFISPNESPLATVGADVSINWDLVREVKSG
KKLVVHSDMDTDVALLRLYPGIPAEQVRAFLQPPLKGVVLETFGSGNG
PTDPELLAALREAAERGLIMVNTSQCLKGSVSPGYTTSLSGANIVSGRD
MTSEAALAKLSYVLGLPGLSLEERLKLLSEPLRGERTLSS

Design A4 sequence: MARVSGDTVRLLLIYTGGTLGMARSGGLLTP
AGGLRELLRELPMFYDKETAEELGLPEDELVLPRASSGPRVVYKVLEM
TPLLDSSDMDIEDWIRIAKLIADNYDKYQGFVVIHGTDTMASGASMLSF
MLENLDKPVILTGAQVPIWELWNDARSNLLGALLIAGQFRIPEVALFMN
DKLYRGNRVTKTDSTSFDAFESPNESPLATVGRDVTINWDLVREVKEG
KPLVVHTDMDTDVALLRLYPGIPAEQVRAFLQPPLKGVVLETFGSGNG

PTDPELLAALREAAERGLIMVNTSQCSRGSVSPGYTTSLSGANIVSGRD
MTSEAALAKLSYVLGLPGLSLEERLRLLSSPLRGESTSSS

Design A5 sequence: MARVSGDVVRLLLIYTGGTLGMSTSGGVLSP
SGGLVELLRELPMFHDRETAEELGLPEDELVLPRAAAGPRVVYKVLELE
PLLDSSDMTIEDWIRIANVIAENYDRYQGFVVIHGTDTMASGASMLSFM
LENLDKPVILTGAQVPIWVLWNDARQNLLGALLIAGQYRIPEVALFMN
WKLYRGNRVQKVDSESFDAFESPNESPLATVGADVSINWGLVRSVKSD
KKLTVHSNMDTDVALLRLYPGIPAERVRAFLRPPLKGVVLETFGSGNGP
TNPELLSALREAAERGKIMVNTSQCVKGKVSPGYSTSLDGANIVSGRD
MTSEAALAKLSYVLGLPNLSIEERLELLSKPLRGERTESS

Design A6 sequence: MARVSGTVRTLLLIYTGGTLGMTTTGGVLSP
SGGLEEYLRELPMFYDKEYAEELGLEEGELALPDLASGPRVVYRVLESR
PLLDSSDMDINDWIQIAELIAEHYDKYQGFVVIHGTDTMASGASMLSFM
LENLDKPVILTGAQVPIWVLWNDATSNLLGALAIAGQYNIPEVALFMN
DKLYRGNRVSKVDSNSFDAFTSPNESPLAIIGSDVFINWDKVAKVSGDK
PLVVRSDMDSDVALLRLYPGIPAHVVKRFLEPPLKGVVLETFGSGNGPT
NPELLETLREAAERGLIMVNTSQCEKGSVAPGYRTSLSGANIVSGRDMT
SEAALAKLSYVLGLPGLTLEERLKLLAEPLRGELTADE

Design A7 sequence: MARASGETRHLLLIYTGGTLGMSRSGGVLTP
SGGLRELLAELPMFNDKEFAKELGLPEDELALPPAAAGPRVVYKVLEC
RPLLDSSDMTIEDWKRIAELIAEHYEKYQGFVVIHGTDTMASGASMLSF
MLENLDKPVILTGAQVPIRELWNDARENLLGALLVAGQFNIPEVCLFM
NSQLFRGNRVTKVDSDLFEAFVSPNLSPLATVGADVSIAWDLVREVKK
NKKLVVHLEMEEDVALLRLYPGIPAEQVRAFLQPPLKGVVLETFGSGN
GPSDPELLAALREAAERGLIMVNCSQCLRGSVSPGYATSLSGANIVSGR
DMTSEAALAKLSYVLGLPDLSLERRQELLSKDLRGEMTLPR

Design A8 sequence: MARASGEVRHLLLIYTGGTLGMSTSGGVLTP
SGGLVELLRELPMFYDKEFAAELGLPEDELALPPAAAGPRVVYKVLEC
EPLLDSSDMTIEDWKRIAELIAEHYEEYQGFVVIHGTDTMASGASMLSF
MLENLDKPVILTGAQVPIRVLWNDARENLLGALLVAGQFNIPEVCLFM

NSQLFRGNRVTKVDSDLFEAFISPNLSPLATVGADVTIAWDLVREVKKG
EKLVVHLEMEEDVALLRLYPGIPAEQVRAFLQPPLKGVVLETFGSGNGP
SDPELLAALREAAERGLIMVNCSQCLRGSVSPGYATSLSGANIVSGRDM
TSEAALAKLSYVLGLPGLSLERRQELLSKDLRGEMTEDR

Design A9 sequence: MARASGDVRHLLLIYTGGTLGMATAGGVL
TPAGGLVELLRELPMFYDKEFAEELGLPENELALPPAAAGPRVVYTVLE
CQPLLDSSDMTIDDWKRIAELIAKHYEDYQGFVVIHGTDTMASGASML
SFMLENLDKPVILTGAQVPIRELWNDARENLLGALLVAGQFNIPEVCLF
MNSQLFRGNRVTKTDSHLFEAFISPNLSPLATVGADVTIAWDLVKEVK
DDEPLVVHSEMESDVALLRLYPGIPAELVRAFLQPPLKGVVLETFGSGN
GPSDPDLLETLREAAERGLIMVNCSQCLRGSVSPGYATSLSGANIVSGR
DMTSEAALAKLSYVLGLPDLSLERRQELLSKDLRGEMTEPR

Design A10 sequence: MARASGETRHLLLIYTGGTLGMSTSGGVL
TPSGGLEELLRELPMFYDKEYAKKLGLPENELALPPAAAGPRVVYTVLE
CQPLLDSSDMTIEDWIRIAKLIAEHYEKYQGFVVIHGTDTMASGASMLS
FMLENLDKPVILTGAQVPIRLLWNDARENLLGALLVAGQFNIPEVCLFM
NSQLFRGNRVTKTDSELFEAFTSPNLSPLATVGAEVSIAWDLVRKTRSD
KPLVVHSDMETDVALLRLYPGIPAERVRAFLQPPLKGVVLETFGSGNGP
SDPELLDTLREAAERGLIMVNCSQCLRGKVQPGYATSLSGANIVSGRD
MTSEAALAKLSYVLGLPGLSLERRQELLSKDLRGEMTEPR

Design A11 sequence: MARASGDVRHLLLIYTGGTLGMSTSGGVLT
PSGGLRELLAELPMFHDKEYAKELGLPENELALPPASAGPRVVYTVLEC
TPLLDSSDMTIEDWKRIAELIAEHYEDYQGFVVIHGTDTMASGASMLSF
MLENLDKPVILTGAQVPIRELWNDARENLLGALLVAGQFNIPEVCLFM
NSQLFRGNRVTKVNSTLFEAFVSPNLSPLATVGATVSVAWDLVKEVKS
DKPLVVHTDMEEDVALLRLYPGIPAEQVRAFLQPPLKGVVLETFGSGN
GPSNPELLDTLREAAERGLIMVNCSQCLRGSVSPGYATSLSGANIVSGR
DMTSEAALAKLSYVLGLPGLSLERRQELLSKDLRGEMTEER

Design A12 sequence: MARASGETRHLLLIYTGGTLGMSTAGGVLT

PSGGLKELLAELPMFYDKEYADELGLPEDSLVLPPAEAGPRVVYTVLEC
RPLLDSSDMTIEDWKRIAELIAEHYEKYQGFVVIHGTDTMASGASMLSF
MLENLNKPVILTGAQVPIRTLWNDARENLLGALLVAGQFTIPEVCLFMN
SQLFRGNRVTKVDSSLFEAFISPNLSPLATVGASVDIAWDLVKSVESNTP
LVVHTDMEDDVALLRLYPGIPAEQVRAFLQPPLKGVVLETFGSGNGPS
DPELLAALREAAERGLIMVNCSQCLRGSVAPGYATSLSGANIVSGRDM
TSEAALAKLSYVLGLPELSLERRQELLSKDLRGEMTESR

Table 4.16. Reference names, monomeric extinction coefficients and molecular weight of the native sequence and eleven designs calculated using ProtParam tool.

Design	Name	Extinction coef. ($M^{-1} \cdot cm^{-1}$)	Mw ($g \cdot mol^{-1}$)
Native-hC	A1	53900	60801.87
m1_1	A2	55140	60903.48
m1_3	A3	60640	60859.45
m1_6	A4	55140	60761.14
m1_21	A5	60640	60882.41
m1_47	A6	59610	60672.96
m2_2	A7	46785	60798.48
m2_5	A8	48275	60619.25
m2_18	A9	48275	60646.10
m2_19	A10	49765	60825.50
m2_21	A11	48275	60532.99
m2_42	A12	49765	60501.97

4.4.5.2 Protein expression and purification

Genes encoding each protein variant were ordered as two separate fragments: a designed N-terminal region (1086 bp) and a fixed C-terminal region corresponding to the h_C domain of hASNase1 (576 bp). These fragments were ordered from Integrated DNA Technologies (IDT) as E-Blocks and joined into plasmids using Golden Gate (GG) assembly using the NdeI and BamHI restriction sites at the 5′ and 3′ ends, respectively. The corresponding genes were

subcloned into a Lm1371 vector containing N-terminal His6 tag. The assembled plasmids were subsequently inserted into E. coli BL21 (DE3) competent cells through heat shock transformation, by briefly exposing the cells to 42°C to facilitate plasmid uptake. The transformation reactions were used to inoculate starter cultures in 1 mL of super optimal broth (SOC) media and kanamycin. After shaking overnight at 37°C, the dense cultures were diluted 100-fold into 250 mL of 2YT medium with 100 μg/mL kanamycin. These cultures were incubated at 37°C while shaking, until the optical density (OD) reached 0.6–0.8 at 600 nm, at which point protein expression was induced by the addition of 0.1-0.15 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG). The cultures were then kept incubating overnight at 18°C. Next day, the cells were harvested by centrifugation for 10 min at 4000g at 4°C. The pellets were resuspended in lysis buffer (25 mm Tris-HCl, 400 mm KCl, 10 mm $MgCl_2$, 10 mm imidazole, 10% glycerol, 1% Triton X-100, 1 mm PMSF, pH 7.5) and disrupted by sonication using a Qsonica Q500 instrument with a four-pronged horn at 80% amplitude, using 10 second pulses followed by 20 second rest intervals, in order to prevent overheating, for a total of 5 minutes. Soluble fractions were centrifugated for 45 min at 14,000g at 4°C. The supernatant was purified by affinity chromatography passing it through HisTrap Ni^{2+}-NTA resin (Qiagen) column on a vacuum manifold. The column was first equilibrated with a buffer containing 25 mM Tris-HCl, 200 mM KCl, 10 mM $MgCl_2$, 30 mM Imidazole, pH 7.5. The second wash was done then with a buffer containing 25 mM Tris-HCl, 200 mM KCl, 10 mM $MgCl_2$, 50 mM Imidazole, pH 7.5. Finally, the column was washed with a buffer containing only 25 mM Tris-HCl pH 7.5 before eluted with 20 mM Tris-HCl, 500 mM Imidazole, pH 7.5.

Eluted protein samples were filtered and injected into an autosampler-equipped ÄKTA pure system on a Superose 6 10/300 column (Cytiva) at room temperature using SEC running buffer (20mM Tris-HCl, 100 mM KCl, 2 mM DDT, pH 7.5). The His6-tag was not removed during the purification process. The eluted samples were collected and concentrated using spin filters (3 kDa molecular weight cutoff; Amicon; Millipore Sigma) and stored at 4 °C.

The concentration of each protein was determined by measuring the absorbance at 280 nm using a NanoDrop spectrophotometer (Thermo Fisher Scientific) and using the molecular weights and extinction coefficients obtained from the amino acid sequences using the ProtParam tool.

4.4.5.3 Native gel and mass photometry

All mass photometry (MP) measurements were performed using a TwoMP (Refeyn) mass photometer at the room temperature (25ºC). Prior to measurement, samples were diluted to a desirable concentration to prevent overcrowding in the field of view caused by the too concentrated solutions. The laser was then oriented to the center of the sample well, and the camera was focused. Next, 10 μL of the sample was introduced into the droplet, and 1-minute videos were recorded using large field of view in AcquireMP. The masses distributions were processed using DiscoverMP.

For mass calibration, a 20 nM sample of Beta-amylase (BA) containing monomers (56 kDa), dimers (112 kDa), and tetramers (224 kDa) in equilibrium was used, allowing contrast values to be converted into mass values across tested designs. Expected masses were calculated by multiplying monomer masses by the number of subunits in different oligomeric configurations. The resulting distributions were fitted to the peaks to estimate observed oligomer masses and mass errors using the normfit function.

Given that in MP diluted protein samples were used, native polyacrylamide gel electrophoresis (PAGE) was used to assess whether protein oligomerization is influenced by protein concentration. The experiment was conducted under non-denaturing (native) conditions to allow for a rapid evaluation of molecular mass. Protein samples were diluted with the Native Sample Buffer for Protein Gels (BioRad) and loaded them onto Any kD™ Criterion™ TGX Stain-Free™ Protein gel (Bio-Rad). Electrophoresis was carried at 150 V for about 3 h at 4ºC and using 10x Tris/Glycine buffer for native gels (BioRad) containing 25 mM Tris, 192 mM glycine, pH 8.3 diluted with MiliQ water. Gels were imaged using a Chemidoc XRS+ (Bio-Rad). A molecular weight ladder NativeMark™

Unstained Protein Standard (ThermoDisher Scientific) containing markers of 20 kDa, 66 kDa, 146 kDa, 242 kDa, 480 kDa, 720 kDa, 1048 kDa and 1236 kDa was included in the first lane to estimate the molecular masses of the protein species.

4.4.5.4 Thermal stability and urea denaturation curves

The thermal stability of the designs was analyzed using circular dichroism (CD) in a Jasco J-1500 CD spectrometer, equipped with a Peltier system (EXOS) for temperature regulation. Measurements were conducted in quartz cells with a 0.1 cm optical path length, covering a wavelength range of 190–260 nm. The CD signal was expressed as molar ellipticity (θ). Thermal unfolding was monitored by observing changes in ellipticity at 222 nm as a function of temperature. Protein denaturation was induced by heating at a rate of 1°C per minute, from 20 to 95°C.

To obtain urea denaturation curves, the CD signal was measured at 222 nm for protein samples incubated with varying urea concentrations. Each sample contained the same protein concentration but was exposed to urea concentrations ranging from 0 to 9 M. After adding urea, the samples were incubated for 24 hours at 25°C to ensure equilibrium. Then the CD spectra of each sample were recorded to assess structural changes induced by urea, allowing for the analysis of protein unfolding as a function of denaturant concentration.

4.4.5.5 Kinetic activity essay

The activity of L-asparaginase was determined using a modified version of the Wriston method.[227] L-asparaginase catalyzes the conversion of L-asparagine to L-aspartic acid and ammonia, which reacts with Nessler's reagent (K_2HgI_4) to form an orange-colored product.

The enzyme assay mixture (60 mM L-asparagine in 50 mM Tris-HCl buffer (pH 7.5) and 20 mM of enzyme) was incubated at 37°C for 60 minutes, after which the reaction was stopped by adding 100 µL of 15% trichloroacetic acid (TCA). The mixture was then centrifuged at 10,000g for 5 minutes at 4°C to remove precipitates. The ammonia released into the supernatant was measured using a

colorimetric method by adding 100 µL of Nessler's reagent to 100 µL of supernatant and 800 µL of distilled water. The sample was vortexed and incubated at room temperature for 10 minutes, and the optical density (OD) with the NanoDrop spectrophotometer (Thermo Fisher Scientific) at 425 nm. Ammonia concentration in each sample was determined using a standard calibration curve prepared previously with the by known concentrations of $(NH_4)_2SO_4$.

Chapter 5: General conclusions and future goals

Each chapter of this doctoral thesis provides its own set of conclusions related to the problems addressed in each one. Therefore, in this chapter, I will rather focus on the general conclusions derived from my work and I will also outline potential future directions in ASNases modelling for leukaemia treatments.

From a methodological perspective, our results highlight the effectiveness of hybrid quantum mechanics/molecular mechanics molecular dynamics (QM/MM MD) simulations in the study of enzyme dynamics and reaction mechanisms. In particular, the Adaptive String Method has proven to be a highly efficient tool for exploring complex reaction pathways in biological systems such as ASNases, allowing the use of high-level Hamiltonians that would be computationally prohibitive with other approaches, such as multidimensional Umbrella Sampling. Although a relatively high number of collective variables were explored in some cases, leading to multidimensional configurational spaces, this did not reduce the overall effectiveness of the method. In addition to providing insights into the reaction mechanism, ASM has also been employed in this doctoral thesis to model the conformational change of flexible loops. Other free energy methods, such as Thermodynamic Integration, has been employed to predict pK_a values of enzymatic residues, as well as relative binding free energies.

Using this computational toolkit we have elucidated the complete enzymatic cycles and key intermediate steps of both type 1 and type 3 ASNases. Our findings indicate that both type 1 and type 3 ASNases follow a stepwise acyl-enzyme formation mechanism. While ASNase2 enzymes were not explicitly studied, their classification as class I ASNases, as the studied ASNase1, suggests they likely share a similar reaction pathway. Therefore, given the consistent behaviour across the studied ASNases, we are tempted to propose that the stepwise mechanism may be common to all ASNase types, though further studies are needed to confirm that ASNase2 also follows the same pathway.

Our findings establish a crucial connection between the oligomeric forms of the studied ASNases and their enzymatic function. To the best of our knowledge, no theoretical framework has yet been published studying the connection between the reaction mechanisms and the active forms of ASNases, highlighting a significant gap in the current understanding of ASNases. These findings are very significant because, in both cases, the active site residues originate exclusively from one monomer, suggesting that each subunit could potentially function as an active form. For ASNase type 3, which exists as a dimer, we demonstrated that a single monomer does not retain a catalytically active configuration. For type 1,

which is a homotetramer, we demonstrated experimentally that a dimeric form of ASNase1 does not exhibit catalytic activity. MD simulations provided an interpretation of the lack of activity of the dimer based on protomer-protomer interactions needed to keep an active configuration.

Our research has significantly advanced the understanding of the enzymatic properties of gpASNase1. We provided a detailed description of the reaction mechanisms and the impact of the surrounding protein in terms of the electric field created by its residues and structural motifs. Additionally, we contributed to the understanding of the absence of glutaminase activity in gpASNase1, untangling its substrate specificity and identifying the key residues responsible for this selectivity. Finally, we also investigated the role of different residues in the conformational change of the loop closing the active site. By studying both gpASNase1 and hASNase1, we established a theoretical explanation for the mutations introduced in recently patented humanized chimeras. In addition to clarifying the observed properties of these chimeras, our findings also provided valuable insights into potential approaches for further optimizing other engineered variants.

Considering the remarkable progress in protein design during recent years, future research on ASNases in leukemia therapy should, in my opinion, focus on two main directions: advancing the redesign of the native backbone of ASNases to improve their stability and expresibility and the *de novo* design to get smaller and easier to obtain ASNases. A part of this doctoral thesis focuses on the redesign of the native sequence of the gpASNase1 enzyme using ProteinMPNN. We have shown that the expressibility can be enhanced through the introduction of mutations that, in principle, should not affect the catalytic activity. Unfortunately, our designs failed to condensate into the active tetrameric form. This ongoing project focuses on optimizing the redesigned variants by stabilizing key residues at the enzyme interfaces, ensuring the formation of an active oligomer and assessing the catalytic activity of the new designs. The main aim is to enhance the stability and expressibility of ASNase, with the goal of developing more effective therapeutic candidates. Furthermore, it would be quite promising to explore the coupling of ProteinMPNN predictions with Major histocompatibility complex (MHC) binding affinity data, allowing for the design of ASNase variants with reduced immunogenicity, which would enhance their therapeutic potential.

Beyond redesigning the native enzyme, an interesting avenue for future exploration also lies in the *de novo* design, particularly at the light of the significant advances made by the Baker Lab in the *de novo* design of other hydrolases. Given the relatively large size of natural ASNases, developing of a more compact enzyme with optimized catalytic properties could provide substantial advantages for leukemia therapy. Utilizing methods for the *de novo* protein design, it may be possible to create ASNases that are not only more stable and expressible but that also could potentially exhibit reduced immunogenicity, enhancing their suitability for therapeutic applications in leukemia treatment.

Chapter 6: Resumen

6.1 Introducción

La leucemia es una neoplasia hematológica que afecta a las células sanguíneas y la médula ósea, caracterizada por una proliferación y diferenciación descontrolada de los glóbulos blancos.[1, 2] Esto provoca un desplazamiento progresivo de las células sanguíneas normales y saludables, lo que resulta en una acumulación de células disfuncionales en el organismo. En 2018, se clasificó como el decimoquinto cáncer más diagnosticado, con una estimación de más de trescientas mil muertes anuales en los Estados Unidos debido a esta enfermedad. Algunos de los factores de riesgo de la leucemia incluyen la exposición a la radiación, síndromes hereditarios, el tabaquismo, la edad y algunos factores desconocidos.[3]

Según las células afectadas, la leucemia se clasifica principalmente en cuatro tipos principales: leucemia mieloide aguda (LMA), leucemia mieloide crónica (LMC), leucemia linfoblástica aguda (LLA) y leucemia linfocítica crónica (LLC). En términos generales, las células madre hematopoyéticas primarias se diferencian en células madre mieloides o linfoides, que posteriormente dan lugar a diversos tipos celulares especializados (Figura 6.1). Las células madre linfoides originan linfoblastos, los cuales se diferencian en células asesinas naturales (NK), linfocitos T o linfocitos B. Por otro lado, las células madre mieloides se desarrollan en eritrocitos, plaquetas (trombocitos) o mieloblastos. A su vez, los mieloblastos se diferencian en granulocitos (neutrófilos, eosinófilos o basófilos). Además, los linfocitos B y T, las células NK y los granulocitos constituyen los glóbulos blancos, fundamentales para el sistema inmunológico.

En la leucemia mieloide (LM), los mieloblastos se dividen de manera incontrolada, lo que da lugar a la formación de glóbulos blancos granulocíticos anormales que no pueden desempeñar sus funciones normales. Este tipo de leucemia también se conoce como leucemia granulocítica o no linfoide.[4] Como sus nombres indican, la leucemia mieloide aguda progresa rápidamente, mientras que la leucemia mieloide crónica se desarrolla de manera más lenta.

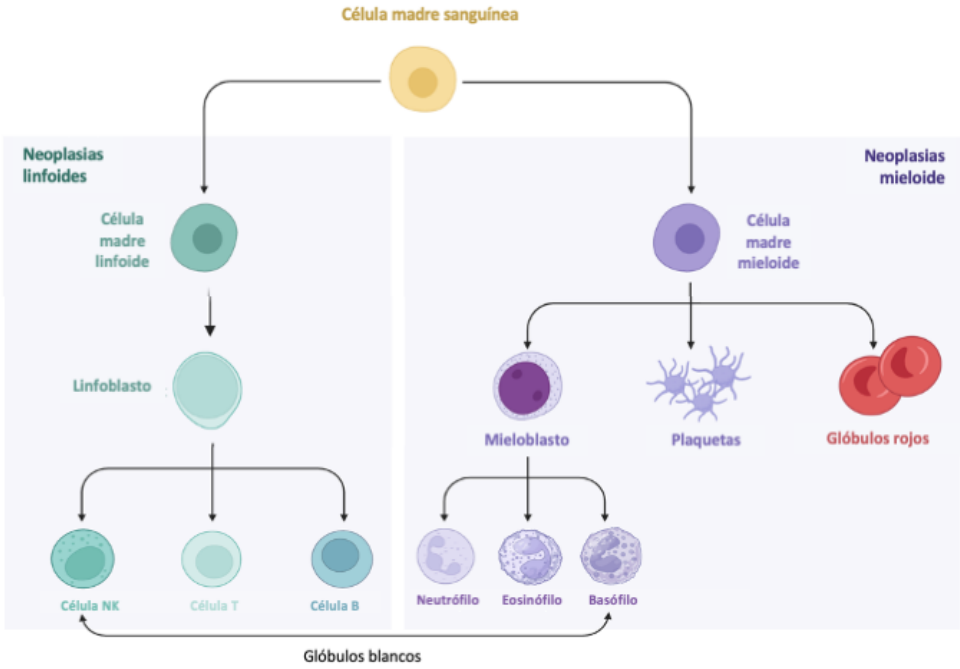


Figura 6.1. Diferenciación de las células madre sanguíneas.

De manera similar, la leucemia linfoblástica (LL) se caracteriza por la proliferación incontrolada de glóbulos blancos en la médula ósea. Dependiendo del subtipo de glóbulo blanco afectado (B, T o NK), la LL se puede subdividir, aunque las más comunes son la leucemia linfoblástica B y T. Nuevamente, la LLA se desarrolla muy rápidamente, causando síntomas inmediatos y graves, mientras que la LLC progresa más lentamente. Aunque puede presentarse en adultos, la LLA afecta principalmente a niños entre 1 y 7 años de edad.[5] También se ha observado una mayor incidencia en niños de ascendencia nativa americana, nativa de Alaska e hispana, así como casi el doble en niños blancos en comparación con los niños negros.[6]

Las L-Asparaginasas (L-ASNasas o simplemente ASNasas, EC 3.5.1.1) son una clase de enzimas hidrolasas que catalizan la transformación de asparagina (Asn) en aspartato (Asp) y amoníaco (Figura 6.2).

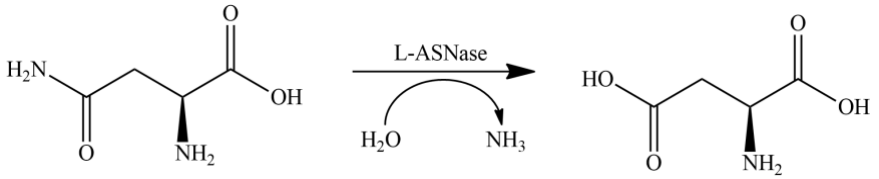


Figura 6.2. Representación esquemática de la reacción de L-Asparaginasa.

Estas enzimas han sido un componente clave en el tratamiento de la leucemia linfoblástica aguda (LLA) y el linfoma no Hodgkin desde la década de 1960.[15] La conexión entre las L-asparaginasas y sus efectos antileucémicos se descubrió por primera vez en la década de 1950, cuando Kidd demostró que el suero de cobaya podía inducir la regresión de linfomas trasplantados en ratones.[16] Posteriormente, a principios de la década de 1960, Broome identificó a las ASNasas como el componente clave detrás de la actividad terapéutica de dicho suero.[17] Pocos años después, Yellin y Wriston lograron aislar una L-ASNasa a partir del suero de cobaya y proporcionaron evidencia directa de su eficacia contra la leucemia.[18] Este avance condujo a su aplicación clínica ese mismo año, marcando un hito significativo en la terapia de la LLA. La gran importancia de las ASNasas en la práctica clínica también se evidencia en su inclusión en la Lista de Medicamentos Esenciales de la Organización Mundial de la Salud.

El efecto terapéutico de las ASNasas en el tratamiento de la LLA se debe a la incapacidad de las células blásticas de la leucemia para sintetizar asparagina de manera independiente. En concreto, estas células presentan poca o ninguna actividad detectable de la enzima asparagina sintetasa (ASNS).[19] La ASNS cataliza la síntesis de asparagina (Asn) a partir de aspartato mediante una reacción dependiente de ATP. Como resultado, la supervivencia de las células malignas depende completamente del suministro exógeno de L-asparagina proveniente del suero del paciente. Por lo tanto, la administración intramuscular o intravenosa de una enzima ASNasa agota la Asn circulante en la sangre, privando a las células blásticas de este nutriente esencial. La inhibición de la síntesis de proteínas debido a la falta de Asn finalmente desencadena la apoptosis de las células blásticas.[20] Es importante destacar que las células sanguíneas normales no se ven afectadas por este tratamiento, ya que tienen la capacidad de sintetizar L-

asparagina. Esto ha sentado una base sólida para el desarrollo de lo que se conoce como Terapia contra el Cáncer por Depleción de Aminoácidos.[21]

Basándose en la fuente de la primera enzima descubierta, las ASNasas se clasificaron inicialmente en tres clases canónicas: bacteriana, vegetal y de la clase Rhizobium etli (ver Figura 6.3).

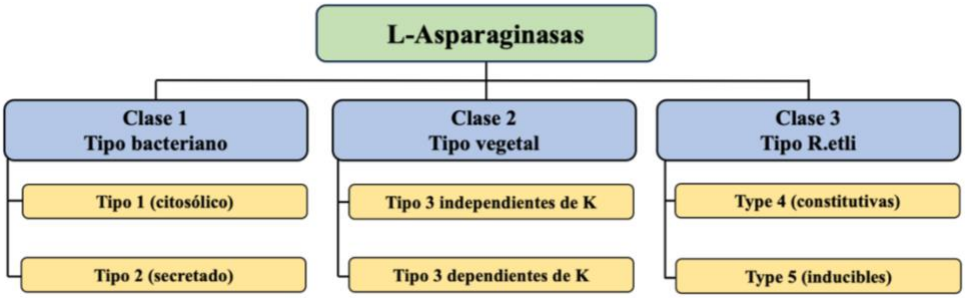


Figura 6.3. Clasificación de L-asparaginasas basada en las estructuras. Adaptado de [23].

Sin embargo, esta clasificación pronto resultó engañosa, ya que diversos organismos producen distintos tipos de ASNasas que se distribuyen en varias clases.[23] Por ejemplo, aunque algunas ASNasas de *E. coli* se clasifican como de tipo bacteriano, otras ASNasas de *E. coli* son más bien de tipo vegetal. Por esta razón, hoy en día estas enzimas han sido renombradas como clases 1, 2 y 3, con referencia a las clases canónicas mencionadas anteriormente.[24] Además, las ASNasas también se dividen en tipos 1, 2, etc., según sus similitudes estructurales. En esta tesis doctoral, se seguirá esta propuesta de clasificación dada por Silva et al. [24] (Figura 6.3). Esta clasificación fue originalmente establecida para las enzimas de *E. coli* y posteriormente se extendió a otras proteínas con una arquitectura similar. Por ejemplo, la enzima con mayor similitud estructural con la ASNasa tipo 1 de *E. coli* se clasifica, por lo tanto, como una enzima de tipo 1.

Existen diferentes formas en las que las ASNasas se denominan en la literatura. En esta tesis doctoral, nos referiremos la mayoría de las veces a los tipos de

ASNasa y no a las clases. Por lo tanto, utilizaremos la siguiente regla: el nombre de una xASNasa se construye de tal manera que x (las primeras letras) corresponde al organismo de origen (por ejemplo, *Escherichia coli* – Ec, humano – h, cobaya – gp, etc.) y, después del nombre, se añade un número (1, 2, etc.) para representar el tipo de enzima al que pertenece. Por ejemplo, EcASNase3 hace referencia a la asparaginasa tipo 3 de *E. coli*.

6.2 Objetivos

El objetivo de esta tesis doctoral es investigar las propiedades catalíticas, los mecanismos de reacción y la dinámica de las L-asparaginasas (ASNasas) de diferentes tipos. Al racionalizar los orígenes de las propiedades catalíticas favorables e identificar los factores que afectan la actividad enzimática, los resultados de esta tesis buscan guiar tanto la ingeniería racional como el diseño de novo de enzimas con mayor eficacia terapéutica.

Mediante simulaciones de dinámica molecular clásica, pretendemos explorar el comportamiento dinámico, los cambios conformacionales y las formas fisiológicamente activas de diversos tipos de ASNasas, aspectos que hasta la fecha no han sido completamente explicados. A través de simulaciones multiescala, analizaremos los factores clave que influyen en la actividad de las ASNasas, con un enfoque en resolver discrepancias entre las propuestas mecanísticas existentes para los distintos tipos de ASNasas. Además, planeamos investigar los valores de pK_a de los residuos catalíticos clave utilizando métodos de energía libre, ya que estos son cruciales para la comprensión mecanística.

Asimismo, examinaremos quimeras desarrolladas recientemente mediante recombinación dirigida por ADN con el fin de demostrar que nuestros métodos pueden predecir con precisión las posiciones de los aminoácidos retenidos por evolución dirigida para conservar una actividad enzimática óptima.

Esta tesis también busca integrar estos conocimientos con métodos de vanguardia para el rediseño de secuencias nativas, como ProteinMPNN. El objetivo es rediseñar la secuencia nativa de la ASNasa más prometedora para generar variantes con mayor expresabilidad, eficiencia catalítica, estabilidad y selectividad, obteniendo así variantes con mejores propiedades terapéuticas. Finalmente, una vez validadas experimentalmente, revisitaremos los métodos de química computacional y proporcionaremos un marco teórico para explicar las variaciones observadas en la estabilidad y actividad de las variantes rediseñadas.

6.3 Metodología

Simulaciones de Dinámica Molecular

Las simulaciones de Dinámica Molecular (MD) permiten estudiar sistemas de gran tamaño y procesos complejos, como interacciones bioquímicas, proporcionando información detallada sobre los procesos estudiados. El principio básico detrás de estas simulaciones es la ecuación de movimiento de Newton. Una fuerza F_i que actúa sobre cada átomo de masa m_i, causando que el átomo se acelere de acuerdo con:

$$\boldsymbol{F_i} = m_i \cdot \frac{d^2\boldsymbol{r_i}}{dt^2} \tag{6.1}$$

donde $\boldsymbol{r_i}$ representa el vector de posición de cada átomo, mientras que su segunda derivada con respecto al tiempo *t* representa la aceleración. En las simulaciones de MD esta fuerza se deriva del gradiente negativo de la energía potencial (E) del sistema:

$$\boldsymbol{F_i} = -\nabla_i \cdot E \tag{6.2}$$

Los modelos matemáticos de las funciones de energía potencial, también conocidos como campos de fuerza, tienen en cuenta tanto las interacciones de enlace (estiramiento de enlaces, flexión de ángulos, torsiones) como las interacciones intermoleculares (fuerzas de van der Waals y electrostáticas). Al combinar las ecuaciones (6.1) y (6.2) e integrar las ecuaciones del movimiento de Newton a lo largo de pequeños intervalos de tiempo, se pueden derivar tanto las velocidades como las posiciones de cada átomo a cada paso de tiempo. Debido a la complejidad de las funciones de energía potencial, las ecuaciones de Newton no pueden resolverse analíticamente, por lo que se utilizan algoritmos numéricos como el método de Verlet [81] para integrar las trayectorias atómicas de manera eficiente, conservando la estabilidad del sistema. Para evitar efectos no físicos en sistemas de tamaño finito, se aplican condiciones de contorno periódicas (PBCs), que simulan un entorno infinito al hacer que las partículas que salen de un límite

reingresen por el lado opuesto de la caja de simulación. Esto permite representar propiedades en fases condensadas sin introducir artefactos de frontera no deseados.

Métodos de energía libre

Los cálculos de energía libre son fundamentales en la simulación de sistemas biológicos y químicos, ya que nos permiten estimar diferencias de energía libre entre estados y predecir la estabilidad de estructuras y procesos.

Las diferencias de energía libre relativas entre dos estados son más accesibles mediante simulaciones que la energía libre absoluta. En muchos casos, estas diferencias relativas son suficientes para abordar cuestiones termodinámicas y cinéticas, como la unión de ligandos, cambios conformacionales o rutas de reacción. En termodinámica, la elección entre la energía libre de Helmholtz (A) y la de Gibbs (G) depende de las condiciones del sistema. Mientras que la energía de Gibbs es relevante para sistemas a presión y temperatura constantes (NPT), en simulaciones de MD generalmente se usa la energía de Helmholtz, ya que los sistemas suelen mantenerse a volumen y temperatura constantes (NVT). Sin embargo, en muchas reacciones enzimáticas, los cambios de volumen son pequeños, lo que permite aproximar la energía de Gibbs mediante la de Helmholtz.

Perturbación de Energía Libre

Este método se basa en la ecuación de perturbación de Zwanzig [228], la cual permite calcular la diferencia de energía libre de Helmholtz entre dos estados (I y II) mediante la relación exponencial:

$$\Delta A = -kTln\langle e^{-\frac{H_{II}-H_I}{kT}}\rangle_I \tag{6.3}$$

Aquí, H_I y H_{II} son los Hamiltonianos de los estados I y II, k es la constante de Boltzmann y T la temperatura. La media se toma sobre las configuraciones del estado inicial.

Al muestrear configuraciones del estado I, el estado II puede ser visitado con frecuencia si la diferencia de energía entre los dos estados es comparable con la energía térmica (kT). Sin embargo, cuando la diferencia de energía entre los dos estados excede esta, la probabilidad de transiciones se vuelve prácticamente nula, lo cual es común en simulaciones de sistemas bioquímicos. Esto lleva a un muestreo deficiente del estado II, lo que hace que las configuraciones de este estado contribuyan mínimamente al promedio del conjunto. Para superar este problema, se pueden introducir estados intermedios, con pequeñas diferencias de energía entre estados consecutivos, que efectivamente cierran la brecha energética y permiten calcular la diferencia de energía libre como la suma de incrementos más pequeños y manejables.

Integración Termodinámica

La integración termodinámica se basa en calcular la derivada de la energía libre respecto a un parámetro de acoplamiento λ, que modula la transición entre dos estados:

$$\Delta A = \int_0^1 \langle \frac{\partial H}{\partial \lambda} \rangle_\lambda \, d\lambda \tag{6.4}$$

Por lo tanto, el cambio de energía libre se puede calcular mediante la realización de simulaciones a diferentes valores de λ y computando el valor promedio de la derivada del Hamiltoniano. La integración generalmente se realiza utilizando algunas técnicas numéricas comunes, como la regla de Simpson o la cuadratura gaussiana.

Perfil de Energía Potencial de Media Fuerza

Un PMF (Potencial de Fuerza Media, por sus siglas en inglés) permite determinar el cambio de energía libre en función de una coordenada de reacción (ξ):

$$W(\xi) = C' - kTln\rho(\xi) \tag{6.5}$$

Es decir, se puede simplemente medir la probabilidad de que un sistema simulado visite una configuración con la coordenada seleccionada tomando un valor entre ξ y ξ+Δξ durante la simulación de dinámica molecular y, a partir de ahí, obtener el PMF.

Este método, como los anteriores, también puede encontrar dificultades cuando la energía libre entre dos valores de la coordenada es significativamente mayor que *kT*, ya que ambos valores no serán adecuadamente muestreados a lo largo de una sola simulación. Por lo tanto, se han desarrollado algunas técnicas de muestreo mejorado, como el método de Muestreo con Potencial de Fuerza Sesgada (Umbrella Sampling, US).[90]

Muestreo con Potencial de Fuerza Sesgada (Umbrella Sampling, US)

En este método se aplica un sesgo (V_{umb}), un término adicional de energía que depende solo de la coordenada de reacción ξ, con el fin de mejorar el muestreo en el vecindario de un valor particular de la coordenada ξ. La nueva función de energía total corresponde a la suma del potencial no sesgado $H(r)$ y el potencial de sesgo:

$$H_{biased}(r) = H(r) + V_{umb}(\xi) \tag{6.6}$$

El muestreo sesgado introduce potenciales de restricción artificiales para mejorar el muestreo en regiones energéticamente desfavorables. Sin embargo, utilizando métodos como WHAM (el método de análisis de histograma ponderado) [91] se puede encontrar una distribución de probabilidad no sesgada y, finalmente, la energía libre se puede determinar cómo:

$$A_{unbiased}(\xi) = -\frac{1}{kT} ln\rho(\xi) \tag{6.7}$$

Donde $\rho(\xi)$ es la distribución de probabilidad de la coordenada de reacción, corregida por el peso del potencial aplicado en cada ventana de muestreo.

Métodos de Mecánica Molecular Poisson−Boltzmann y Generalized Born Surface Area (MMGBSA y MMPBSA)

Los enfoques más conocidos para calcular las energías de los puntos finales son Molecular Mechanics Poisson−Boltzmann Surface Area (MM/PBSA) y Molecular Mechanics Generalized Born Surface Area (MM/GBSA), aplicados por primera vez por Kollman et al.[92] Por ejemplo, consideremos el caso de un ligando (*L*) que se une a un receptor proteico (*R*), formando un complejo receptor-ligando (*RL*):

$$L + R \rightarrow RL \tag{6.8}$$

La energía libre de unión ΔG_{bind} se estima a partir de las energías libres de los reactivos y productos como un conjunto de promedios:

$$\Delta G_{bind} = \langle G_{RL} - G_L - G_R \rangle_{RL} \tag{6.9}$$

Donde G_{RL}, G_L y G_R son las energía libres del complejo, del ligando y del receptor, respectivamente. Estrictamente, los promedios deberían determinarse a partir de simulaciones separadas del complejo, el ligando libre y el receptor no unido. Sin embargo, es una práctica común determinar las energías libres a partir de la simulación de un complejo receptor-ligando, eliminando átomos.

Análisis del Potencial Eléctricostatico y del Campo Eléctrico

Las fuerzas electrostáticas pueden ser cruciales para muchos procesos biológicos, especialmente debido que las fuerzas electrostáticas suelen ajustar el entorno del sitio activo para estabilizar los estados de transición e intermedios.[95, 96] Es por eso que el análisis del campo eléctrico en las simulaciones de MD puede proporcionar valiosos conocimientos sobre los orígenes de la estabilización del estado de transición (TS) y las fuerzas impulsoras detrás de la catálisis enzimática. [95, 96]

El campo eléctrico $\boldsymbol{E}$ de N cargas puntuales en un punto específico se puede calcular como:

$$\boldsymbol{E} = \sum_{i=1}^{N} \frac{1}{4\pi\varepsilon_0} \frac{Q_i}{r_i^2} \cdot \boldsymbol{u_r} \tag{6.10}$$

donde ε_0 representa la permitividad eléctrica, Q_i es la carga parcial del átomo, $\boldsymbol{r_i}$ es la distancia entre el átomo i en el punto en el espacio y $\boldsymbol{u_r}$ es el vector unitario de la distancia. Al definir un punto en el espacio (llamado "sonda") e iterar sobre todos los átomos N, en cada configuración de la trayectoria, se puede calcular la magnitud y la dirección de los campos eléctricos promediados sobre la simulación. Este análisis también se puede realizar a nivel de cada residuo, lo cual es un método efectivo para identificar los "residuos más importantes" que determinan las propiedades electrostáticas de una enzima. Además, es posible combinar las contribuciones individuales de los residuos dentro de un motivo estructural al que pertenecen (por ejemplo, una hélice alfa o una hoja beta), proporcionando información sobre el papel de los motivos estructurales de las proteínas en la estabilización del estado de transición.

Simulaciones de Mecánica Cuántica / Mecánica Molecular (QM/MM)

El análisis de reacciones químicas requiere una descripción explícita de los electrones involucrados en el proceso de ruptura/formación de enlaces, por lo tanto, un tratamiento cuántico (QM). Por otro lado, los cálculos QM para un sistema enzimático completo pueden ser extremadamente costosos computacionalmente, especialmente cuando se deben muestrear muchas configuraciones para obtener promedios estadísticos. Por lo tanto, una de las soluciones adoptadas es el enfoque híbrido de Mecánica Cuántica/Mecánica Molecular (QM/MM), un esquema multiescala concurrente que combina las fortalezas de los métodos QM y MM. [96-99] En este marco, la descripción cuántica se enfoca en el subsistema reactivo, incluyendo el sitio activo de una enzima y el sustrato. Existen dos enfoques QM/MM: el esquema sustractivo y el aditivo.

El enfoque sustractivo calcula la energía restando la energía de la región QM, calculada por un método MM de la energía total del sistema, calculada a nivel MM y luego sumando la energía de la región QM. En el enfoque aditivo, la

energía total se calcula como la suma de la energía de la región QM, de acuerdo con el método QM, la energía de la región MM, de acuerdo con el método MM, y la interacción entre las regiones QM y MM. El acoplamiento de los métodos de mecánica cuántica (QM) y mecánica molecular clásica (MM), también conocido como embebido, asegura que las propiedades electrónicas de la región QM estén influenciadas por el entorno clásico, lo que lleva a una representación más precisa. Se pueden utilizar tres esquemas principales de embebido: (*i*) esquema de embebido mecánico, (*ii*) esquema de embebido electrostático y (*iii*) esquema de embebido de polarización. Sin embargo, para sistemas enzimáticos, hasta ahora se utilizan preferencialmente los esquemas electrostáticos, permitiendo incorporar los efectos electrostáticos del entorno sobre el sistema reactive-

Explorando los caminos de energía libre de reacción: El Método de Cuerda Adaptativa

Algunos métodos basados en caminos proyectan la superficie de energía libre multidimensional (FES, por sus siglas en inglés) sobre una trayectoria de menor dimensionalidad. Idealmente, esta trayectoria se define a partir de una única coordenada de reacción (RC) representativa, capaz de guiar al sistema a lo largo del proceso de interés. Estos métodos asumen que una reacción tiene más probabilidades de seguir un camino específico o "tubo de reacción". El "tubo de reacción" se refiere a una región estrecha que conecta los pozos de los reactivos y el producto, e incluye la mayoría de las trayectorias reactivas. Esto permite proyectar el espacio de energía libre sobre una sola variable colectiva del camino (CV del camino) que define el progreso a lo largo del mismo. Existen ligeras diferencias en las definiciones de los caminos, y los métodos utilizados para determinar un camino varían en consecuencia. En general, para un tubo de reacción estrecho y una FES suave, el camino de mínima energía libre (MFEP, por sus siglas en inglés) y otras definiciones de camino se espera que sean similares.[146] Uno de los métodos de optimización de caminos más utilizados y bien establecidos es el método de la cuerda. En este método, se ejecutan dinámicas para los estados intermedios, los nodos de la cuerda, que se mueven según los gradientes de energía libre mientras se mantienen equidistantes. De esta manera, la cuerda converge hacia el Camino de Mínima Energía Libre (MFEP).

En esta tesis doctoral, se utiliza el Método de cuerda Adaptativa (ASM, por sus siglas en inglés) para explorar los paisajes de energía libre de las reacciones enzimáticas.[147, 193]

La convergencia de la cuerda se evalúa utilizando la desviación cuadrática media (RMSD) de las CVs, con respecto a las estructuras del paso anterior. Una vez que la cuerda está convergida, se define la variable colectiva del camino, a lo largo de la cual se utiliza el método US para evaluar el perfil de energía libre a lo largo de esta coordenada de reacción.

La predicción de la estructura de proteínas y el diseño de proteínas de novo

Las proteínas se sintetizan como cadenas lineales de aminoácidos, pero su funcionalidad depende del plegamiento en estructuras tridimensionales específicas. Un buen plegamiento garantiza su estabilidad, actividad y función biológica, mientras que el mal plegamiento puede generar enfermedades. Según la hipótesis de Anfinsen, la información de plegamiento está codificada en el paisaje energético de la proteína, donde el estado nativo es la conformación de menor energía libre.[151]

Con el crecimiento de nuevas secuencias proteicas y la dificultad de la cristalización, la determinación de estructuras ha sido un desafío. Sin embargo, avances recientes en técnicas computacionales, como AlphaFold [153] y RoseTTAFold [154], han mejorado significativamente la predicción estructural. También existen métodos híbridos y de aprendizaje profundo que optimizan la predicción de estructuras, aunque con algunas limitaciones.[113]

El diseño de proteínas *de novo* permite explorar nuevas secuencias más allá de las encontradas durante la evolución, lo que abre nuevas posibilidades para la creación de enzimas con funciones específicas no presentes en la naturaleza. Este campo es muy prometedor, especialmente considerando los logros recientes del Instituto de Diseño de Proteínas y el laboratorio de Baker. El desarrollo de software para la generación de estructuras de espina dorsal, como RFdiffusion

[175], junto con herramientas de diseño de secuencias como ProteinMPNN [178], ha abierto nuevas posibilidades en el diseño de proteínas *de novo*.

6.4 Resultados principales y conclusiones

Asparaginasa humana tipo 3 (hASNase3)

Los resultados de las simulaciones clásicas de dinámica molecular identificaron interacciones clave entre el sustrato y los residuos del sitio activo en el complejo de Michaelis. Las simulaciones demuestran que la forma dimérica de hASNase3 es esencial para la actividad catalítica. Aunque los residuos del sitio activo provienen de un solo protómero, las interacciones entre protómeros son críticas para mantener la posición del sustrato y estabilizar el sitio activo. Se encontró que el lazo de unión al sodio era estable durante la simulación y el análisis de interacciones reveló su papel crucial en la estabilización del residuo C-terminal generado después de la escisión de la proteína (Gly167), lo que, a su vez, mantiene al residuo nucleófilo N-terminal (Thr168) correctamente posicionado durante la reacción catalítica.

Los resultados de las simulaciones de dinámica molecular del complejo de Michaelis identificaron a Thr168 como el posible nucleófilo. Para que su grupo hidroxilo sea activado, se ha propuesto que su propio grupo amino resta un protón. Para estudiar esta posibilidad, se ha determinado el pK_a de Thr168 mediante técnicas de energía libre. A pH 7.5, el grupo N-terminal de Thr168 está predominantemente protonado, pero un pK_a más bajo en la enzima que en disolución debido a la menor capa de hidratación. En consecuencia, el costo en energía libre determinado para la deprotonación es mínimo (1.38 kcal·mol^{-1}), lo que sugiere que el grupo amino podría activar realmente el grupo hidroxilo del mismo residuo.

El mecanismo de reacción fue explorado utilizando simulaciones QM/MM multiescala y ASM. Las propuestas más prometedoras fueron recalculadas a nivel de teoría B3LYPD3/6-31+G*/MM. El ciclo completo de reacción consta de 3 etapas principales: formación del acilo-enzima, hidrólisis y regeneración de hASNase3, y 7 estados de transición se forman a lo largo del proceso catalítico. La primera etapa comienza con un grupo N-terminal desprotonado en Thr168 que abstrae un protón del grupo hidroxilo del mismo residuo. La reacción luego

procede con el ataque nucleofílico del átomo Oγ de Thr168 al Cγ del sustrato con la posterior liberación de amoníaco. El mecanismo propuesto para la formación del complejo acilo-enzima explica su potencial para la cristalización, ya que esta estructura aparece como un mínimo global en el perfil de energía libre de la reacción. Además, el complejo acilo-enzima obtenido computacionalmente se alinea bien con la estructura de rayos X [37]. En la segunda etapa, una molécula de agua se activa por el grupo amino del nucleófilo N-terminal para atacar el Cγ del sustrato, alcanzando de esta manera el paso limitante de la velocidad del proceso global, con la barrera de energía libre siendo 20.9 $kcal \cdot mol^{-1}$. Finalmente, el estado de protonación de la enzima se regenera mediante una transferencia de protón del ácido aspártico al grupo amino de Thr168, seguida de una caída monótona en la energía libre, resultando en un proceso exergónico con una energía libre de reacción de -7.2 $kcal \cdot mol^{-1}$. Esta propuesta también explica la falta de reactividad en los mutantes Thr219Val, Thr219Ala y Thr186Val. Específicamente, se encontró que Thr186 estabiliza la carga negativa en el nucleófilo activado, mientras que Thr219 actúa como un residuo de "oxyanion hole", estabilizando la acumulación de carga negativa en el átomo de oxígeno del sustrato.

Asparaginasa de cobaya tipo 1 (gpASNase1) y humana tipo 1 (hASNase1)

El análisis de las simulaciones de dinámica molecular del complejo de Michaelis con asparagina presente en el sitio activo identificó interacciones clave para la actividad catalítica. Encontramos que, a pH neutro, el residuo Lys188 está protonado y no puede funcionar fácilmente como base catalítica. En su lugar, Thr19 puede ser activado para actuar como el nucleófilo a través de la transferencia de protón a Tyr308', que a su vez se desprotona después de un proceso mediado por moléculas de agua hasta el residuo Asp117. La proximidad cercana del átomo Cγ del sustrato al oxígeno hidroxilo de Thr19 confirma su rol como nucleófilo, esencial para la formación del complejo acilo-enzima. El paso limitante de la velocidad corresponde al ataque nucleofílico de Thr19 sobre el sustrato, acoplado a su desprotonación por Tyr308', lo que presenta una energía libre de activación de 18.6 $kcal \cdot mol^{-1}$ calculada a nivel B3LYP-D3/6-31+G(d)/MM, en buen acuerdo con el valor derivado experimentalmente. El

análisis del campo eléctrico reveló que los residuos Gly18, Asp84, Thr116 y Tyr308', junto con las hélices α_1 y α_4, juegan un papel crucial en la estabilización de la carga negativa en el oxígeno carbonílico del sustrato durante el estado de transición limitante de la velocidad, mejorando la eficiencia catalítica enzimática.

Investigamos la dinámica conformacional del lazo de Tyr, que contiene el residuo catalítico Tyr308' crucial para la activación del nucleófilo durante la catálisis. Los resultados muestran que el cierre del lazo es favorecido tras la unión del sustrato, siendo la forma cerrada más estable en la holoenzima y la forma abierta en la apoenzima. Además, el análisis de la energía de interacción por residuo identificó residuos clave involucrados en la estabilización tanto de las conformaciones abiertas como cerradas del lazo.

Los cálculos de integración termodinámica estimaron la energía libre de unión relativa entre la asparagina y la glutamina, alineándose con la observación experimental de la falta de actividad glutaminasa en gpASNase1. El análisis de las contribuciones de la energía de interacción por residuo reveló que residuos clave del sitio activo, como Asp84, Ser85, Thr116 y Asp117, juegan un papel significativo en la unión de la asparagina al interactuar con los grupos carboxilato y amino cargados del sustrato. Además, Asp190 y Ala142 contribuyen a interacciones con la porción no zwitteriónica de la asparagina, favoreciendo aún más su unión en lugar de la glutamina.

Considerando todos los resultados obtenidos, junto con la comparación de las contribuciones por residuo a las energías libres de unión de hASNase1 y gpASNase1, se pudieron explicar alrededor del 40% de las mutaciones en las quimeras obtenidas tras el proceso de recombinación dirigida de AND entre enzimas gpASNase1 y hASNase1.[29] A pesar de que las quimeras presentaban una alta identidad de secuencia con hASNase1, su actividad catalítica se mantenía comparable a la de gpASNase1. Es importante destacar que en esta tesis doctoral se identificaron seis mutaciones adicionales (Arg52Gln, Arg54Gln, Glu58Asp, Asp59His, Arg68His e Ile72Val) que podrían mejorar las propiedades de las quimeras. También se predijo que los dos motivos estructurales, las hélices

α_1 y α_4, conservados en el MSA y cruciales para la estabilización del estado de transición, no causarían reacciones alérgicas significativas.

Diseño de novo de hidrolasa de epóxido soluble (sEH)

Las simulaciones MD resultaron ser un recurso valioso para guiar el diseño de enzimas *de novo*. Al modelar las interacciones enzima-sustrato a nivel atómico, las simulaciones MD permitieron una comprensión más profunda de las interacciones entre los residuos del sitio activo y los sustratos, lo que llevó a decisiones sobre el diseño mejor fundamentadas.

Siguiendo el estándar de la línea de diseño de proteinas *de novo* desarrollado en el Instituto de Diseño de Proteínas (RFdiffusion [175], Protein and LigandMPNN [178, 179]), y con la ayuda del análisis de las simulaciones MD clásicas, logramos diseñar seis hidrolasas de epóxido funcionales que exhibieron k_{cat}/K_M comparables con las de las proteínas nativas.

Rediseño con MPNN de la estructura nativa del esqueleto de gpASNase1

El rediseño de gpASNase1 proporcionó información sobre su forma activa y, más ampliamente, sobre las formas activas de las ASNasas de tipo 1, aunque no se han diseñado variantes catalíticamente activas con éxito. Dado que los residuos de la interfaz entre protómerosno se preservaron, las variantes rediseñadas formaron estructuras diméricas en lugar de tetraméricas. Las simulaciones de gpASNase1 dimérica ofrecieron una base teórica para entender la observación experimental de que estas variantes carecen de actividad enzimática. El análisis de estas simulaciones reveló que la ausencia del protómero adyacente causa que el bucle que contiene al residuo Gln143 se desplace lejos del sitio activo, alterando la disposición catalítica del centro activo y promoviendo una conformación no reactiva.

Chapter 7: References

1. Pui C-H, Robison LL, Look AT (2008) Acute lymphoblastic leukaemia. Lancet 371:1030–1043. https://doi.org/10.1016/S0140-6736(08)60457-2

2. Pagliaro L, Chen S-J, Herranz D, et al (2024) Acute lymphoblastic leukaemia. Nat Rev Dis Primers 10:1–28. https://doi.org/10.1038/s41572-024-00525-x

3. Schmidt J-A, Hornhardt S, Erdmann F, et al (2021) Risk Factors for Childhood Leukemia: Radiation and Beyond. Front Public Health 9:805757. https://doi.org/10.3389/fpubh.2021.805757

4. Short NJ, Rytting ME, Cortes JE (2018) Acute myeloid leukaemia. The Lancet 392:593–606. https://doi.org/10.1016/S0140-6736(18)31041-9

5. Vitale A, Guarini A, Chiaretti S, Foà R (2006) The changing scene of adult acute lymphoblastic leukemia. Curr Opin Oncol 18:652–659. https://doi.org/10.1097/01.cco.0000245317.82391.1b

6. Ekpa QL, Akahara PC, Anderson AM, et al A Review of Acute Lymphocytic Leukemia (ALL) in the Pediatric Population: Evaluating Current Trends and Changes in Guidelines in the Past Decade. Cureus 15:e49930. https://doi.org/10.7759/cureus.49930

7. Gale RP (2023) Radiation and leukaemia: Which leukaemias and what doses? Blood Rev 58:101017. https://doi.org/10.1016/j.blre.2022.101017

8. Stryckmans P, Debusscher L (1991) Chemotherapy of adult acute lymphoblastic leukaemia. Baillière's Clinical Haematology 4:115–130. https://doi.org/10.1016/S0950-3536(05)80287-2

9. Shyr D, Davis KL, Bertaina A (2023) Stem cell transplantation for ALL: you've always got a donor, why not always use it? Hematology Am Soc Hematol Educ Program 2023:84–90. https://doi.org/10.1182/hematology.2023000423

10. Zerra P, Bergsagel J, Keller FG, et al (2016) Maintenance Treatment With Low-Dose Mercaptopurine in Combination With Allopurinol in Children With Acute Lymphoblastic Leukemia and Mercaptopurine-Induced Pancreatitis. Pediatr Blood Cancer 63:712–715. https://doi.org/10.1002/pbc.25841

11. Kang MH, Kang YH, Szymanska B, et al (2007) Activity of vincristine, L-ASP, and dexamethasone against acute lymphoblastic leukemia is enhanced by the BH3-mimetic ABT-737 in vitro and in vivo. Blood 110:2057–2066. https://doi.org/10.1182/blood-2007-03-080325

12. Inaba H, Pui C-H (2010) Glucocorticoid use in acute lymphoblastic leukemia: comparison of prednisone and dexamethasone. Lancet Oncol 11:1096–1106. https://doi.org/10.1016/S1470-2045(10)70114-5

13. Sakura T, Hayakawa F, Sugiura I, et al (2018) High-dose methotrexate therapy significantly improved survival of adult acute lymphoblastic leukemia: a phase III study by JALSG. Leukemia 32:626–632. https://doi.org/10.1038/leu.2017.283

14. Zhao Y, Song H, Ni L (2019) Cyclophosphamide for the treatment of acute lymphoblastic leukemia. Medicine (Baltimore) 98:e14293. https://doi.org/10.1097/MD.0000000000014293

15. Kobrinsky NL, Sposto R, Shah NR, et al (2001) Outcomes of treatment of children and adolescents with recurrent non-Hodgkin's lymphoma and Hodgkin's disease with dexamethasone, etoposide, cisplatin, cytarabine, and L-asparaginase, maintenance chemotherapy, and transplantation: Children's cancer group. J Clin Oncol 19:2390–2396. https://doi.org/10.1200/JCO.2001.19.9.2390

16. Kidd JG (1953) Regression of transplanted lymphomas induced in vivo by means of normal guinea pig serum. I. Course of transplanted cancers of various kinds in mice and rats given guinea pig serum, horse serum, or rabbit serum. J Exp Med 98:565–582. https://doi.org/10.1084/jem.98.6.565

17. Broome JD (1963) Evidence that the L-asparaginase of guinea pig serum is responsible for its antilymphoma effects. I. Properties of the L-asparaginase of guinea pig serum in relation to those of the antilymphoma substance. The Journal of experimental medicine 118:99–120. https://doi.org/10.1084/jem.118.1.99

18. Yellin TO, Wriston JC (1966) Antagonism of purified asparaginase from guinea pig serum toward lymphoma. Science 151:998–999. https://doi.org/10.1126/science.151.3713.998

19. Chiu M, Taurino G, Bianchi MG, et al (2020) Asparagine Synthetase in Cancer: Beyond Acute Lymphoblastic Leukemia. Front Oncol 9:1–10. https://doi.org/10.3389/fonc.2019.01480

20. Appel IM, Kazemier KM, Boos J, et al (2008) Pharmacokinetic, pharmacodynamic and intracellular effects of PEG-asparaginase in newly diagnosed childhood acute lymphoblastic leukemia: results from a single agent window study. LEUKEMIA 22:1665–1679. https://doi.org/10.1038/leu.2008.165

21. Butler M, van der Meer LT, van Leeuwen FN (2021) Amino Acid Depletion Therapies: Starving Cancer Cells to Death. Trends Endocrinol Metab 32:367–381. https://doi.org/10.1016/j.tem.2021.03.003

22. Michalska K, Jaskolski M (2006) Structural aspects of L-asparaginases, their friends and relations. Acta Biochim Pol 53:627–640. https://doi.org/10.18388/abp.2006_3291

23. Loch JI, Jaskolski M (2021) Structural and biophysical aspects of L -asparaginases: a growing family with amazing diversity. IUCrJ 8:514–531. https://doi.org/10.1107/s2052252521006011

24. da Silva LS, Doonan LB, Pessoa A, et al (2022) Structural and functional diversity of asparaginases: Overview and recommendations for a revised nomenclature. Biotech and App Biochem 69:503–513. https://doi.org/10.1002/bab.2127

25. Campbell HA, Mashburn LT, Boyse EA, Old LJ (1967) Two L-asparaginases from Escherichia coli B. Their separation, purification, and antitumor activity. Biochemistry 6:721–730. https://doi.org/10.1021/bi00855a011

26. Ho PPK, Milikin EB, Bobbitt JL, et al (1970) Crystalline l-Asparaginase from Escherichia coli B. J Biol Chem 245:3708–3715. https://doi.org/10.1016/S0021-9258(18)62984-9

27. Aghaiypour K, Wlodawer A, Lubkowski J (2001) Structural basis for the activity and substrate specificity of Erwinia chrysanthemi L-asparaginase. Biochemistry 40:5655–5664. https://doi.org/10.1021/bi0029595

28. Schalk AM, Nguyen HA, Rigouin C, Lavie A (2014) Identification and structural analysis of an L-asparaginase enzyme from guinea pig with putative tumor cell killing properties. J Biol Chem 289:33175–33186. https://doi.org/10.1074/jbc.M114.609552

29. Rigouin C, Nguyen HA, Schalk AM, Lavie A (2017) Discovery of human-like L-asparaginases with potential clinical use by directed evolution. Scientific Reports 7:. https://doi.org/10.1038/s41598-017-10758-4

30. Swain AL, Jaskólski M, Housset D, et al (1993) Crystal structure of Escherichia coli L-asparaginase, an enzyme used in cancer therapy. Proc Natl Acad Sci U S A 90:1474–1478. https://doi.org/10.1073/pnas.90.4.1474

31. Maggi M, Meli M, Colombo G, Scotti C (2021) Revealing Escherichia coli type II l-asparaginase active site flexible loop in its open, ligand-free conformation. Sci Rep 11:18885. https://doi.org/10.1038/s41598-021-98455-1

32. Sánchez L, Medina FE, Mendoza F, et al (2022) Elucidation of the Reaction Mechanism of Cavia porcellus l-Asparaginase: A QM/MM Study. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.2c01122

33. Schalk AM, Antansijevic A, Caffrey M, Lavie A (2016) Experimental data in support of a direct displacement mechanism for type I/II L-asparaginases. J Biol Chem 291:5088–5100. https://doi.org/10.1074/jbc.M115.699884

34. Karamitros CS, Konrad M (2014) Human 60-kDa Lysophospholipase Contains an N-terminal l-Asparaginase Domain That Is Allosterically Regulated by l-Asparagine. Journal of Biological Chemistry 289:12962–12975. https://doi.org/10.1074/jbc.M113.545038

35. Yun M-K, Nourse A, White SW, et al (2007) Crystal Structure and Allosteric Regulation of the Cytoplasmic Escherichia coli L-Asparaginase I. J Mol Biol 369:794–811. https://doi.org/10.1016/j.jmb.2007.03.061

36. Schalk AM, Lavie A (2014) Structural and kinetic characterization of guinea pig l-asparaginase type III. Biochemistry 53:2318–2328. https://doi.org/10.1021/bi401692v

37. Nomme J, Su Y, Konrad M, Lavie A (2012) Structures of apo and product-bound human l-asparaginase: Insights into the mechanism of autoproteolysis and substrate hydrolysis. Biochemistry 51:6816–6826. https://doi.org/10.1021/bi300870g

38. Oinonen C, Rouvinen J (2000) Structural comparison of Ntn-hydrolases. Protein Sci 9:2329–2337. https://doi.org/10.1110/ps.9.12.2329

39. Loch JI, Klonecka A, Kądziołka K, et al (2022) Structural and biophysical studies of new l-asparaginase variants: lessons from random mutagenesis of the prototypic Escherichia coli Ntn-amidohydrolase. Acta Crystallogr D Biol Crystallogr 78:911–926. https://doi.org/10.1107/S2059798322005691

40. Michalska K, Hernandez-Santoyo A, Jaskolski M (2008) The Mechanism of Autocatalytic Activation of Plant-type L-Asparaginases. Journal of Biological Chemistry 283:13388–13397. https://doi.org/10.1074/jbc.M800746200

41. Bejger M, Imiolczyk B, Clavel D, et al (2014) Na+/K+ exchange switches the catalytic apparatus of potassium-dependent plant l-asparaginase. Acta Cryst D 70:1854–1872. https://doi.org/10.1107/S1399004714008700

42. Guillén-Navarro K, Araíza G, García-de los Santos A, et al (2005) The Rhizobium etli bioMNY operon is involved in biotin transport. FEMS Microbiology Letters 250:209–219. https://doi.org/10.1016/j.femsle.2005.07.020

43. Panosyan EH, Grigoryan RS, Avramis IA, et al (2004) Deamination of glutamine is a prerequisite for optimal asparagine deamination by asparaginases in vivo (CCG-1961). Anticancer Res 24:1121–1125

44. Offman MN, Krol M, Patel N, et al (2011) Rational engineering of L-asparaginase reveals importance of dual activity for cancer cell toxicity. Blood 117:1614–1621. https://doi.org/10.1182/blood-2010-07-298422

45. Avramis VI, Sencer S, Periclou AP, et al (2002) A randomized comparison of native Escherichia coli asparaginase and polyethylene glycol conjugated asparaginase for treatment of children with newly diagnosed standard-risk acute lymphoblastic leukemia: a Children's Cancer Group study. Blood 99:1986–1994. https://doi.org/10.1182/blood.v99.6.1986

46. Plourde PV, Jeha S, Hijiya N, et al (2014) Safety profile of asparaginase Erwinia chrysanthemi in a large compassionate-use trial. Pediatric Blood & Cancer 61:1232–1238. https://doi.org/10.1002/pbc.24938

47. Vrooman LM, Kirov II, Dreyer ZE, et al (2016) Activity and Toxicity of Intravenous Erwinia Asparaginase Following Allergy to E. coli-Derived Asparaginase in Children and Adolescents With Acute Lymphoblastic Leukemia. Pediatr Blood Cancer 63:228–233. https://doi.org/10.1002/pbc.25757

48. Beckett A, Gervais D (2019) What makes a good new therapeutic l-asparaginase? World J Microbiol Biotechnol 35:152. https://doi.org/10.1007/s11274-019-2731-9

49. Sands S, Ladas EJ, Kelly KM, et al (2017) Glutamine for the treatment of vincristine-induced neuropathy in children and adolescents with cancer. Support Care Cancer 25:701–708. https://doi.org/10.1007/s00520-016-3441-6

50. Han Y, Zhang F, Wang J, et al (2016) Application of Glutamine-enriched nutrition therapy in childhood acute lymphoblastic leukemia. Nutrition Journal 15:65. https://doi.org/10.1186/s12937-016-0187-4

51. van Zanten ARH, Hofman Z, Heyland DK (2015) Consequences of the REDOXS and METAPLUS Trials: The End of an Era of Glutamine and Antioxidant Supplementation for Critically Ill Patients? JPEN J Parenter Enteral Nutr 39:890–892. https://doi.org/10.1177/0148607114567201

52. Moe-Byrne T, Brown JV, McGuire W (2016) Glutamine supplementation to prevent morbidity and mortality in preterm infants. Cochrane Database Syst Rev 2016:CD001457. https://doi.org/10.1002/14651858.CD001457.pub6

53. Wernerman J (2015) How to understand the results of studies of glutamine supplementation. Crit Care 19:385. https://doi.org/10.1186/s13054-015-1090-7

54. Van Trimpont M, Schalk AM, Hofkens K, et al (2025) A human-like glutaminase-free asparaginase is highly efficacious in ASNSlow leukemia and solid cancer mouse xenograft models. Cancer Letters 611:217404. https://doi.org/10.1016/j.canlet.2024.217404

55. Ashok A, Doriya K, Rao JV, et al (2019) Microbes Producing L-Asparaginase free of Glutaminase and Urease isolated from Extreme Locations of Antarctic Soil and Moss. Sci Rep 9:1423. https://doi.org/10.1038/s41598-018-38094-1

56. Prihanto AA, Yanti I, Murtazam MA, Jatmiko YD (2020) Optimization of glutaminase-free L-asparaginase production using mangrove endophytic Lysinibacillus fusiformis B27. F1000Res 8:1938. https://doi.org/10.12688/f1000research.21178.2

57. Doriya K, Kumar DS (2016) Isolation and screening of l-asparaginase free of glutaminase and urease from fungal sp. 3 Biotech 6:239. https://doi.org/10.1007/s13205-016-0544-1

58. Ollenschläger G, Roth E, Linkesch W, et al (1988) Asparaginase-induced derangements of glutamine metabolism: the pathogenetic basis for some drug-related side-effects. Eur J Clin Invest 18:512–516. https://doi.org/10.1111/j.1365-2362.1988.tb01049.x

59. Ho PP, Milikin EB, Bobbitt JL, et al (1970) Crystalline L-asparaginase from Escherichia coli B. I. Purification and chemical characterization. J Biol Chem 245:3708–3715

60. Moola ZB, Scawen MD, Atkinson T, Nicholls DJ (1994) Erwinia chrysanthemi L-asparaginase: epitope mapping and production of antigenically modified enzymes. Biochem J 302:921–927. https://doi.org/10.1042/bj3020921

61. Iida K, Li Y, McGrath BC, et al (2007) PERK eIF2 alpha kinase is required to regulate the viability of the exocrine pancreas in mice. BMC Cell Biol 8:38. https://doi.org/10.1186/1471-2121-8-38

62. Burke MJ, Zalewska-Szewczyk B (2022) Hypersensitivity reactions to asparaginase therapy in acute lymphoblastic leukemia: immunology and clinical consequences. Future Oncol 18:1285–1299. https://doi.org/10.2217/fon-2021-1288

63. Vrooman LM, Supko JG, Neuberg DS, et al (2010) Erwinia Asparaginase after Allergy to E coli Asparaginase in Children with Acute Lymphoblastic Leukemia. Pediatr Blood Cancer 54:199–205. https://doi.org/10.1002/pbc.22225

64. Molineux G (2003) Pegylation: engineering improved biopharmaceuticals for oncology. Pharmacotherapy 23:3S-8S. https://doi.org/10.1592/phco.23.9.3s.32886

65. Torres-Obreque K, Meneguetti GP, Custódio D, et al (2019) Production of a novel N-terminal PEGylated crisantaspase. Biotechnology and Applied Biochemistry 66:281–289. https://doi.org/10.1002/bab.1723

66. Xu F, Oruna-Concha M-J, Elmore JS (2016) The use of asparaginase to reduce acrylamide levels in cooked food. Food Chemistry 210:163–171. https://doi.org/10.1016/j.foodchem.2016.04.105

67. Matsumura M, Becktel WJ, Levitt M, Matthews BW (1989) Stabilization of phage T4 lysozyme by engineered disulfide bonds. Proceedings of the National Academy of Sciences 86:6562–6566. https://doi.org/10.1073/pnas.86.17.6562

68. Lee C-W, Wang H-J, Hwang J-K, Tseng C-P (2014) Protein Thermal Stability Enhancement by Designing Salt Bridges: A Combined Computational and Experimental Study. PLoS One 9:e112751. https://doi.org/10.1371/journal.pone.0112751

69. Das R, Gerstein M (2000) The stability of thermophilic proteins: A study based on comprehensive genome comparison. Functional & integrative genomics 1:76–88. https://doi.org/10.1007/s101420000003

70. Feng Y, Liu S, Pang C, et al (2019) Improvement of catalytic efficiency and thermal stability of l-asparaginase from *Bacillus subtilis* 168 through reducing the flexibility of the highly flexible loop at N-terminus. Process Biochemistry 78:42–49. https://doi.org/10.1016/j.procbio.2019.01.001

71. Pritsa AA, Kyriakidis DA (2001) L-asparaginase of Thermus thermophilus: purification, properties and identification of essential amino acids for its catalytic activity. Mol Cell Biochem 216:93–101. https://doi.org/10.1023/a:1011066129771

72. Chohan SM, Rashid N, Sajed M, Imanaka T (2019) Pcal_0970: an extremely thermostable l-asparaginase from Pyrobaculum calidifontis with no detectable glutaminase activity. Folia Microbiol 64:313–320. https://doi.org/10.1007/s12223-018-0656-6

73. Pedroso A, Herrera Belén L, Beltrán JF, et al (2023) In Silico Design of a Chimeric Humanized L-asparaginase. Int J Mol Sci 24:7550. https://doi.org/10.3390/ijms24087550

74. Ln R, Doble M, Rekha VPB, Pulicherla KK (2011) In silico Engineering of L-Asparaginase to Have Reduced Glutaminase Side Activity for Effective Treatment of Acute Lymphoblastic Leukemia. Journal of Pediatric Hematology/Oncology 33:617. https://doi.org/10.1097/MPH.0b013e31822aa4ec

75. Ghadikolaei MS, Asad S, Hassan-Zadeh V (2024) In Silico-Driven Engineering of Halomonas elongata L-Asparaginase: Towards Enhanced Proteolytic Resistance in Lymphoblastic Leukemia. 2024.06.07.597648. http://dx.doi.org/10.2139/ssrn.4871035

76. Lubkowski J, Vanegas J, Chan WK, et al (2020) Mechanism of Catalysis by l -Asparaginase. Biochemistry 59:1927–1945. https://doi.org/10.1021/acs.biochem.0c00116

77. Gesto DS, Cerqueira NMFSA, Fernandes PA, Ramos MJ (2013) Unraveling the enigmatic mechanism of l-asparaginase II with QM/QM calculations. J Am Chem Soc 135:7146–7158. https://doi.org/10.1021/ja310165u

78. Palm GJ, Lubkowski J, Derst C, et al (1996) A covalently bound catalytic intermediate in Escherichia coli asparaginase: Crystal structure of a Thr-89-Val mutant. FEBS Lett 390:211–216. https://doi.org/10.1016/0014-5793(96)00660-6

79. Guimarães AVF, Frota NF, Lourenzoni MR (2021) Molecular dynamics simulations of human L-asparaginase1: Insights into structural determinants of enzymatic activity. Journal of Molecular Graphics and Modelling 109:108007. https://doi.org/10.1016/j.jmgm.2021.108007

80. Brannigan JA, Dodson G, Duggleby HJ, et al (1995) A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. Nature 378:416–419. https://doi.org/10.1038/378416a0

81. Verlet L (1967) Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. Phys Rev 159:98–103. https://doi.org/10.1103/PhysRev.159.98

82. Schlick T (2010) Molecular Modeling and Simulation: An Interdisciplinary Guide: An Interdisciplinary Guide. Springer, New York, NY

83. Ryckaert J-P, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. Journal of Computational Physics 23:327–341. https://doi.org/10.1016/0021-9991(77)90098-5

84. Gallavotti G (2016) Ergodicity: a historical perspective. Equilibrium and Nonequilibrium. EPJ H 41:181–259. https://doi.org/10.1140/epjh/e2016-70030-8

85. Prigogine I, Nicolis G (1985) Self-Organisation in Nonequilibrium Systems: Towards A Dynamics of Complexity. In: Hazewinkel M, Jurkovich R, Paelinck JHP (eds) Bifurcation Analysis: Principles, Applications and Synthesis. Springer Netherlands, Dordrecht, pp 3–12

86. Groot SRD, Mazur P (2013) Non-Equilibrium Thermodynamics. Courier Corporation

87. Mey ASJS, Allen BK, Bruce Macdonald HE, et al (2020) Best Practices for Alchemical Free Energy Calculations [Article v1.0]. LiveCoMS 2:. https://doi.org/10.33011/livecoms.2.1.18378

88. Steinbrecher T, Joung IS, Case DA (2011) Soft-Core Potentials in Thermodynamic Integration. Comparing One- and Two-Step Transformations. J Comput Chem 32:3253–3263. https://doi.org/10.1002/jcc.21909

89. Duboué-Dijon E, Hénin J (2021) Building intuition for binding free energy calculations: Bound state definition, restraints, and symmetry. J Chem Phys 154:204101. https://doi.org/10.1063/5.0046853

90. Kästner J (2011) Umbrella sampling. WIREs Computational Molecular Science 1:932–942. https://doi.org/10.1002/wcms.66

91. Kumar S, Bouzida D, Swendsen RH, et al (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. J Comput Chem 13:1011–1021. https://doi.org/10.1002/jcc.540130812

92. Kollman PA, Massova I, Reyes C, et al (2000) Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. Acc Chem Res 33:889–897. https://doi.org/10.1021/ar000033j

93. Roux B, Chipot C (2024) Editorial Guidelines for Computational Studies of Ligand Binding Using MM/PBSA and MM/GBSA Approximations Wisely. J Phys Chem B 128:12027–12029. https://doi.org/10.1021/acs.jpcb.4c06614

94. Fried SD, Boxer SG (2017) Electric Fields and Enzyme Catalysis. Annu Rev Biochem 86:387–415. https://doi.org/10.1146/annurev-biochem-061516-044432

95. Jabeen H, Beer M, Spencer J, et al (2024) Electric Fields Are a Key Determinant of Carbapenemase Activity in Class A β-Lactamases. ACS Catal 14:7166–7172. https://doi.org/10.1021/acscatal.3c05302

96. Ruiz-Pernía JJ, Świderek K, Bertran J, et al (2024) Electrostatics as a Guiding Principle in Understanding and Designing Enzymes. J Chem Theory Comput 20:1783–1795. https://doi.org/10.1021/acs.jctc.3c01395

97. Dubey KD, Stuyver T, Shaik S (2022) Local Electric Fields: From Enzyme Catalysis to Synthetic Catalyst Design. J Phys Chem B 126:10285–10294. https://doi.org/10.1021/acs.jpcb.2c06422

98. Polêto MD, Lemkul JA (2022) TUPÃ: Electric field analyses for molecular simulations. Journal of Computational Chemistry 43:1113–1119. https://doi.org/10.1002/jcc.26873

99. Leven I, Hao H, Tan S, et al (2021) Recent Advances for Improving the Accuracy, Transferability, and Efficiency of Reactive Force Fields. J Chem Theory Comput 17:3237–3251. https://doi.org/10.1021/acs.jctc.1c00118

100. Warshel A, Levitt M (1976) Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. Journal of Molecular Biology 103:227–249. https://doi.org/10.1016/0022-2836(76)90311-9

101. Groenhof G (2013) Introduction to QM/MM Simulations. In: Monticelli L, Salonen E (eds) Biomolecular Simulations: Methods and Protocols. Humana Press, Totowa, NJ, pp 43–66

102. Ryde U (2016) QM/MM Calculations on Proteins. Methods Enzymol 577:119–158. https://doi.org/10.1016/bs.mie.2016.05.014

103. van der Kamp MW, Mulholland AJ (2013) Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. Biochemistry 52:2708–2728. https://doi.org/10.1021/bi400215w

104. Cao L, Ryde U (2018) On the Difference Between Additive and Subtractive QM/MM Calculations. Front Chem 6:89. https://doi.org/10.3389/fchem.2018.00089

105. Field MJ, Bash PA, Karplus M (1990) A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. Journal of Computational Chemistry 11:700–733. https://doi.org/10.1002/jcc.540110605

106. Warshel A, Russell ST (1984) Calculations of electrostatic interactions in biological systems and in solutions. Quarterly Reviews of Biophysics 17:283–422. https://doi.org/10.1017/S0033583500005333

107. Lee FS, Chu ZT, Warshel A (1993) Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIX programs. Journal of Computational Chemistry 14:161–185. https://doi.org/10.1002/jcc.540140205

108. Yu H, van Gunsteren W (2005) Accounting for polarization in molecular simulation. https://doi.org/10.1016/j.cpc.2005.01.022']

109. Bondanza M, Nottoli M, Cupellini L, et al (2020) Polarizable embedding QM/MM: the future gold standard for complex (bio)systems? Phys Chem Chem Phys 22:14433–14448. https://doi.org/10.1039/D0CP02119A

110. Singh UC, Kollman PA (1986) A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the CH3Cl + Cl− exchange reaction and gas phase protonation of polyethers. Journal of Computational Chemistry 7:718–730. https://doi.org/10.1002/jcc.540070604

111. Maseras F, Morokuma K (1995) IMOMM: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures

and transition states. Journal of Computational Chemistry 16:1170–1179. https://doi.org/10.1002/jcc.540160911

112. Woo TK, Cavallo L, Ziegler T (1998) Implementation of the IMOMM methodology for performing combined QM/MM molecular dynamics simulations and frequency calculations. THEORETICAL CHEMISTRY ACCOUNTS 100:307–313. https://doi.org/10.1007/s002140050391

113. Eichler U, Kölmel CM, Sauer J (1997) Combining ab initio techniques with analytical potential functions for structure predictions of large systems: Method and application to crystalline silica polymorphs. Journal of Computational Chemistry 18:463–477. https://doi.org/10.1002/(SICI)1096-987X(199703)18:4<463::AID-JCC2>3.0.CO;2-R

114. Lee C, Yang W, Parr RG (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37:785–789. https://doi.org/10.1103/PhysRevB.37.785

115. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648–5652. https://doi.org/10.1063/1.464913

116. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ (1994) Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. J Phys Chem 98:11623–11627. https://doi.org/10.1021/j100096a001

117. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. https://cdnsciencepub.com/doi/10.1139/p80-159. Accessed 16 Dec 2024

118. Zhao Y, Truhlar DG (2008) The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other function. Theor Chem Acc 120:215–241. https://doi.org/10.1007/s00214-007-0310-x

119. Perdew JP, Burke K, Ernzerhof M (1997) Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. Phys Rev Lett 78:1396–1396. https://doi.org/10.1103/PhysRevLett.78.1396

120. Winget P, Horn AHC, Selçuki C, et al (2003) AM1* parameters for phosphorus, sulfur and chlorine. J Mol Model 9:408–414. https://doi.org/10.1007/s00894-003-0156-7

121. Winget P, Clark T (2005) AM1* parameters for aluminum, silicon, titanium and zirconium. J Mol Model 11:439–456. https://doi.org/10.1007/s00894-005-0236-y

122. Kayi H, Clark T (2007) AM1* parameters for copper and zinc. J Mol Model 13:965–979. https://doi.org/10.1007/s00894-007-0214-7

123. Gaus M, Cui Q, Elstner M (2011) DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). J Chem Theory Comput 7:931–948. https://doi.org/10.1021/ct100684s

124. Bannwarth C, Ehlert S, Grimme S (2019) GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. J Chem Theory Comput 15:1652–1671. https://doi.org/10.1021/acs.jctc.8b01176

125. Lu L, Hu H, Hou H, Wang B (2013) An improved B3LYP method in the calculation of organic thermochemistry and reactivity. Computational and Theoretical Chemistry 1015:64–71. https://doi.org/10.1016/j.comptc.2013.04.009

126. Copper Oxidation/Reduction in Water and Protein: Studies with DFTB3/MM and VALBOND Molecular Dynamics Simulations | The Journal of Physical Chemistry B. https://pubs.acs.org/doi/10.1021/acs.jpcb.5b09656. Accessed 24 Feb 2025

127. Kansari M, Eichinger L, Kubař T (2023) Extended-sampling QM/MM simulation of biochemical reactions involving P–N bonds. Physical Chemistry Chemical Physics 25:9824–9836. https://doi.org/10.1039/D2CP05890A

128. Christensen AS, Elstner M, Cui Q (2015) Improving intermolecular interactions in DFTB3 using extended polarization from chemical-potential equalization. J Chem Phys 143:084123. https://doi.org/10.1063/1.4929335

129. Roston D, Demapan D, Cui Q (2019) Extensive free-energy simulations identify water as the base in nucleotide addition by DNA polymerase. Proceedings of the National Academy of Sciences 116:25048–25056. https://doi.org/10.1073/pnas.1914613116

130. Nurhuda M, Perry CC, Addicoat MA (2022) Performance of GFN1-xTB for periodic optimization of metal organic frameworks. Phys Chem Chem Phys 24:10906–10914. https://doi.org/10.1039/D2CP00184E

131. Pracht P, Grant DF, Grimme S (2020) Comprehensive Assessment of GFN Tight-Binding and Composite Density Functional Theory Methods for Calculating Gas-Phase Infrared Spectra. J Chem Theory Comput 16:7044–7060. https://doi.org/10.1021/acs.jctc.0c00877

132. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules | Journal of the American Chemical Society. https://pubs.acs.org/doi/10.1021/ja00124a002. Accessed 24 Feb 2025

133. Case DA, Ben-Shalom IY, Brozell SR, et al (2024) AMBER 2024

134. MacKerell Jr. AD, Banavali NK (2000) All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. Journal of Computational Chemistry 21:105–120. https://doi.org/10.1002/(SICI)1096-987X(20000130)21:2<105::AID-JCC3>3.0.CO;2-P

135. MacKerell AD Jr, Bashford D, Bellott M, et al (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J Phys Chem B 102:. https://pubs.acs.org/doi/10.1021/jp973084f

136. Scott WRP, Hunenberger PH, Tironi IG, et al (1999) The GROMOS biomolecular simulation program package. The Journal of Physical Chemistry A 103:3596–3607. https://pubs.acs.org/doi/10.1021/jp984217f

137. Chipot C, Ángyán J (2005) Continuing challenges in the parametrization of intermolecular force fields. Towards an accurate description of electro…. N J Chem 3:411–420. https://doi.org/10.1039/B414280M

138. Jorgensen WL (2002) OPLS Force Fields. In: Encyclopedia of Computational Chemistry. John Wiley & Sons, Ltd

139. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J Am Chem Soc 118:11225–11236. https://doi.org/10.1021/ja9621760

140. Maier JA, Martinez C, Kasavajhala K, et al (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput 11:3696–3713. https://doi.org/10.1021/acs.jctc.5b00255

141. Jorgensen WL, Chandrasekhar J, Madura JD, et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935. https://doi.org/10.1063/1.445869

142. Love O, Pacheco Lima MC, Clark C, et al (2023) Evaluating the accuracy of the AMBER protein force fields in modeling dihydrofolate reductase structures: misbalance in the conformational arrangements of the flexible loop domains. Journal of Biomolecular Structure and Dynamics 41:5946–5960. https://doi.org/10.1080/07391102.2022.2098823

143. Peters B (2017) Chapter 20 - Reaction coordinates and mechanisms. In: Peters B (ed) Reaction Rate Theory and Rare Events Simulations. Elsevier, Amsterdam, pp 539–571

144. Barducci A, Bonomi M, Parrinello M (2011) Metadynamics. WIREs Computational Molecular Science 1:826–843. https://doi.org/10.1002/wcms.31

145. Eric Vanden-Eijnden and Maddalena Venturoli. Revisiting the nite temperature string method for the calculation of reaction tubes and free energies. The Journal of Chemical Physics, 130(19):194103, 2009.

146. Maragliano L, Roux B, Vanden-Eijnden E (2014) Comparison between Mean Forces and Swarms-of-Trajectories String Methods. J Chem Theory Comput 10:524–533. https://doi.org/10.1021/ct400606c

147. Zinovjev K, Tuñón I (2017) Adaptive Finite Temperature String Method in Collective Variables. J Phys Chem A 121:9764–9772. https://doi.org/10.1021/acs.jpca.7b10842

148. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G (2006) String method in collective variables: Minimum free energy paths and isocommittor surfaces. The Journal of Chemical Physics 125:024106. https://doi.org/10.1063/1.2212942

149. Sugita Y, Kitao A, Okamoto Y (2000) Multidimensional replicaexchange method for free-energy calculations. The Journal of Chemical Physics, 113:6042–6051. https://doi.org/doi.org/10.1063/1.1308516

150. Anson ML, Mirsky AE (1930) Protein Coagulation and its Reversal: The Preparation of Insoluble Globulin, Soluble Globulin and Heme. J Gen Physiol 13:469–476. https://doi.org/10.1085/jgp.13.4.469

151. Anfinsen CB, Haber E, Sela M, White FH (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci U S A 47:1309–1314. https://doi.org/10.1073/pnas.47.9.1309

152. Anfinsen CB (1973) Principles that Govern the Folding of Protein Chains. Science 181:223–230. https://doi.org/10.1126/science.181.4096.223

153. Jumper J, Evans R, Pritzel A, et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2

154. Baek M, DiMaio F, Anishchenko I, et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–876. https://doi.org/10.1126/science.abj8754

155. Meier A, Söding J (2015) Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. PLoS Comput Biol 11:e1004343. https://doi.org/10.1371/journal.pcbi.1004343

156. Kaczanowski S, Zielenkiewicz P (2010) Why similar protein sequences encode similar three-dimensional structures? Theor Chem Acc 125:643–650. https://doi.org/10.1007/s00214-009-0656-3

157. Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389

158. Dhingra S, Sowdhamini R, Cadet F, Offmann B (2020) A glance into the evolution of template-free protein structure prediction methodologies. Biochimie 175:85–92. https://doi.org/10.1016/j.biochi.2020.04.026

159. Handl J, Knowles J, Vernon R, et al (2012) The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. Proteins 80:490–504. https://doi.org/10.1002/prot.23215

160. Leaver-Fay A, Tyka M, Lewis SM, et al (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487:545–574. https://doi.org/10.1016/B978-0-12-381270-4.00019-6

161. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80:1715–1735. https://doi.org/10.1002/prot.24065

162. AlQuraishi M, Sorger PK (2021) Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. Nat Methods 18:1169–1180. https://doi.org/10.1038/s41592-021-01283-4

163. Qin Y, Chen Z, Peng Y, et al (2024) Deep learning methods for protein structure prediction. MedComm – Future Medicine 3:e96. https://doi.org/10.1002/mef2.96

164. Pearce R, Zhang Y (2021) Deep learning techniques have significantly impacted protein structure prediction and protein design. Curr Opin Struct Biol 68:194–207. https://doi.org/10.1016/j.sbi.2021.01.007

165. Khan S, Khan M, Iqbal N, et al (2023) Enhancing Sumoylation Site Prediction: A Deep Neural Network with Discriminative Features. Life 13:2153. https://doi.org/10.3390/life13112153

166. Lauko A, Pellock SJ, Anischanka I, et al (2024) Computational design of serine hydrolases. bioRxiv 2024.08.29.610411. https://doi.org/10.1101/2024.08.29.610411

167. Torres SV, Valle MB, Mackessy SP, et al (2024) De novo designed proteins neutralize lethal snake venom toxins. Res Sq rs.3.rs-4402792. https://doi.org/10.21203/rs.3.rs-4402792/v1

168. Sumida KH, Núñez-Franco R, Kalvet I, et al (2024) Improving Protein Expression, Stability, and Function with ProteinMPNN. J Am Chem Soc 146:2054–2061. https://doi.org/10.1021/jacs.3c10941

169. Yeh AH-W, Norn C, Kipnis Y, et al (2023) De novo design of luciferases using deep learning. Nature 614:774–780. https://doi.org/10.1038/s41586-023-05696-3

170. Rettie SA, Juergens D, Adebomi V, et al (2024) Accurate de novo design of high-affinity protein binding macrocycles using deep learning. bioRxiv 2024.11.18.622547. https://doi.org/10.1101/2024.11.18.622547

171. Röthlisberger D, Khersonsky O, Wollacott AM, et al (2008) Kemp elimination catalysts by computational enzyme design. Nature 453:190–195. https://doi.org/10.1038/nature06879

172. Siegel JB, Zanghellini A, Lovick HM, et al (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329:309–313. https://doi.org/10.1126/science.1190239

173. Marcos E, Silva D-A (2018) Essentials of de novo protein design: Methods and applications. WIREs Computational Molecular Science 8:e1374. https://doi.org/10.1002/wcms.1374

174. Scaffolding protein functional sites using deep learning | Science. https://www.science.org/doi/10.1126/science.abn2100. Accessed 11 Feb 2025

175. Watson JL, Juergens D, Bennett NR, et al (2023) De novo design of protein structure and function with RFdiffusion. Nature 620:1089–1100. https://doi.org/10.1038/s41586-023-06415-8

176. Saharia C, Chan W, Saxena S, et al (2022) Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. https://doi.org/10.48550/arXiv.2205.11487

177. Norn C, Wicky BIM, Juergens D, et al (2021) Protein sequence design by conformational landscape optimization. Proceedings of the National Academy of Sciences 118:e2017228118. https://doi.org/10.1073/pnas.2017228118

178. Dauparas J, Anishchenko I, Bennett N, et al (2022) Robust deep learning–based protein sequence design using ProteinMPNN. Science 378:49–56. https://doi.org/10.1126/science.add2187

179. Dauparas J, Lee GR, Pecoraro R, et al (2023) Atomic context-conditioned protein sequence design using LigandMPNN. Preprint 2023.12.22.573103. https://doi.org/10.1101/2023.12.22.573103

180. Nomme J, Su Y, Lavie A (2014) Elucidation of the specific function of the conserved threonine triad responsible for human l-Asparaginase autocleavage and substrate hydrolysis. J Mol Biol 426:2471–2485. https://doi.org/10.1016/j.jmb.2014.04.016

181. Andjelkovic M, Zinovjev K, Ramos-Guzmán CA, et al (2023) Elucidation of the Active Form and Reaction Mechanism in Human Asparaginase Type III Using Multiscale Simulations. J Chem Inf Model 63:5676–5688. https://doi.org/10.1021/acs.jcim.3c00900

182. Smith PK, Gorham AT, Smith ERB (1942) Thermodynamic Properties of Solutions of Amino Acids and Related Substances. J Biol Chem 144:737–745. https://doi.org/10.1016/s0021-9258(18)72499-x

183. Hanson KR, Havir EA (1972) 3 The Enzymic Elimination of Ammonia. In: Boyer PD (ed) The Enzymes. pp 75–166

184. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH (2011) PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput 7:525–537. https://doi.org/10.1021/ct100578z

185. Le Grand S, Götz AW, Walker RC (2013) SPFP: Speed without compromise - A mixed precision model for GPU accelerated molecular dynamics simulations. Comput Phys Commun 184:374–380. https://doi.org/10.1016/j.cpc.2012.09.022

186. Götz AW, Williamson MJ, Xu D, et al (2012) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. generalized born. J Chem Theory Comput 8:1542–1555. https://doi.org/10.1021/ct200909j

187. Horn AHC (2014) A consistent force field parameter set for zwitterionic amino acid residues. J Mol Model 20:1–14. https://doi.org/10.1007/s00894-014-2478-z

188. Case DA, Aktulga HM, Belfon K, et al (2023) AmberTools. J Chem Inf Model 63:6183–6191. https://doi.org/10.1021/acs.jcim.3c01153

189. Miyamoto S, Kollman PA (1992) Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. J Comput Chem 13:952–962. https://doi.org/10.1002/jcc.540130805

190. Essmann U, Perera L, Berkowitz ML, et al (1995) A smooth particle mesh Ewald method. The Journal of Chemical Physics 103:8577–8593. https://doi.org/10.1063/1.470117

191. He X, Liu S, Lee TS, et al (2020) Fast, Accurate, and Reliable Protocols for Routine Calculations of Protein-Ligand Binding Affinities in Drug Design Projects Using AMBER GPU-TI with ff14SB/GAFF. ACS Omega 5:4611–4619. https://doi.org/10.1021/acsomega.9b04233

192. Grimme S, Antony J, Ehrlich S, Krieg H (2010) A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. J Chem Phys 132:154104(1)-154104(19). https://doi.org/10.1063/1.3382344

193. Zinovjev K Adaptive String Method. https://github.com/kzinovjev/string-amber

194. Gaussian 16, Revision C.01, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.;

Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian, Inc., Wallingford CT, 2016.

195. Grossfield A WHAM: the weighted histogram analysis method. http://membrane.urmc.rochester.edu/wordpress/?page_id=126 (accessed 2023-05-04)

196. Andjelkovic M, Zinovjev K, Ruiz-Pernía JJ, Tuñón I (2024) Unveiling the Catalytic Mechanism and Conformational Dynamics of Guinea Pig L-Asparaginase Type 1 for Leukemia Drug Design. ChemRxiv. https://doi.org/10.26434/chemrxiv-2024-dtcbv

197. Zinovjev K, Guénon P, Ramos-Guzmán CA, et al (2024) Activation and friction in enzymatic loop opening and closing dynamics. Nat Commun 15:2490. https://doi.org/10.1038/s41467-024-46723-9

198. Evans R, O'Neill M, Pritzel A, et al (2022) Protein complex prediction with AlphaFold-Multimer. 2021.10.04.463034

199. Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9:173–175. https://doi.org/10.1038/nmeth.1818

200. (2016) Schrödinger Release 2016-3: Maestro

201. Bayly CI, Cieplak P, Cornell WD, Kollman PA (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. J Phys Chem 97:10269–10280. https://doi.org/10.1021/j100142a004

202. Miller BRI, McGee TDJr, Swails JM, et al (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. J Chem Theory Comput 8:3314–3321. https://doi.org/10.1021/ct300418h

203. Gohlke H, Kiel C, Case DA (2003) Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. J Mol Biol 330:891–913. https://doi.org/10.1016/s0022-2836(03)00610-7

204. Jurtz V, Paul S, Andreatta M, et al (2017) NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. The Journal of Immunology 199:3360–3368. https://doi.org/10.4049/jimmunol.1700893

205. Fernandez CA, Smith C, Yang W, et al (2014) HLA-DRB1*07:01 is associated with a higher risk of asparaginase allergies. Blood 124:1266–1276. https://doi.org/10.1182/blood-2014-03-563742

206. Chacos N, Capdevila J, Falck JR, et al (1983) The reaction of arachidonic acid epoxides (epoxyeicosatrienoic acids) with a cytosolic epoxide hydrolase. Arch Biochem Biophys 223:639–648. https://doi.org/10.1016/0003-9861(83)90628-8

207. Yu Z, Xu F, Huse LM, et al (2000) Soluble epoxide hydrolase regulates hydrolysis of vasoactive epoxyeicosatrienoic acids. Circ Res 87:992–998. https://doi.org/10.1161/01.res.87.11.992

208. Lonsdale R, Harvey JN, Mulholland AJ (2010) Compound I reactivity defines alkene oxidation selectivity in cytochrome P450cam. J Phys Chem B 114:1156–1162. https://doi.org/10.1021/jp910127j

209. Reetz MT, Bocola M, Wang L-W, et al (2009) Directed evolution of an enantioselective epoxide hydrolase: uncovering the source of enantioselectivity at each evolutionary stage. J Am Chem Soc 131:7334–7343. https://doi.org/10.1021/ja809673d

210. Fretland AJ, Omiecinski CJ (2000) Epoxide hydrolases: biochemistry and molecular biology. Chemico-Biological Interactions 129:41–59. https://doi.org/10.1016/S0009-2797(00)00197-6

211. Mowbray SL, Elfström LT, Ahlgren KM, et al X-ray structure of potato epoxide hydrolase sheds light on substrate specificity in plant enzymes - Mowbray - 2006 - Protein Science - Wiley Online Library. Protein Sci 7:1628–1637. https://doi.org/10.1110/ps.051792106

212. Arand M, Cronin A, Adamska M, Oesch F (2005) Epoxide Hydrolases: Structure, Function, Mechanism, and Assay. In: Methods in Enzymology. Academic Press, pp 569–588

213. Arand M, Muller F, Mecky A, et al (1999) Catalytic triad of microsomal epoxide hydrolase: replacement of Glu404 with Asp leads to a strongly increased turnover rate - PubMed. Biochem J 1:37–43. https://doi.org/10.1016/S0009-2797(00)00197-6

214. Hopmann KH, Himo F Insights into the Reaction Mechanism of Soluble Epoxide Hydrolase from Theoretical Active Site Mutants. The Journal of Physical Chemistry B 110:21299–21310. https://doi.org/10.1021/jp063830t

215. Jones PD, Wolf NM, Morisseau C, et al (2005) Fluorescent substrates for soluble epoxide hydrolase and application to inhibition studies. Analytical Biochemistry 343:66–75. https://doi.org/10.1016/j.ab.2005.03.041

216. Tyka MD, Keedy DA, André I, et al (2011) Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol 405:607–618. https://doi.org/10.1016/j.jmb.2010.11.008

217. Anishchenko I, Kipnis Y, Kalvet I, et al (2024) Modeling protein-small molecule conformational ensembles with ChemNet. 2024.09.25.614868. https://doi.org/10.1101/2024.09.25.614868

218. Elfström LT, Widersten M (2005) Catalysis of potato epoxide hydrolase, StEH1. Biochem J 390:633–640. https://doi.org/10.1042/BJ20050526

219. Richter F, Leaver-Fay A, Khare SD, et al (2011) De Novo Enzyme Design Using Rosetta3 | PLOS ONE. PLoS ONE 6:e19230. https://doi.org/10.1371/journal.pone.0019230

220. Zanghellini A, Jiang L, Wollacott AM, et al (2006) New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 15:2785–2794. https://doi.org/10.1110/ps.062353106

221. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. Journal of Molecular Graphics and Modelling 25:247–260. https://doi.org/10.1016/j.jmgm.2005.12.005

222. Dang B, Mravic M, Hu H, et al (2019) SNAC-tag for sequence-specific chemical protein cleavage. Nat Methods 16:319–322. https://doi.org/10.1038/s41592-019-0357-3

223. Lander ES, Linton LM, Birren B, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. https://doi.org/10.1038/35057062

224. Expasy - ProtParam. https://web.expasy.org/protparam/. Accessed 2 Feb 2025

225. Prism - GraphPad. https://www.graphpad.com/prism. Accessed 2 Feb 2025

226. Shaw KL, Scholtz JM, Pace CN, Grimsley RG (2009) Determining the Conformational Stability of a Protein Using Urea Denaturation Curves. In: Shriver JW (ed) Protein Structure, Stability, and Interactions. Humana Press, Totowa, NJ, pp 41–55

227. Wriston JC (1970) [98] Asparaginase. In: Methods in Enzymology. Academic Press, pp 732–742

228. Zwanzig RW (1954) High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. The Journal of Chemical Physics 22:1420–1426. https://doi.org/10.1063/1.1740409