**Henrique Daniel Figueiredo Carvalho**

Mestre em Bioquímica

# Computational design and experimental characterization of metallopeptides as proteases for bioengineering applications

Dissertação para obtenção do Grau de Doutor em Bioengenharia (MIT Portugal)

Orientador: Olga Iranzo, Doutora., Aix-Marseille Université

Co-orientadores: Ana Cecília Roque, Prof. Doutora., Univ. NOVA de Lisboa

Ricardo J. F. Branco, Doutor, Univ. NOVA de Lisboa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro 2017

**Computational design and experimental characterization of metallopeptides as proteases for bioengineering applications**

# Agradecimentos

Este trabalho marca o final de uma importante etapa no meu percurso académico e pessoal e portanto gostaria de agradecer a todos aqueles que de alguma forma estiveram envolvidos. À minha orientadora Doutora Olga Iranzo e co-orientadores Professora Doutora Ana Cecília A. Roque e Ricardo J. F. Branco por propocionarem as condições necessárias à realização deste trabalho nos grupo de Engenharia Biomolecular (UCIBIO/REQUIMTE, Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa) e no grupo BiosCiences (Institut des Sciences Moléculaires de Marseille UMR CNRS 7313, Aix-Marseille Université), pela supervisão e apoio crítico na realização tarefas e discussão de resultados. Ao Doutor João Galamba Correia por me ter recebido no grupo de Ciências Radiofarmacológicas (Centro de Ciências e Tecnologias Nucleares, Instituto Superior Técnico) para produção química de péptidos. Ao Doutor Manolis Matzapetakis por me ter recebibo do Laboratório de RMN Biomolecular (Instituto de Tecnologia Química e Biológica António Xavier, Universidade NOVA de Lisboa) para realização de experiências de RMN. À Professora Doutora Rita Delgado e restantes membros do grupo de Química de Coordenação e Supramolecular (Instituto de Tecnologia Química e Biológica António Xavier, Universidade NOVA de Lisboa) pela disponibilização de recursos e auxílio na realização de tarefas. Aos meus colegas e minhas colegas do grupo de Engenharia Biomolecular pelo companheirismo e convivência, assim como pela troca de ideias e auxílio na resolução de problemas. À minha família e amigos, pelo seu apoio incondicional e fundamental para que este trabalho pudesse ser realizado. Gostatia por ultimo de agradecer ao Programa Doutoral em Bioengenharia MIT/Portugal pela formação e à Fundação para a Ciência e Tecnologia pelo financiamento através da bolsa SFRH/BD/90644/2012 e ERA-IB-2.

*“The map is not the territory”*

Alfred Korzybski

# Resumo

Enzimas são versáteis catalisadoras presentes em sistemas biológicos e com elevado potencial tecnológico. Metaloproteases de zinco são uma classe de enzimas com aplicação corrente na indústria *e.g.* alimentar, biofarmacêutica e detergentes. A versatilidade química de diferentes metais combinada com modificações na sequência de metaloenzimas podem ser exploradas de forma a aumentar a sua robustez e leque de aplicações biológicas e tecnológicas,

Neste trabalho, novos metaloproteases de zinco foram desenhados computacionalmente de forma a testar se atividade proteolítica pode ser reproduzida em proteínas pequenas adaptadas para bioengenharia. Análise de aspetos estruturais/dinâmicos de metaloproteases permitiram identificar interações cataliticamente relevantes. Modelos do centro ativo de zinco foram desenvolvidos para examinar com o software Rosetta um conjunto de 43 pequenas estruturas quanto ao seu potencial em recapitular a função nativa de enzimas. Duas estruturas foram selecionadas para design e caracterização experimental, nomeadamente o "zinc-finger" 2 da proteína Sp1 e o subdomínio cabeça da vilina. Embora coordenação com o metal tenha sido alcançada (constantes de afinidade $K_{ZnP,app}$ na ordem $10^5\ M^{-1}$), as estruturas apresentam baixa estabilidade (temperatura de desnaturação inferior a 50 °C), refletindo perturbações provavelmente causadas por 4-10 modificações de sequência. Os metalopéptidos apresentam actividade catalítica para ésteres semelhante aos valores de literatura obtidos para outras pequenas estruturas (constantes de segunda ordem $k_2$ na ordem $10^{-1}\ M^{-1}s^{-1}$).

A metodologia desenvolvida neste trabalho foi bem sucedida em desenhar metalopéptidos catalíticos, embora a atividade alvo de metaloprotease não tenha sido alcançada. Simulações de dinâmica molecular na escala de microsegundo foram usadas posteriormente para detectar falhas nos designs relacionadas com elevada flexibilidade estrutural. Este trabalho contribui para o melhoramento de métodos de design computacional de enzimas, ao demonstrar a necessidade de considerar aspectos dinâmicos dos designs em escalas de tempo maiores, e no desenvolvimento de métodos rápidos para classificar e avaliar um vasto leque de potenciais biocatalisadores.

**Termos-chave**: design de enzimas; biocatalisadores: bioengenharia; péptidos; metaloproteases

# Abstract

Enzymes are highly versatile catalysts present in biological systems and with high technological potential. Zinc metalloproteases are a major class of enzymes currently being employed in *e.g.* food, detergent, biopharmaceutical industries. In order to increase their robustness and range of biological and technological applications, metalloenzymes can be redesigned by exploring the chemical versatility of different metals along with protein sequence modifications.

In this work, the computational design of new zinc metalloproteases was approached to test if proteolytic activity can be recapitulated in small scaffolds tailored for bioengineering applications. Structural and dynamical aspects of metalloproteases were first addressed to identify catalytically-relevant interactions. Models of the active site were developed to screen a set of 43 small scaffolds with the Rosetta software for their ability to recapitulate the native enzyme functionality. Two candidate scaffolds were selected for enzyme design and experimental characterization, namely the Sp1 zinc finger 2 and the villin headpiece subdomain. While metal coordination was achieved (binding constants $K_{ZnP,app}$ in the $10^5$ $M^{-1}$ range), the scaffolds presented low stabilities (thermal unfolding bellow 50 °C) most likely due to perturbations introduced by the 4 to 10 sequence modifications. The metallopeptides presented catalytic activity towards ester substrates within the range of values found for other small scaffolds in the literature (second-order rate constants $k_2$ in the $10^{-1}$ $M^{-1}s^{-1}$ range).

The design approach developed in this work was successful in achieving catalytically-active metallopeptides, although target metalloprotease activity could not be achieved. Molecular dynamics simulations in microsecond regimes were subsequently used to detect design flaws related with high scaffold flexibility. This work contributes to the improvement of the computational enzyme design approaches by pointing out the need for a dynamical treatment of the designs in longer time-scales, and through the development of fast methods to rank and evaluate a large number of potential biocatalysts.

**Keywords:** enzyme design; biocatalysts; bioengineering; peptides; metalloproteases

# Table of Contents

# Index of Figures

# Index of Tables

# List of Abbreviations

| | |
|---|---|
| 4-nP | 4-nitrophenol |
| 4-nPA | 4-nitrophenyl acetate |
| ACN | Acetonitrile |
| AS | Active Site |
| Boc | *t*-butyloxycarbonyl |
| CaDa | Cationic Dummy Atom |
| CAPS | N-cyclohexyl-3-aminopropanesulfonic acid |
| CDE | Computational Enzyme Design |
| CD | Circular Dichroism |
| CHES | N-Cyclohexyl-2-aminoethanesulfonic acid |
| DCM | Dichloromethane |
| DE | Designed Enzyme |
| DFT | Density Functional Theory |
| DIEA | N,N-diisopropylethylamine |
| DMF | Dimethylformamide |
| EDTA | Ethylenediaminetetraacetic acid |
| ESI | Electrospay Ionisation |
| Fmoc | 9-fluorenylmethyloxycarbonyl |
| HBTU | 2-(1H-Benzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate |
| HEPES | 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid |
| HOBt | 1-Hydroxybenzotriazole |
| HP35 | Human villin headpiece C-terminal subdomain |
| HPLC | High-Pressure Liquid Chromatography |
| MALDI | Matrix-assisted laser desorption/ionization |
| MBHA | 4-methylbenzhydrylamine |
| MD | Molecular Dynamics |
| MM | Molecular Mechanics |
| MMP | Matrix Metalloprotease |
| MP | Metalloprotease |
| MS | Mass spectrometry |
| NMP | N-methyl-2-pyrrolidone |
| NMR | Nuclear Magnetic Resonance |
| PDB | Protein Data Bank |

| | |
|---|---|
| pNA | *p*-nitroaniline |
| QM/MM | Quantum-Mechanics/Molecular-Mechanics |
| S.E. | Standard error |
| SCOPe | Structural Classification Of Proteins - extended |
| Sp1f2 | Finger 2 of human Sp1 transcription factor |
| SPPS | Solid-phase peptide synthesis |
| TFA | Trifluoroacetic acid |
| TFE | 2,2,2-triflouroethanol |
| TIS | 1,2-ethanedithiol, triisopropylsilane |
| TOF | Time-of-flight |
| TRIS | Tris(hydroxymethyl)aminomethane |
| TS | Transition-state |
| TSA | Transition-state analogue |
| UM | Unique Match |
| UV | Ulta-violet |
| X-pNA | Amino acid (X) p-nitroanilide |
| ZBG | Zinc Binding Group |
| ZF | Zinc Finger |
| Zi | 2-carboxy-2′-hydroxy-5′-sulfoformazylbenzene |

1

# 1. General Introduction

## 1.1 Computational Enzyme Design

Enzymes are highly specific and active protein catalysts found in biological systems. The versatile chemical properties of these catalysts find use in many fields of research and industry by allowing the creation of biologically-active or valuable chemicals.[1,2] Enzymes are a result of millions of years of evolution, where fitness to natural environments does not necessarily translate into direct applicability and efficiency in industrial processes, in particular for non-natural occurring transformations. The ability to design enzymes for a given chemical reaction or to optimize its properties for process-based applications holds great potential for industry and biotechnology and has been the subject of multi-disciplinary efforts.

Metalloenzymes have been common targets in rational design of protein biocatalysts over the last decades.[3,4] The chemical versatility of metal ions can be explored along with protein sequence modifications to increase the range of biological and technological applications of these biocatalysts. Approaches used include metal substitution, inclusion of unnatural amino acids or co-factor replacement for the redesign of native metalloenzymes or the design of novel metalloenzymes from non-active scaffolds. Computational enzyme design (CED) approaches have also contributed significantly to the development of metalloenzymes by bringing together topics of chemistry, biochemistry and biophysics with computational methods to design tailored proteins sequences.[5–8] Recent examples include design of proteins with artificial organometallic co-factors [9,10] and iron-cluster centres [11]. Design of Zn(II) metalloenzymes has been extensively addressed [12] since this metal ion plays a catalytic role in all classes of enzymes.

CED methodologies allow to redesign metal centres in native systems or to *de novo* design of alternative scaffolds with metal centres to obtain active biocatalysts from scratch. The ROSETTA (Rosetta) software suite [13] have shown to be a particularly suitable tool to develop new metalloenzymes, as it will be discussed throughout this work in the context of design of Zn(II) proteins. Rosetta employs a molecular-mechanics (MM)-based force field including knowledge-based potentials to guide the design of sequence modifications in a given protein structure.[14–16] Common features of Rosetta and other CED algorithms such as those implement in DEZYMER [17], ORBIT [18] and OSPREY [19] software suites include: *i)* a coarse-grained representation of scaffolds with lower degrees of freedom, where discrete sets of side chain rotamers or atom types are considered; *ii)* an assumption of low backbone flexibility; *iii)* scoring functions based on scaffold features and physics-based potentials and; *iv)* a classical treatment of metal-protein interactions. These considerations allow for the design of protein structures to be computationally tractable using heuristic approximations, since finding the optimal solution for the conformational space of side chain and backbone atoms in a polypeptide chain is NP-hard.[20] The robustness of employing Rosetta protocols to introduce new functionalities in protein scaffolds has also been

shown with the development of new biocatalysts that can perform non-native chemical reactions.[21–23] Moreover, CED approaches allow to integrate both simulations and directed evolution techniques to optimize initial designs with residual catalytic activity. [24]

## 1.2 Metalloproteases

Metalloproteases (MPs), also termed metallopeptidases or metalloproteinases, are metalloenzymes spread throughout all kingdoms of life. According to the MEROPS database of peptidases, MPs are found in fourteen clans and over sixty protein families, being encoded in around 2% of the genes in all organisms (further details of their classification will be given in Chapter 2). [25–27] They are located either intracellularly or bound to cell membranes, or can be secreted to the periplasm or extracellular environment. Other types of proteases are known besides MPs, which differ from the former by their catalytic apparatus. These are the serine, threonine, aspartic, glutamic proteases.

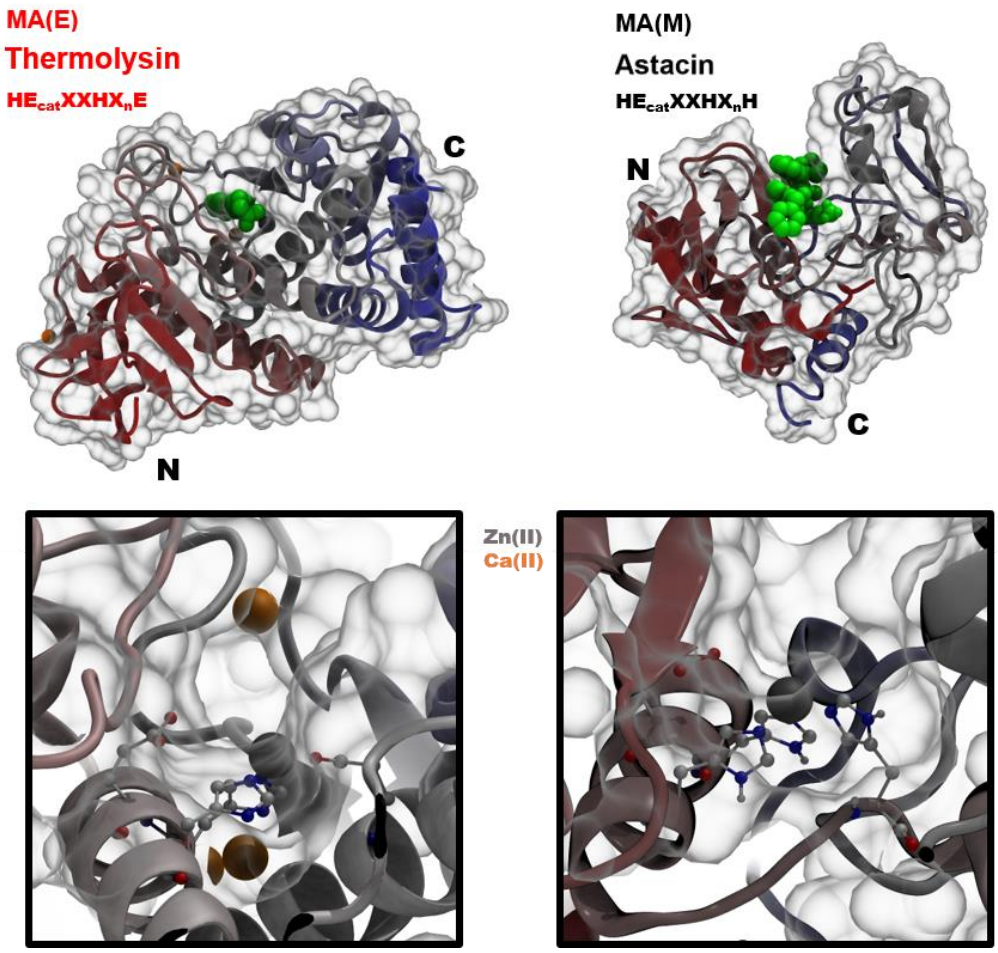
Cleavage of the peptide bond in the MP type of hydrolases is achieved by the nucleophilic attack of a water molecule activated by a metal ion present at the active site (AS), such as Zn(II), Co(II), Mn(II), Ni(II) or Cu(II). The most common type are MPs containing Zn(II), where the AS contains a single catalytic ion (often referred as zincins) or two ions acting co-catalytically. Three protein ligands coordinate the single metal ion at the AS, while in MPs containing co-catalytic Zn(II) ions the number of ligands is raised to five, with one ligand coordinating both metal ions.

Most of known MPs are grouped in the MA clan (clan classification according to MEROPS database), which is characterized by a single catalytic Zn(II) ion, the conserved HEXXH sequence motif and a common fold architecture - a two-domain catalytic unit where the AS is located in-between (Scheme 1.1). Division of this clan into subclans is based on the set of Zn(II) ligands present at the AS. These subclans are represented by their corresponding archetype proteins, astacin in MA(M) and thermolysin in MA(E), the latter being also the archetype protein for the entire MA clan. In MA(M) subclan members (metzincins) three histidine residues bind to the metal ion. A conserved methionine residue located in a β-turn underlying directly the AS, “met-turn”, is on the origin of the metzincin term.[1] In MA(E) subclan members (gluzincins) two histidines and one glutamate are the coordinating residues (the glutamate is on the origin of the gluzincin term).

Variations in domain composition and topology occur between families of the two MA subclans. The conserved N-terminal domain where the HEXXH motif is located contains both conserved α-helices and six-stranded β-sheets. The two Zn(II)-coordinating histidines common to both subclans are part of a conserved α-helix in this domain, together with the catalytic glutamate (details below). The C-terminal, which contains the third Zn(II) ligand, can vary greatly in terms of size and topology across families. This distance can either be eighteen or more residues in the

[1] There are cases where an aspartate replaces one histidine ligand.

MA(E) subclan, where the glutamate is contained in a α-helix, or six residues in the MA(M) subclan where the histidine is in a conserved β-turn. The MC clan in which the well-studied carboxypeptidase A is the archetype protein has a distinct fold but the AS organization is similar to the MA(M) subclan, where variations occur mostly at the level of the Zn(II)-coordinating nitrogen atoms from the histidine side chains. Due to such variations between distinct MP members, only the N-terminal domain is commonly used for classification. Conservation of AS features beyond residue composition is best captured in terms of their structural organization, as it will be further explored in Chapter 2.




**Scheme 1.1 –** Representative members of the two MA clan of MPs.
Thermolysin (PDB ID: 3FDG) from the MA(E) subclan and astacin (PDB ID: 1QJI) from the MA(M) subclan. Top: Backbone in cartoon representation, color-coded by residue index. Inhibitor molecules mimicking protein-substrate interactions in green. Bottom: detail of AS composition and structure for each subclan. Conserved sequence motif (n > 18 residues in MA(E), n=6 residues in MA(M)). Zn(II) coordinated to two histidines and one glutamate (MA(E)) or three histidines (MA(M)).

MPs from the MA clan have a wide range of substrates, spanning small peptides to full-sized proteins, and acting as exopeptidases (carboxypeptidases or aminopeptidases) if peptide bond cleavage is performed in the termini of the substrates, or endopeptidases if cleavage is performed

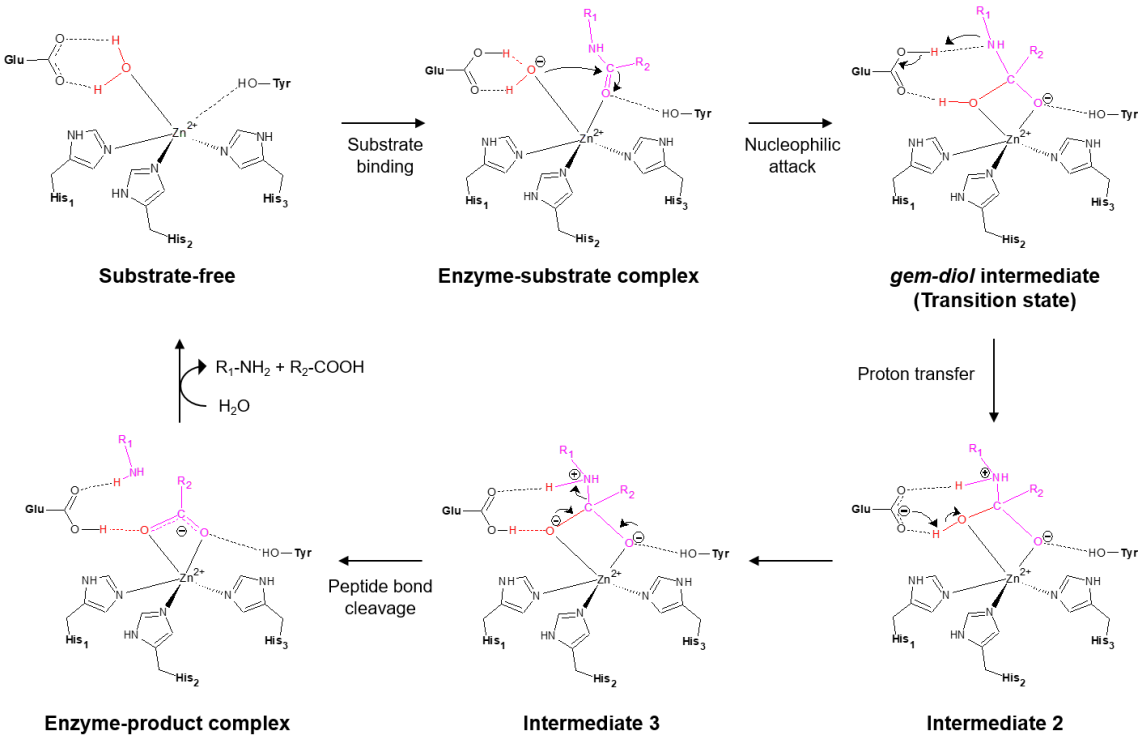
along the polypeptide chain. MA(M) members act strictly as endopeptidases while MA(E) members present both types of activities. Selectivity in MPs is a result of well-defined pockets placed along the corresponding AS clefts that mediate specific interactions between the enzyme and the substrate. MPs play many biological roles, such as virulence factors (*P. aeruginosa* elastase, the anthrax lethal factor from *Bacillus anthracis* and bacterial collagenases secreted from *Vibrio* and *Clostridium* strains) [28–31]; cell defence through bacteriolytic activity (staphylolysin) [32]; nutrition by degradation of other proteins [33]; regulation and homeostasis through tissue maintenance (matrix MPs); regulation through activation of small bioactive peptides, autocalytic activation through degradation of pro-peptides. [34,35] As it will be discussed in the following two sections, the catalytic role played by the Zn(II) ion allows for the cleavage of peptide bonds to occur under mild conditions (close to neutral pH in aqueous media), which makes MPs interesting target for technological applications.

## 1.3 Catalytic mechanism of metalloproteases

Different levels of theory have been used to describe the catalytic mechanism of metalloenzymes, such as the commonly used density functional theory (DFT) at quantum mechanical level and hybrid quantum mechanics/molecular-mechanics (QM/MM) methods. [36–38] In such studies a set of atoms from the enzyme are defined to represent the AS and its electron density treated quantum mechanically. When the protein environment and solvent contributions are considered to play an important role in the enzymatic mechanism, hybrid QM/MM methods are employed where the active site is treated accurately at quantum mechanical level and the remaining enzyme is treated with classical MM simulation methods. An overview of these methods is far beyond the scope of the current work, but it should be noted that valuable contributions to the understanding of MP enzymatic mechanisms have been obtained through their application.[39–47] While the mechanisms often vary for different atomic sets or even for different levels of theory employed, there are common features that apply for several MPs, including the well-studied MA clan representatives astacin and thermolysin. An overview of the catalytic mechanism of astacin based on DFT studies (with the B3LYP hybrid functional) is given in Scheme 1.2 and will be described below. [41]

The enzyme resting-state (substrate-free) contains Zn(II) bound to three structural ligands and one water molecule in a tetrahedral geometry. In astacin, the tyrosine OH group is considered as a fifth ligand in a trigonal bipyramidal geometry. The catalytic glutamate acts as a second coordination sphere ligand through positioning/interaction with the bound water via hydrogen bonds. The enzyme-substrate complex is formed when the substrate binds at the AS, where the substrate carbonyl oxygen replaces tyrosine as the fifth Zn(II) ligand, the so called “tyrosine switch” in MA(M) members.[48] Stabilization of this new interaction is mediated by the tyrosine residue or alternatively by a proximal histidine in thermolysin. [49] The bound water molecule is displaced

towards the catalytic glutamate residue, which acts as a general base by accepting one proton. The resulting hydroxide anion is polarized between the negatively-charged carboxyl group and the positively-charged Zn(II) ion, increasing the electronic density in the oxygen atom.




**Scheme 1.2 –** Enzymatic mechanism of a MP from the MA(M) subclan (Adapted from ref. [41]).

The reaction then proceeds by nucleophilic attack of the bound hydroxide to the substrate carbonyl carbon. Concomitant with this process is the formation of a hydrogen bond between the protonated glutamate and the nitrogen atom from the peptide bond, with the former acting as a general acid (proton shuttle mechanism). The resulting *gem-diol* tetrahedral intermediate is thus formed with an energy barrier of 19 kcal/mol, corresponding to the most energetic step of the reaction coordinate - transition-state (TS). In this intermediate, the Zn(II) ion is penta-coordinated to the three structural ligands and to the *gem*-diolate moiety acting as a bidentate ligand, with one oxygen from the carbonyl group and another from the hydroxide anion. Electronic charge accumulation at the carbonyl oxygen is stabilized by interaction with the tyrosine residue via a hydrogen bond and by the Zn(II) ion. However, it should be noted that this form of oxyanion stabilization is not conserved in the subclan: In stromelysin 1 and other matrix metalloproteases with surface-exposed AS pockets, an additional water molecule plays a similar electrophilic role as tyrosine. [46,50] In Snapalysin, the smallest MA(M) member known with 132 residues, the tyrosine is positioned away from the AS. In adamalysin II, MMP-8 and leishmanolysin this residue is replaced by a superimposable proline.[51]

The remaining steps of the reaction coordinate occur downhill the energy landscape. In intermediate 2 the hydrogen bond between the hydroxide anion and glutamate is strengthened due to negative charge accumulation in the latter, while there is positive charge build up in the nitrogen atom due to its protonation. For astacin, an additional proton transfer between the hydroxide anion and the catalytic glutamate before N-C peptide bond elongation has been proposed in the form of intermediate 3. Upon its formation, charge stabilization of the Zn(II)-bound carbonyl leads to disruption of the peptide bond. Finally, the enzyme-product complex is formed and the glutamate proton is transferred back to the bound hydroxide. The catalytic cycle is closed by detachment of the two product fragments by replacement with a solvent water molecule.

It should be noted that the three-proton exchange process is described for astacin, while in MA(E) members a single-proton transfer from the glutamate to the nitrogen concomitant with N-C bond cleavage is described instead.[40,52,53] In any case, the nucleophilic attack of an activated water molecule leading to formation of the *gem-diol* intermediate is commonly considered to be the most energetic step of the reaction for both subclans. The differences in catalytic environments and specific electronic readjustments occurring throughout the reaction in different MA members is an opportunity to test by CED if MP activity could be reproduced in alternative scaffolds.

## 1.4 Technological applications of metalloproteases

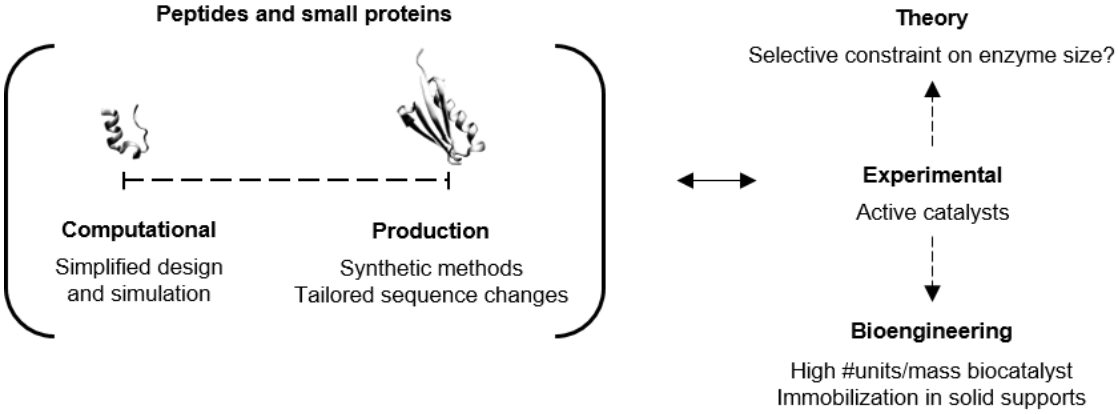
The MA clan contains most of the industrially-relevant MPs, constituting up to half of the total enzyme sales in the market. [1,54,55]. Examples are Thermoase PC10F (thermolysin - Amano Enzyme Inc., Japan) and Neutrase® (*Bacillus amyloliquefaciens* neutral protease - Novozymes Corp., Denmark) used in the food industry to produce flavour-enhancing peptides and hydrolysed food proteins from soy, wheat, milk and meat. Such flavour enhancers are used in soups, sauces and cheese products. Other applications include brewing [56] and leather [57] industries and also in the production of high-value functional foods given the probiotic, digestive and antioxidant effects of released peptides. [58–60] Reverse proteolysis with MPs can also be used for large scale production of aspartame by thermolysin (DSM, The Netherlands).

Biotechnological applications of MPs can also be explored, such as limited proteolysis for proteomic sequencing [61,62]. Other applications with therapeutic potential include the usage of collagenases in tissue dissociation to isolate different cell types or to remove necrotic tissue from burns and ulcers. [26,63,64] Given the wide range of organisms in which MPs are found (from psychrophiles to mesophiles and thermophiles, from acid to alkaline environments) their potential for biocatalysis is continuously explored, including function in mixed solvents common in process-based applications.[65] The activity towards biological targets and variety of folds, mechanisms and sizes adopted by different MPs makes them good targets to test if proteolytic activity can be recapitulated in alternative scaffolds tailored for process-based applications.

## 1.5 Small scaffolds for bioengineering applications

Peptides and small proteins (Scheme 1.3) are independently folded units whose structures often resemble (or are part of) single protein domains, where native functionality can be reproduced (or retained). The reduced size and complexity facilitates their design and allows for their structural features to be easily accessed through simulations. Moreover, it allows production to be done through synthetic methods, thus allowing full sequence modification and introduction of other non-natural chemical functionalities. Computational and experimental results can therefore be more easily achieved and correlations found between them used to guide the design process iteratively.

From a bioengineering perspective, a high number of functionally active units per mass of biocatalyst can be achieved using small (low molecular weight) scaffolds. Further advantages can be explored such as immobilization in solid supports for reusability, thereby reducing operation costs. [66,67] Magnetic nanoparticles based on iron oxide are particularly suitable given their high surface area to volume ratio that allows for high loading of biocatalysts and superparamagnetic properties that can be explored for extraction/recovery after each reaction cycle.[68]




**Scheme 1.3 –** The potential of peptides and small scaffolds for bioengineering applications.

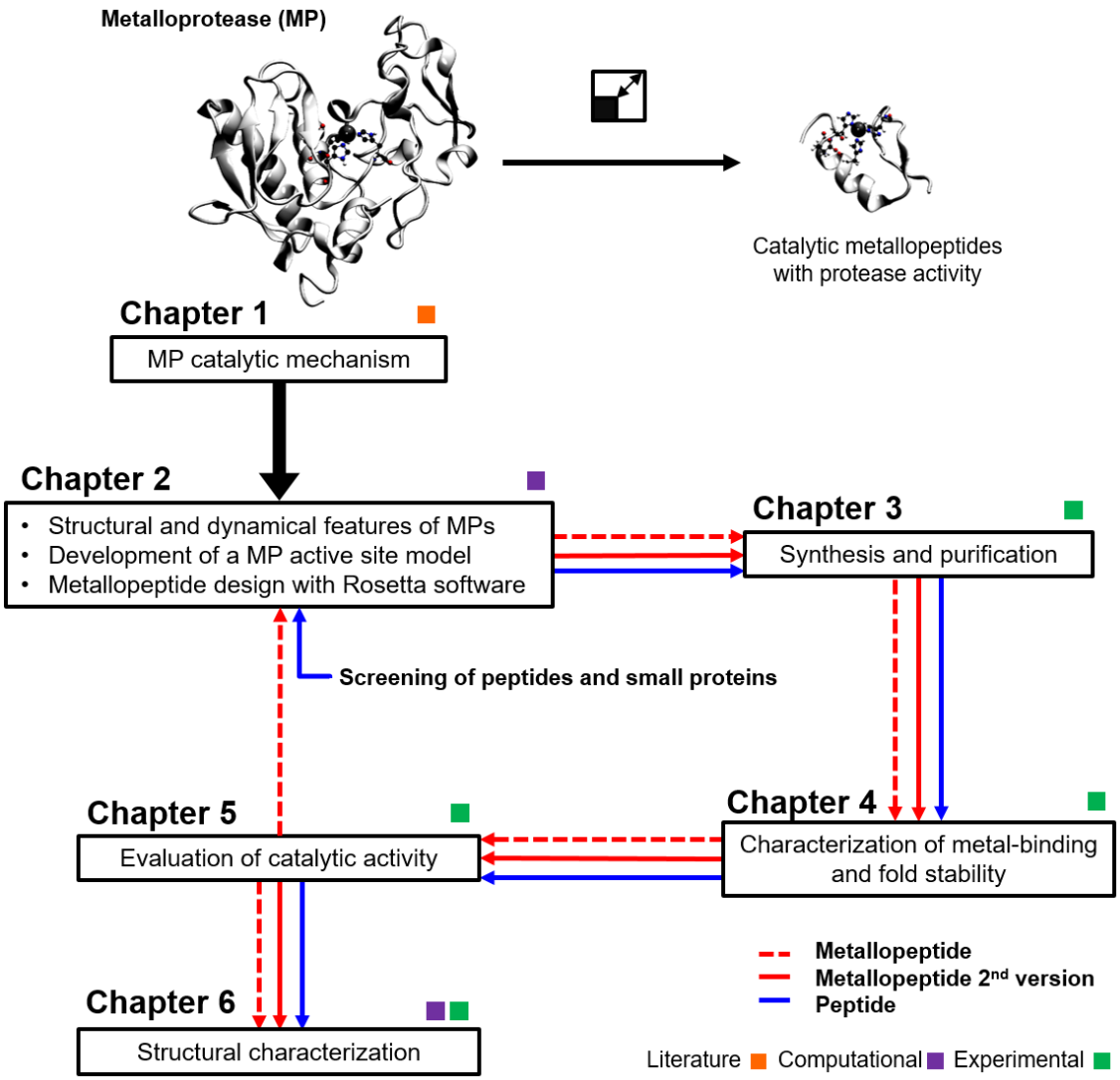
Although production of small scaffolds is easier to achieve once their sequences are defined, there are difficulties from both experimental and bioinformatic approaches to study them. This is because to isolate and access their functionality in biological system is challenging, and methods for prediction of genes are less robust to small DNA sequences. Therefore, our current knowledge of protein function is still too limited to larger systems, leaving the properties of peptides and small scaffolds under-studied. Only recently an increased focus on small genes and their function has been approached.[69–72] In these studies there is a pervasive under-representation of enzymatic function. Indeed, the smallest enzymes known up to date are around one hundred residues long,

falling outside the range of small proteins.[73] The only exception is 4-oxalocrotonate tautomerase with sixty-two residues in a βαβ fold, which nonetheless forms a homopentamer in its functional form.[74,75] This raises the question if there are selective constraints on sequence size for enzymatic function to be achieved in biological contexts or if this is a result of our current underexploration of small protein scaffolds. The current project addresses this issue by testing if MP activity can be achieved in peptide scaffolds. If so, then the range of scaffolds suitable for enzyme-based bioengineering applications can be expanded.

## 1.6 Objectives and project layout

The objective of the current project is to develop catalytic metallopeptides with protease activity (Scheme 1.4). Miniaturization of MPs was attempted by reproducing the functionality of their AS in peptide scaffolds combining computational design and simulation tools. Identification of active scaffolds was planned to be followed by immobilization in magnetic nanoparticles to develop recyclable catalysts for bioengineering applications. The starting point was the catalytic mechanism of MPs already described in the current chapter. In Chapter 2, MPs from the MA clan were characterized in terms of their structural and dynamical features. Conserved structural properties across a subclan were used to develop a computational model of the AS based on the catalytic mechanism described in the current chapter. The Rosetta software was then used to design a metallopeptide scaffold, where the sequence was modified to incorporate the model AS and to optimize interactions with both the Zn(II) metal ion and a model dipeptide substrate. The resulting peptide was synthesized thorough chemical methods as described in Chapter 3.

After purification and identity confirmation, the physicochemical and structural properties of the metallopeptide was addressed. In Chapter 4 both the metal binding affinity and thermal stability of the scaffold were characterized and compared to the native sequence to identify misfolded forms or destabilizing interactions introduced during the design stage. Protease activity of the designed metallopeptide was evaluated in Chapter 5 along with other esterase and amidase substrates. The results obtained in Chapter 4 and 5 were used to iteratively guide the computational methodology developed in Chapter 2. A second version of the metallopeptide was designed, and the approach was extended to screen a set of peptide and small-protein scaffolds. The candidate scaffold was identified and its sequence optimized as before. Both the second version of the metallopeptide and the candidate peptide were again characterized and their catalytic activities evaluated.




**Scheme 1.4 –** Objectives (top) and layout (bottom) of the current project. Details in main text.

Structural features of the designs were explored in Chapter 6 with both simulations and experiments in order to rationalize the low stabilities and low catalytic activities observed in previous chapters. The combination of both computational and experimental methods in this work allowed to identify advantages and inherent limitations of design tools when applied to metallopeptides and considerations for their improvement were discussed throughout the following chapters.

2

# 2. Computational Enzyme Design of Small Scaffolds

### Publication

Carvalho, HF, Roque ACA, Iranzo O, Branco RJF., *Comparison of the Internal Dynamics of Metalloproteases Provides New Insights on Their Function and Evolution,* PLoS ONE, 2015. 10(9): e0138118.

### Oral presentations in conference

Carvalho HF, Branco RJF, Roque ACA, Iranzo O, Encontro de Jovens Investigadores de Biologia Computacional Estrutural 2016 (2016 Oeiras, Portugal).

### Poster presentations in conferences

Carvalho HF, Roque ACA, Iranzo O, Branco RJF, 6th EuCheMS in Life Sciences (2015 Lisbon, Portugal).

Carvalho HF, Roque ACA, Iranzo O, Branco RJF, 29th Annual Symposium of The Protein Society (2015 Barcelona, Spain).

Carvalho HF, Iranzo, Roque ACA, Branco RJF, The Biochemistry and chemistry of biocatalysis: From Understanding to Design (2016 Oulu, Finland).

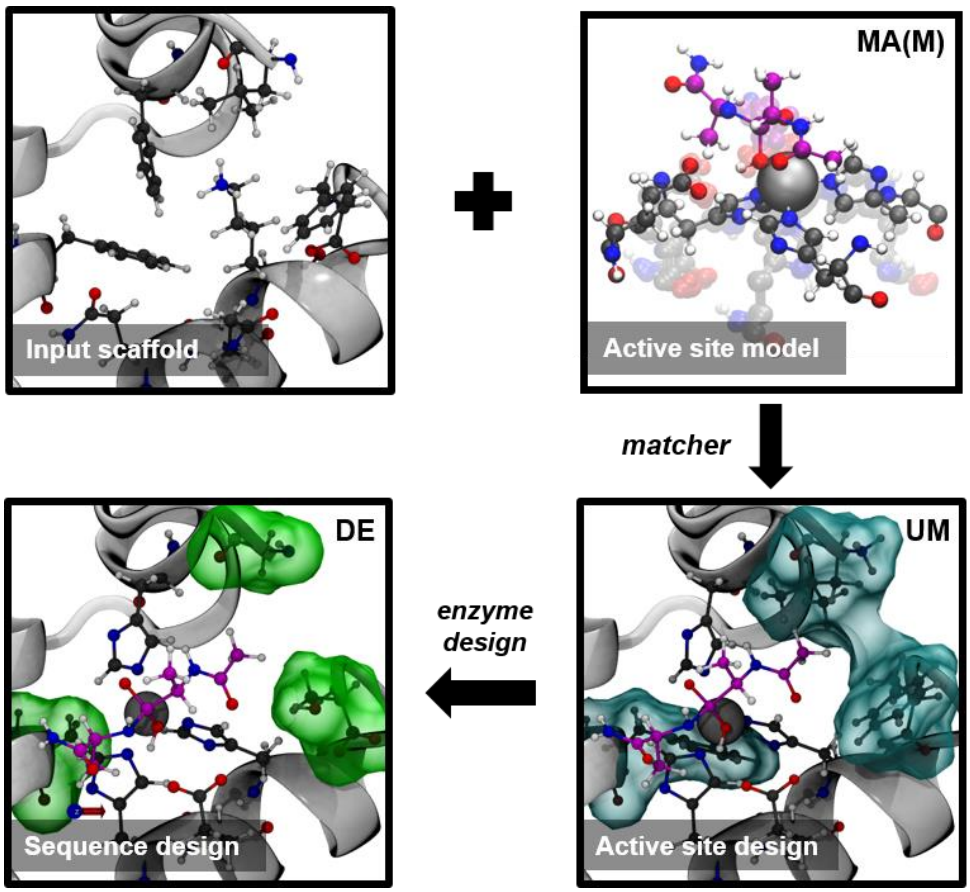
## 2.1 Introduction

In the study and design of enzymes, it is generally assumed that catalytic proficiency stems from their ability to stabilize the most energetic state(s) of a given chemical reaction at the AS.[76,77] As a result of this, reactions with higher activation energies (TS formation) present more challenging targets for designs where the TS stabilization is attempted. In order to produce a description of the underlying chemical reaction, theoretical models of the target enzymatic reactions are built where the enzyme AS is modelled together with the substrate as a TS. Introduced by Ken Houk's group in so called "theozymes", such models correspond to minimalistic descriptions of idealized protein orientations of side chain (or backbone) functional groups in relation to the TS.[78,79] In addition to QM methods to calculate optimized geometries for lowering of reaction activation barriers, models can also be derived from analysis of known enzyme AS. In this respect, MPs stem as ideal design targets given the wealth of information regarding both QM investigations of their enzymatic mechanism, as described in Chapter 1, and the experimental data available in the Protein Data Bank (PDB) [80] and MEROPS database of peptidases.[25] Regarding the latter, MP members are grouped into families based on their sequence similarity and structurally homologous families are combined into clans, therefore providing a description on how members are evolutionarily related at the sequence and structure levels.

In order to develop AS models of MPs, general features of MA clan members were first examined in light of their sequence-structure-dynamics relationships in Section 2.3.1. There has been an increasing interest in addressing functionally-relevant dynamical aspects of proteins, with the development of computational tools allowing to systematically compare proteins based on their dynamical similarity, such as the ALADYN webserver [81] and ProDy software [82] (details of principles and methodologies are fully addressed in Annex 1). Employment of such tools was made in order to correlate information on MP dynamics with their sequence and structure relationships, revealing important aspects of structural AS conservation in the MA clan. Following this, MP-inhibitor complexes were used to characterize the geometrical features of ASs (Section 2.3.2). These structures provide important clues regarding catalytically-relevant interactions, since in most cases the inhibitor molecules are modelled in order to mimic TS interactions, *i.e.* transition-state analogues (TSA).[83,84] TSA molecules typically contain a Zn(II) binding group (ZBG) which interacts with the metal ion as hypothesized upon formation of the *gem-diol* intermediate ($TSA_{ZBG}$). This source of information was chosen over DFT and QM/MM models of MP enzymatic mechanism as the later tend to be limited to a small number of specific case studies and therefore do not capture the heterogeneity of catalytic interactions across the MA clan.

With the characterization of both Zn(II)-coordination geometries and catalytic interactions in the MA(M) subset of MPs, computational models of the corresponding ASs were developed with the Rosetta software package in Section 2.3.3.[13] In addition to the examples given in Chapter

1 of its successful implementation in the design of biocatalysts, Rosetta has also been used recently to design a Zn(II) metal centre with first coordination sphere similar to the one found in native MA(M) members. [85] An overview of its implementation for design of small scaffolds is given in Scheme 2.1. Two programs included in the package are used for screening and design of peptides/small-proteins: *matcher* and *enzyme design*. The *matcher* program uses an arbitrary scaffold as input where backbone coordinates are used as anchor points to build the TS side chain geometry of a given AS model.[86] In cases where this is possible, a positive hit is generated (Unique Match, UM) corresponding to the input structure containing the modelled AS in a set of sequence positions. The *enzyme design* program is used afterwards to optimize catalytic interactions of UMs or other input structures. [87] It performs successive rounds of sequence optimization by designing and repacking residues close to the modelled AS, whose spatial organization is kept fixed. New amino acid entities are selected to minimize repulsions with AS residues and mediate favourable interactions with the substrate molecule.




**Scheme 2.2.1 –** Design of small scaffolds with the AS model of MA(M) subclan using Rosetta.

This process is repeated for a pre-defined number of iterations with the target function of minimizing the energy of the system, with generation of new amino acid entities based on a Monte Carlo sampling algorithm. The final outputted structures (Designed Enzyme, DE) may therefore

vary at each run in the number and type of amino acids introduced, which implies running the executable several times to obtain convergent sequence profiles. Quality of designs is evaluated based on Rosetta scoring function, including penalties for deviations from idealized starting geometries of the AS model. [16]

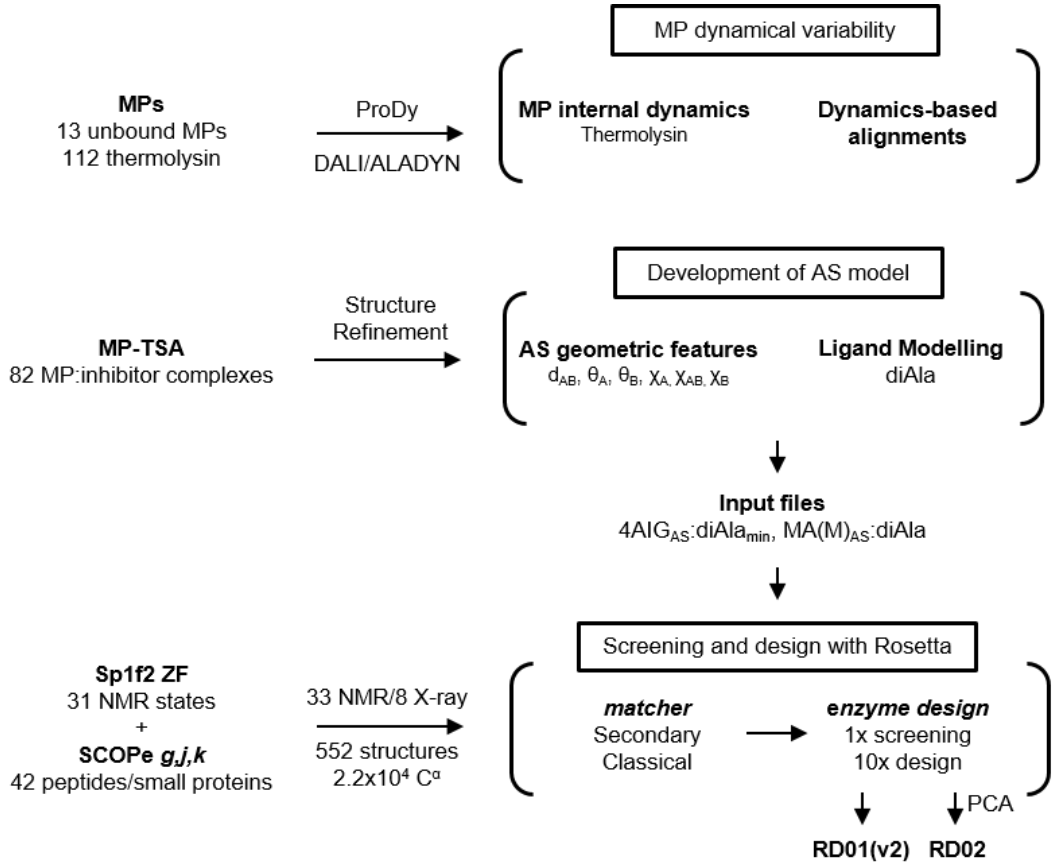
The general AS model of a MP was used in the control design of the MA(M) subclan archetype astacin in Section 2.3.4. After being validated, it was used in Section 2.3.5 for the redesign of a native zinc-finger (ZF) metallopeptide, the finger 2 of the (DNA-binding) human Sp1 transcription factor (Sp1f2). [88] ZFs have been previously used as input scaffolds for rational design of catalytic metallopeptides.[89,90] In the case of Sp1f2, variants with one less coordinating residue presented esterase activity. Differences in Zn(II)-coordination sets (histidines to glycine/alanine, cysteines to glycine/alanine) exhibited distinct pH-dependent catalytic profiles, which indicates that pKa modulation of bound water can be achieved by modifications of first coordination sphere interactions. As it will be described in following sections, the AS model developed contains both first sphere and second sphere interactions with the Zn(II) metal ion. Therefore, the Sp1f2 peptide is a good candidate to test if improved catalytic efficiencies can be obtained by employing computational methods where both coordination spheres are taken into consideration.

In addition to the redesign of a native metallopeptide, other peptides and small-protein scaffolds available in the Structural Classification Of Proteins – extended (SCOPe) database of protein structures were *de novo* designed in Section 2.3.6.[2] [91] The 42 alternative scaffolds varied in terms of chain length, fold and content of secondary structure elements. An additional design challenge was thus introduced - while in ZF redesign a structural Zn(II)-site is converted into a catalytic Zn(II)-site, in *de novo* design the catalytic Zn(II)-site needs to be designed from scratch since the scaffolds do not present native metal ions. Moreover, ZF peptides fold upon binding to the metal ion (metal-coupled folding) while for remaining scaffolds the driving force of folding is different, *e.g.* hydrophobic collapse. [92] Nonetheless, this approach was pursued since it allowed to test if the AS of native MPs could be modelled in other protein architectures. The best candidate resulting from the screening of peptide/small-protein scaffolds was identified and its design was made in Section 2.3.7. The designed scaffolds were synthesized in Chapter 3 and experimentally characterized in Chapter 4 and 5, with the observations made therein being used to iteratively guide the approach developed in the current chapter. Finally, the dynamical features of the designs were subject to further analysis and validation by other computational tools in Chapter 6.

[2] The term *de novo* is employed here in the context of enzyme design, referring to the development of a catalytically active variant of protein scaffold rather than the design from scratch of a new protein scaffold.

## 2.2 Materials and Methods

An overview of the workflow followed in this chapter is given in Scheme 2.2 and will be described below.




**Scheme 2.2 –** Workflow of CED developed in this chapter.

*Selection of structures:* The MEROPS database (release 9.9) was used to manually search and extract information of MP members. [25] The search routine for MPs focused on the MA clan, therefore excluding families containing bi-metallic Zn(II) centres or other metal ions. For ligand-free structures used in Section 2.3.1, full details of selection criteria are given in Annex 1. Sequence alignments were done with Clustal W program, where scores (bits) are related to the score of a given substitution matrix (credit for identity minus penalties for gap insertions and non-identity) and to the fraction between the number of aligned residues and total length of the alignment. [93,94] Network representations of similarity scores were done with the Cytoscape software using the Edge-weighted Spring embedded layout.[95] For MP-inhibitor complexes used in Section 2.3.2, the selected structures were those which *i)* contained one TSA molecule per protein with; *ii)* either a carboxylate or phosphonate ZBG where the distance Zn-O < 3.2 Å and; *iii)* had

an atomic resolution < 2.6 Å.[3] For peptide/small-protein structures used in Section 2.3.6, the SCOPe database (2.05 release) of protein structures was used. The classes *g: Small Proteins, j: Peptides* and *k: Designed Proteins* were manually inspected for structures with *i)* more than a single α-helix or β-sheet elements (minimum topological motif); *ii)* less than 65 residues and; *iii)* no metal ion or disulphide bridges involved in folding. The corresponding .pdb files were retrieved from the PDB database, selecting whenever possible the ones with no mutations, minimal number of NMR conformers, or highest resolution in case only X-ray structures were available.

*Refinement of MP-TSA structures:* For each of the *N* structures selected, coordinates of the Zn(II)-coordinating residues ($His_{1-3}$), the Zn(II) metal ion, catalytic glutamate ($Glu_{cat}$) and the $TSA_{ZBG}$ were extracted from the corresponding .pdb file and used for further analysis in PyMOL visualization software.[96] AS belonging to the MA(M) and MA(E) subclans were treated separately since the Zn(II)-coordinating residues are different (Chapter 1). The AS were then aligned and outliers were removed by an iterative procedure based on pairwise root-mean squared (RMS) deviations for different sets of atoms: *i)* 1st coordination sphere, Zn(II) ion and Nε2 from His residues or Cδ from Glu residue (Oε1 and Oε2 atoms can coordinate to the metal ion); *ii)* Side chain, Nε2, Cε1, Cδ2, Nδ1, Cγ Cβ, Cδ; *iii)* Backbone, C, Cα, N, O. To evaluate heterogeneities at each iteration, a symmetric *NxN* matrix of all pairwise alignments was calculated (*align_all_to_all.py* PyMOL script) for the three sets of atoms and the "similarity" score $S_i = \sum_{i=1}^{N} RMS$ calculated for each *i*th column. The average value *m* and standard deviation δ of $S_i$ values was also calculated. The exclusion rule $S_i > m+2\delta$ was applied to all atom sets in order to identify and remove outliers. For comparison between subclans, $S=S_i/N$ was also determined. For MA(M) this process was repeated until no outliers were obtained (*N* from 14 to 11). For MA(E) the procedure was done twice: first, only for thermolysin structures in order to decrease the total number (*N* from 39 to 16), second with the remaining AS from the subclan (*N* from 29 to 21). The resulting consensus set with no outliers was used to measure the dispersion of values of six geometrical parameters ($d_{AB}$, $\theta_A$, $\theta_B$, $\chi_A$, $\chi_{AB}$, $\chi_B$) between $TSA_{ZBG}$ and *i)* $His_{1-3}$ or *ii)* $Glu_{cat}$ with the PyMOL "measurement" tool (details below). Production of all images was done using VMD [97] and rendering of image files done with Tachyon. [98] Protein residues were typically represented in CPK (combination of bonds and Van der Waals), Zn(II) ion in silver spheres, backbone in ribbon or cartoon or surface representations. Colouring based on atom type: nitrogen in blue, oxygen in red, hydrogen in white, sulphur in yellow, carbon in grey (protein) or purple (ligand).

*Ligand modeling*: A model of the dipeptide Ala-Ala bound to Zn(II) as a *gem*-diol intermediate (diAla) was manually constructed based on atomic coordinates of TSA molecules bound to MPs in PyMOL. For $diAla_{min}$ the starting structure PDB code was 4AIG (Adamalysin II with bound inhibitor N-[(FURAN-2-YL)CARBONYL]-(S)-LEUCYL-(R)-[1-AMINO-2(1H-INDOL-3-YL)ETHYL]-PHOSPHONIC ACID, **FLX**) and for diAla the PDB code was 1QJI (high-resolution structure of

[3] MC and ME clans were also considered but discarded, since the number of members which fulfilled the criteria was significantly lower than those from the MA.

astacin with bound inhibitor CARBOBENZOXY-PRO-LYS-PHE-Y(PO2)-ALA-PRO-OME, **PKF**). In both cases the TSA was trimmed and the phosphate atom replaced by a carbon atom. The PKF molecule is modeled after a pentapeptide, which allowed for reconstruction of alanine side chains in extended conformation. The following steps were done to generate diAla conformers using the Open Babel software [99]: *i)* a *.pdb* file of the ligand without Zn(II) was processed with the *confab* command to generate the corresponding *.sdf* file (6 rotatable bonds identified, 7776 conformers tested and final 306 conformers produced, including the original extended conformer); *ii)* Zn(II) atom manually added to the *.sdf* file (replacement of the last hydrogen atom in file by Zn(II) coordinates and defining its binding to the $O_w$ atom). The resulting *.sdf* file was then converted to *1QJI_diAla_conf.params* file (Annex 2) with the executable *molfile_to_params.py* as part of Rosetta3.5 apps (M_commands file used, containing M ROOT Zn and M NBR Ow). This executable automatically sets the net charge of the molecule to 0 (initial charge 1.145, offset of -0.036 to all 32 atoms). All *.pdb* files of the 306 conformers were concatenated (replacement of END for TER lines), which is identified at the end of the *.params* file with the PDB_ROTAMERS flag. For $diAla_{min}$ the corresponding .pdb file of FLX was converted to a .sdf and subsequently converted to *4AIG_diAla_min.params* file as described for diAla.

*AS modelling:* computational versions of AS models were built using constraint files (*MAM_diAla.cst* and *4AIG_diAla_min.cst*), as shown in Annex 2. These five-column *.cst* files contain instructions to set *i)* the interacting ligand-residue atom pairs - Zn(II) to $O\varepsilon^1$ or $O\varepsilon^2$ atoms for $Glu_{cat}$ and Zn(II) to $N\delta^1$ or $N\varepsilon^2$ atoms for $His_{1\text{-}3}$; *ii)* the measured values (*x0*), deviations (*xtol*) and associated penalty constants (*k*) for all geometrical parameters; *iii*) sampling level (*n*) and; *iv*) periodicity, 360° for all geometrical parameters except for $\chi_{AB}$ in $His_{1\text{-}3}$ where it was set to 11.25°, corresponding to discrete rotation of the imidazole ring around the Cβ-Cγ bond. Interactions were modelled as pseudocovalent for $His_{1\text{-}3}$-Zn(II) and as non-bonded for $Glu_{cat}$-Zn(II), with higher *k* values for distances than for angles and dihedrals to penalize designs with distorted coordination geometries. Two versions of *.cst* files were used to sample the *x0*±*xtol* value range at regular intervals: the first with $n$=1 to sample $2n$+1=5 discrete values for all 6 geometrical parameters; the second with the "rule-of-thumb" *n=xtol/0.1* for $d_{AB}$ and *n=xtol/5* for remaining geometrical parameters (*n<4,* maximum 9 discrete values to keep algorithm running times feasable). The first version of the .cst file was used to screen and design NMR structures; the second version was used to screen and design X-ray structures and to redesign of Sp1f2, since it allows increased conformational sampling of side chain orientations. The *CstfileToTheozymePDB* Rosetta program was used to inspect the constraint files via production of corresponding *.pdb* files, with $n$=0 (*xtol* values only) for clear visual inspection.

*Screening of scaffolds with Rosetta*: The *.pdb* files of the selected scaffolds were used as starting point to build the inverse-rotamer tree of residues and substrate position (Dunbrack library of rotamers 2010, $\chi'_{1,2}$± 2σ side chain sampling for $His_{1\text{-}3}$) with the Rosetta *matcher* executable.[4]

[4] Each UM obtained is automatically split in files containing a maximum of 100 different diAla positions.

For each NMR (total 34, 10-38 states) or X-ray structure (total 8) used as input, the *options_matcher.flags* (Annex 2), scaffold positions file (*.pos*), AS model (*.cst*) and substrate (.params) files were used. The "secondary algorithm",[5] which allows for distorted coordination geometries of $His_{1-3}$ residues, was used in the first step to screen combinatorially all $C^{\alpha}$ positions. UMs where the AS had at least one residue in the termini or two consecutive His residues were discarded since such designs are unlikely to reproduce consensus coordination geometries. The "classical algorithm" was then used in the second step of design to filter the UMs with proper geometry for all $His_{1-3}$ and $Glu_{cat}$ residues. Only combinations of $C^{\alpha}$ positions where $Glu_{cat}$ could be modelled in the first step were screened for each structure. In the redesign of ZF scaffolds both algorithms were used. The two-step design approach is redundant since UMs with proper geometries are obtained in both stages. However, using the faster "secondary" algorithm in the first step allowed to screen a total of $2.2x10^4$ $C^{\alpha}$ positions for 552 structures (544 NMR and 8 X-ray) in a reasonable amount of time (3-4h CPU per structure in $20 < C^{\alpha} < 64$ positions, total ~1900h CPU time), which would not be feasible using first the "classical" algorithm (~1h CPU time per $C^{\alpha}$ position, total $>2.2x10^4$h CPU time).

*Design of scaffolds using Rosetta:* The Rosetta *enzyme design* executable was used to design scaffolds using as input the UMs (*.pdb*), together with information of which AS (*.cst*) and substrate (*.params*) models to use and the *options_enzdes.flags* file (Annex 2).The design step included 4 cycles of repacking and design of residues close to the modelled AS in order to optimize catalytic interactions. The output DEs were evaluated using Rosetta scoring function (Rosetta energy units, REU) with a total of 59 parameters. For principal component analysis (PCA) using the Origin Pro software (2016 edition), only a set of 16 parameters was selected to build the corresponding correlation matrix plus an additional parameter corresponding to chain length (*L*).[6] This set does not contain parameters specific to each scaffold, such as pose metric calculators (*_pm*). For enhanced discrimination of results, the parameters related with constraint penalty *k*, *all_cst* and *SR5_all_cst*, were represented as logarithm value of *k*. The *-native_compare* option flag was not used in order to compare the DEs with native scaffolds, since in case of NMR structures it yielded unrealistic values ($>10^7$ REU).[7] The final step of design did not evaluate repacking in the absence of diAla since Zn(II) was modelled as part of the substrate (charge repulsion in the absence of Zn(II) render the pre-organization of $His_{1-3}$ unfavourable). For screening of peptide/small-protein scaffolds, only one round of design was used (*-nstruct* 1) for a total of 542 UMs (1-4 min for DE production, total $> 10^4$h CPU time). UMs with coplanar H-bond pairs were not designed, as the

[5] The first residue to be screened, $Glu_{cat}$, was built with the classical algorithm, n=1 for all geometrical parameters. The remaining $His_{1-3}$ were built with secondary algorithm, n=1 with only $d_{AB}$ defined.

[6] *total_score* ($Score_{total}$), *fa_rep* (repulsive LJ), *hbond_sc* (H-bond energy), *all_cst* (constraint, k), *tot_pstat_pm* (Packing), *tot_nlsurfaceE_pm* (surface energy), *SR_1_total_score* ($Glu_{cat}$), *SR_2_total_score* ($His_1$), *SR_3_total_score* ($His_2$), *SR_4_total_score* ($His_3$), *SR_5_total_score* ($diAla_{score}$), *SR_5_fa_rep* (diAla repulsive LJ), *SR_5_hbond_sc* (diAla H-bond energy), *SR_5_all_cst* (diAla constraint, k), *SR_5_interf_E_1_2* (diAla interface energy), and *SR_5_dsasa_1_2* (diAla solvent accessible surface area, SASA).

[7] This may due to issues with Rosetta when using multi-state structure files.

*enzyme design* algorithm fails to process and evaluate such structures. For redesign of ZF scaffolds and sequence design of best candidate from peptide/small-proteins screening, 10 rounds were done (*-nstruct* 10) and sequence logos of outputted DEs generated to select the most frequent residue substitutions.[100]

## 2.3 Results and Discussion

### 2.3.1 Structural and dynamical variability of metalloproteases

The MA clan of MPs was first analysed to explore the relationship between sequence-structure-dynamics of the corresponding members. Full details of the analysis are described in Annex 1 and summarized below in Figure 2.1. Members from each subclan formed two distinct structure clusters, which is expected given the criteria of structural similarity to group MP members within clans in the MEROPS database. However, this relationship is not preserved when considering their sequence and dynamical similarity, with no clear identification of sequence or dynamics clusters. This suggests distinct selective pressures acting on different levels for MPs of the MA clan, with no clear correlation found between structural and dynamical similarity.

.

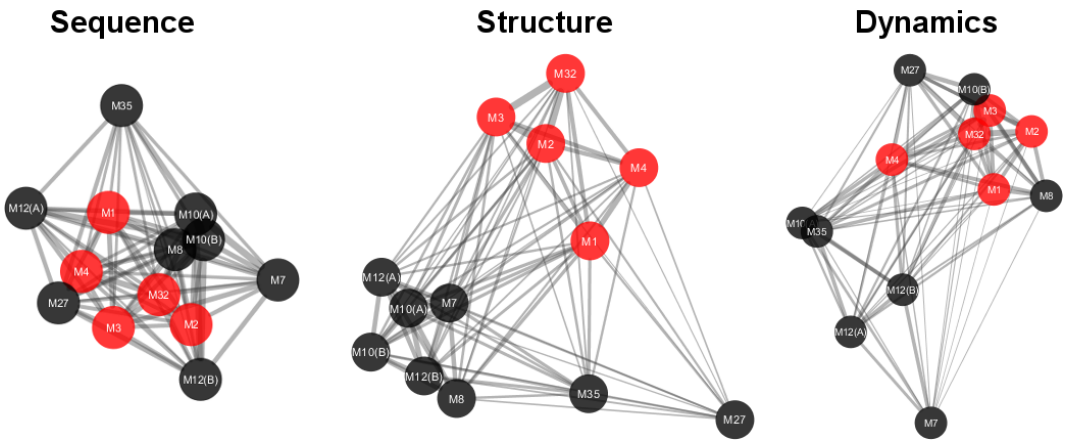



**Figure 2.1 –** Sequence, structural and dynamical variability of the MA clan of MPs.
Network representation of sequence (left), structure (middle) and dynamics (right) similarity between MP family representatives from the MA(M) subclan (black) and MA(E) subclan (red). Nodes correspond to family representatives and edge length proportional to similarity scores, bits for sequence and Z-scores for structure and dynamics (further details in Annex 1). Network topology built using edge-weighted Spring embedded layout, reflecting all pairwise comparisons between nodes.

An important observation resulting from this analysis was that AS tend to be located in regions structurally conserved and with low large-scale flexibility. This is relevant in terms of protein design, since the catalytically-relevant aspects of MP function can be focused on the analysis of the AS from crystallographic structures.

### 2.3.2 Analysis of metalloprotease active sites

The structural conservation and low flexibility of AS regions in MPs observed in previous section indicates that catalytic residues do not undergo major structural fluctuations along the catalytic cycle (reviewed in Chapter 1). Nonetheless, in the case of the clan type thermolysin, binding of inhibitors was associated with small residue fluctuations at the AS pocket. Given that such local fluctuations are catalytically relevant, their characterization was addressed. The selection of corresponding structures of MP-TSA complexes was made and is summarized in Table 2.1 for the MA(M) subclan and in Annex 2 for the MA(E) subclan.

**Table 2.1 –** Structures of MP-TSA complexes used in the analysis of MA(M) subclan AS.
Information of protein members was retrieved from the MEROPS database and selected based on the criteria defined in Section 2.2. Family representative members underlined. Selected structures for further analysis in black, structures excluded during refinement striked. *1CGL was excluded after the refinement step since it presented a highly distorted orientation of the $TSA_{ZBG}$.

| Subclan | Family | Member | MP-TSA complex |
|---|---|---|---|
| MA(M) | M10 | Fibroblast collagenase | 1CGL* |
| | | MMP3 | ~~1ZTQ~~ |
| | | Neutrophil collagenase | 1I76 |
| | | MMP7 | 1MMP |
| | | Stromelysin 1 | 1CIZ, 1CAQ, 1B8Y, 1HFS, ~~1HY7, 1SLN~~ |
| | M12 | Astacin | 1QJI |
| | | Adamalysin II | 4AIG |
| | | Atrolysin C | 1ATL |
| | | ADAM17 | 3G42 |

In the first step, information from MEROPS was used to select from the structures of MP-TSA complexes that followed the criteria defined in Section 2.2. In the second step, a refinement procedure was employed to remove large structural heterogeneities in the MA(M) set at the level of Zn(II)-coordinating residues, as shown in Figure 2.2. Flexibility of $Glu_{cat}$ residue was not considered at this stage since structural fluctuations are a result of local accommodations to different TSA molecules.

Outliers could be identified as “hot spots” in the matrixes of pairwise alignments, particularly in the case of 1st sphere Zn(II)-coordinating atoms and backbone. Iterative refinement of the MA(M) subclan in 3 rounds resulted in a refined set of 11 structures with smaller atomic heterogeneities. Therefore, the procedure allowed to obtain a more well-defined Zn(II) coordination geometry. Comparison of the AS from the two subclans shows that the 1st coordination sphere interactions are tightly preserved across subclans. However, the MA(E) subclan presents higher heterogeneities at the level of side chain orientation, although there are no major differences at the level of the backbone. Visual inspection of MA(E) ASs (not shown) revealed that side chain heterogeneities were related with glutamate coordination to Zn(II), which differ between monodentate and bidentate forms in Zn(II)-dependent enzymes.[101]

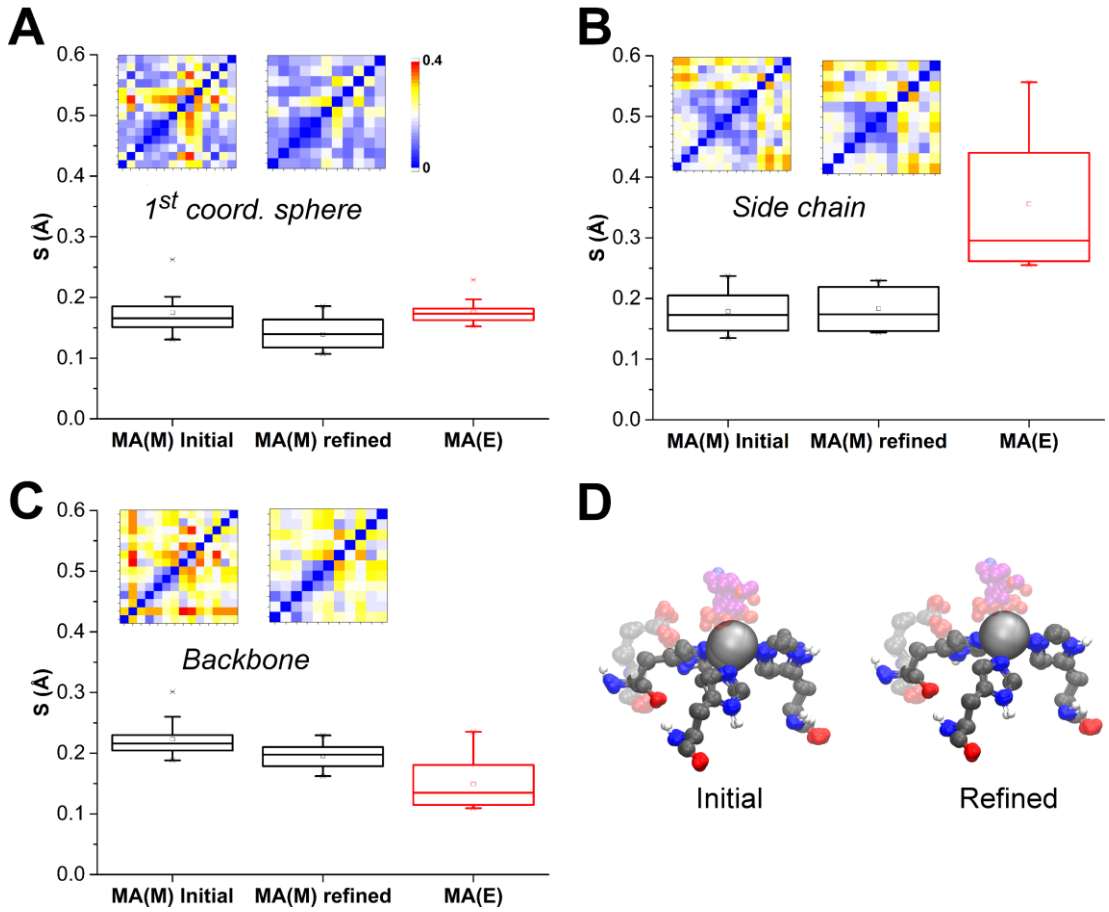
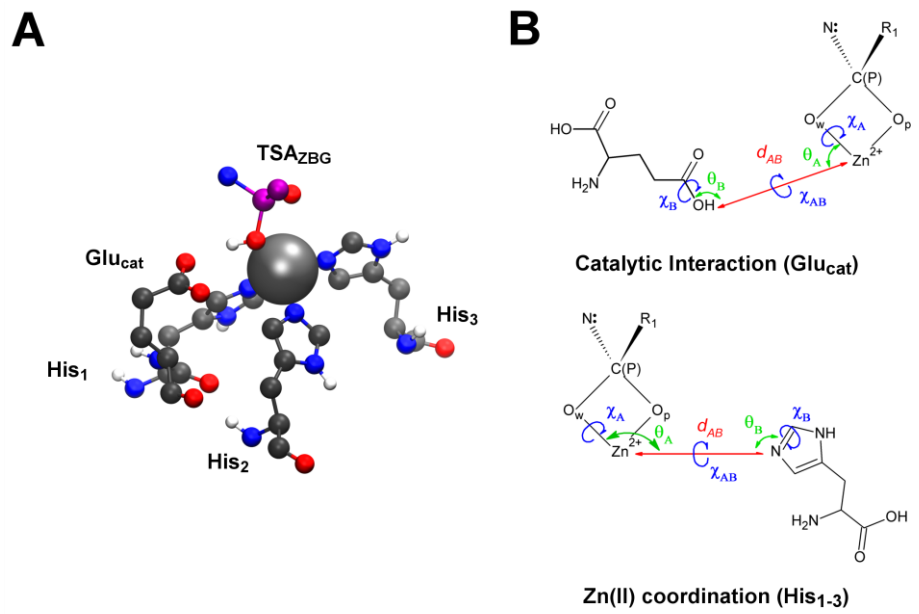



**Figure 2.2 –** Analysis of similarity scores (S) between AS of the MA(M) and MA(E) subclans. AS were structurally aligned based on three atoms sets: A- 1st coordination sphere, B - Side chain, C – Backbone. Details of atoms used for each alignment in Section 2.2. D – Comparison between the initial and refined $MA(M)_{AS}$ set, $Glu_{cat}$ and $TSA_{ZBG}$ transparent since they were not considered during alignments. Box plots correspond to 25%-75% distribution of S values (50% central line), crosses to the 1% and 99%, average value in squares and whiskers to outliers. Matrixes of MA(M) pairwise alignments shown in box plots, with each element corresponding to RMS values of each member ordered as in Table 2.1 and scaled to 0 < RMS < 0.4 for all atom sets. $MA(E)_{AS}$ correspond to the refined set of thermolysin structures and remaining subclan members analysed (details in Annex 2).

### 2.3.3 Active site model

The refined set of MA(M) members obtained in previous section was used to develop the corresponding AS model. This subclan was chosen over the MA(E) since Zn(II) coordination is more well-defined. Consideration for experimental aspects was also taken into account, since design of a Glu-containing AS of the MA(E) subclan in peptide cores may pose additional issues related with increased repulsive electrostatic interactions. Also, as discussed in Chapter 4, protein and peptide designs based on $(His)_3$-Zn(II) coordination motifs are common and thus allow for more direct comparison with the designs developed here. Selected MA(M) structures were analysed in order to characterize the geometric parameters of the AS, as exemplified for astacin in Scheme 2.3.




**Scheme 2.3 –** Characterization of AS geometrical features for the MA(M) subclan.
A -Example of AS from astacin, corresponding to the three Zn(II)-coordinated histidines ($H_1$-$H_3$), catalytic glutamate ($Glu_{cat}$), Zn(II) and bound TSA moiety ($TSA_{ZBG}$). B – Geometrical parameters defined for catalytic interactions (top) and Zn(II) coordination (bottom), corresponding to one distance ($d_{AB}$), two angles ($\theta_A$ and $\theta_B$) and three dihedrals ($\chi_A$, $\chi_{AB}$, $\chi_B$) between protein and TSA atoms.

Two types of interactions were defined, “zinc coordination” between $His_{1-3}$ residues and Zn(II) and “catalytic interaction” between $Glu_{cat}$, Zn(II) and the $TSA_{ZBG}$. The former was used to define the tetrahedral/trigonal bipyramidal coordination geometry of the metal, while the latter was expected to reflect orientation of the $Glu_{cat}$ upon TS formation. Characterization of these geometrical features reflects the specific orientation of TSA molecules along the active site pocket, therefore providing also information on the positioning of the bound carbonyl oxygen atom ($O_p$). The corresponding values of the two types of interactions for all MA(M) structures is summarized in Table 2.2.

MPs present specificity towards different peptide sequences on the basis of interactions of substrate side chains with pockets in the MP matrix. Such interactions modulate substrate access to the AS but are not directly coupled to the catalytic mechanism *per se*. Nonetheless, substrate bulkiness was taken into consideration when modelling the diAla substrate, as shown in Figure 2.3. This simple form of a capped peptide was chosen since it has only one target peptide bond but allows to access side chain and backbone orientations. A minimal representation of diAla was also modelled ($diAla_{min}$), with only one possible conformer, absent portions of the Ala residues (including backbone N- and $O_w$-bound hydrogens) and with no amidation or acetylation of the C- and N-terminals, respectively. The usage of the two substrate models allowed to evaluate the influence of size and conformation in the successful production of designs. These were based on structures of two MA(M) members with similar interactions between Zn(II) and $TSA_{ZBG}$. The high-resolution structure of astacin was used for diAla, while for $diAla_{min}$ the structure of adamalysin II was chosen instead (see below).

**Table 2.2 –** Geometrical parameters of AS from selected MA(M) structures.

| Interaction | | 1I76 | 1MMP | 1CIZ | 1CAQ | 1B8Y | 1HFS | 1QJI[a] | 4AIG[b] | 1ATL | 3G42 | $\bar{x} \pm \frac{\sigma}{\sqrt{n}}$ [c,d] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $d_{AB}$ | $His_1$ | 2 | 2.1 | 2.2 | 2.2 | 2.2 | 1.9 | 2.2 | 2.1 | 2 | 2 | 2.1±0.1 |
| | $His_2$ | 2 | 2.2 | 2.2 | 2.2 | 2.2 | 1.9 | 2.1 | 2.2 | 2.1 | 2 | 2.1±0.1 |
| | $His_3$ | 2 | 2.1 | 2.2 | 2.2 | 2.2 | 1.8 | 1.9 | 2.3 | 2 | 2 | 2.1±0.2 |
| | $Glu_{cat}$ | 5.1 | 4.9 | 5 | 5 | 5.2 | 5.5 | 4.7 | 5.2 | 4.6 | 5.1 | 5.0±0.3 |
| $\theta_A$ | $His_1$ | 100.5 | 91.8 | 93.6 | 91.8 | 94.2 | 93.3 | 105 | 96 | 112.4 | 89.6 | 96.8±7.1 |
| | $His_2$ | 88.4 | 94.9 | 88.4 | 96.6 | 93.3 | 88.2 | 108.8 | 83.9 | 106.5 | 92.6 | 94.2±8.1 |
| | $His_3$ | 142.2 | 150.6 | 151.4 | 147.2 | 144.2 | 154.4 | 130.6 | 144.7 | 131.9 | 143 | 144.0±7.8 |
| | $Glu_{cat}$ | 33 | 31 | 30.4 | 36.1 | 36.5 | 19.8 | 48.2 | 31.5 | 51.1 | 33.4 | 35.1±9.0 |
| $\theta_B$ | $His_1$ | 126.7 | 126.1 | 129.4 | 128.2 | 132.5 | 129.3 | 133.1 | 128.8 | 131.3 | 118.9 | 128.4±4.1 |
| | $His_2$ | 123.9 | 126.1 | 121.2 | 121.5 | 114.9 | 131.5 | 126.8 | 121 | 126.5 | 124.9 | 123.8±4.5 |
| | $His_3$ | 126.6 | 117.3 | 119.1 | 122.6 | 121.6 | 124.7 | 126.2 | 131.4 | 122.9 | 130.5 | 124.3±4.5 |
| | $Glu_{cat}$ | 88.5 | 96.1 | 93.1 | 94.7 | 90.8 | 83.3 | 85.3 | 94 | 98.8 | 93.3 | 91.8±4.8 |
| $\chi_A$ | $His_1$ | 248.4 | 261.1 | 247.7 | 240.4 | 237.4 | 249 | 243.3 | 246.2 | 281.4 | 252.7 | 250.8±12.6 |
| | $His_2$ | 143.9 | 158.3 | 142.1 | 134.4 | 131.9 | 145.7 | 134.2 | 143.5 | 172.3 | 153.2 | 146.0±12.4 |
| | $His_3$ | 36.2 | 34.2 | 37.4 | 17.5 | 1.7 | 37 | 10.7 | 37.7 | 53.6 | 38.3 | 30.4±15.5 |
| | $Glu_{cat}$ | 184.7 | 188.5 | 173.8 | 160.6 | 156.3 | 187.2 | 166.1 | 186 | 220 | 190.3 | 181.4±18.4 |
| $\chi_{AB}$ | $His_1$ | 165.5 | 149.8 | 167.8 | 169.1 | 170.4 | 156.5 | 148.9 | 167.3 | 129.1 | 157.6 | 158.2±12.9 |
| | $His_2$ | 208.6 | 197.9 | 191.3 | 198 | 198.3 | 198.3 | 197.4 | 201.9 | 198.9 | 191.2 | 198.2±4.9 |
| | $His_3$ | 4.6 | 22.9 | 2.8 | 16.2 | 29.6 | 8.7 | 39.2 | 4.3 | 21.1 | 19.8 | 16.9±12.0 |
| | $Glu_{cat}$ | 79.6 | 74.9 | 67.2 | 67.3 | 70.1 | 95.6 | 72.4 | 82 | 79.1 | 66.1 | 75.4±9.1 |
| $\chi_B$ | $His_1$ | 168 | 174.8 | 160.6 | 155.6 | 154.7 | 176.9 | 164.6 | 161.9 | 182.3 | 171.2 | 167.1±9.2 |
| | $His_2$ | 186 | 190.5 | 190.2 | 184.4 | 185.2 | 186.9 | 195 | 193 | 198.3 | 194.9 | 190.4±4.8 |
| | $His_3$ | 162.6 | 163.4 | 173 | 176.8 | 163.9 | 167.5 | 164.9 | 171.3 | 165.1 | 158.1 | 166.7±5.6 |
| | $Glu_{cat}$ | 138.7 | 145.7 | 140.3 | 141.3 | 143.8 | 148.4 | 151.6 | 148 | 151.5 | 141.2 | 145.1±4.7 |

a – AS of subclan archetype astacin used in diAla modelling. b – AS of adamalysin II used in $diAla_{min}$ modelling and RD01 design. c - Average values of each geometrical parameter for the n=10 analysed AS, used as *x0* values in the general MA(M) AS model. d – Corresponding standard error of values used as *xtol* in both the Adamalysin II and general MA(M) AS models.

The Zn(II) metal ion was modelled as part of the substrate molecule in order to combine the design of the metal site and substrate, as done similarly in the redesign of adenosine deaminase.[102] As discussed in Section 2.3.6, this proved to be useful in the design of peptide scaffolds lacking structural metal sites.[8]

The computational models of the AS were built using the Rosetta software package, as depicted in Figure 2.4.

[8] Design of only metal sites is possible by usage of virtual atoms bound to Zn(II) ion to define proper coordination geometries.

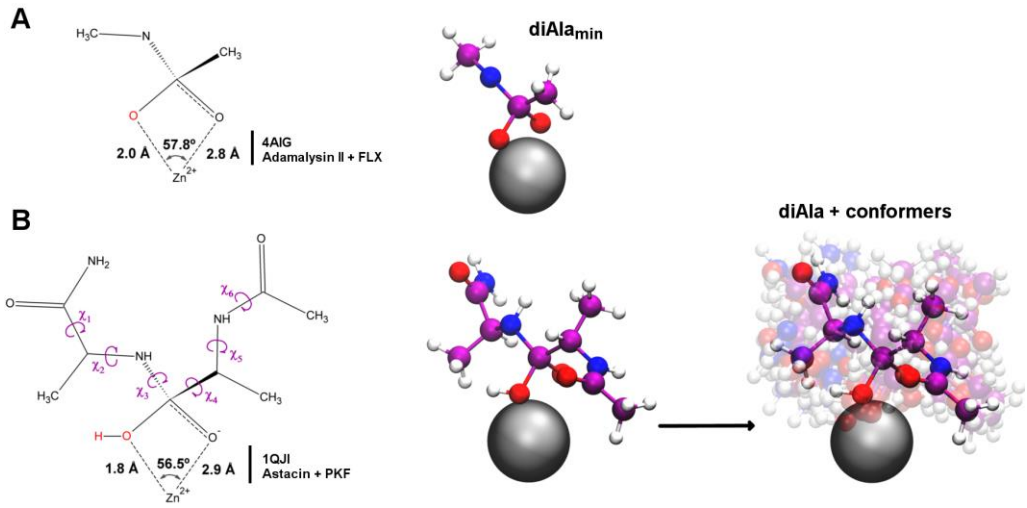



**Figure 2.3 -** Modelling of diAla$_{min}$ and diAla substrates.
A – Modelling of diAla$_{min}$ based on the structure of adamalysin II-FLX complex (PDB 4AIG). B – modelling of diAla based on the structure of astacin-PFK complex (PDB 1QJI). Displayed distances and angle in two-dimensional models (left) correspond to values found in the respective crystallographic structures. $O_w$ atoms identified in red and $\chi_{1\text{-}6}$ rotatable bonds of diAla in purple. Single conformer for diAla$_{min}$ and 306 conformers for diAla represented in three-dimensional models (right). Transparent representation of conformers used for clarity.

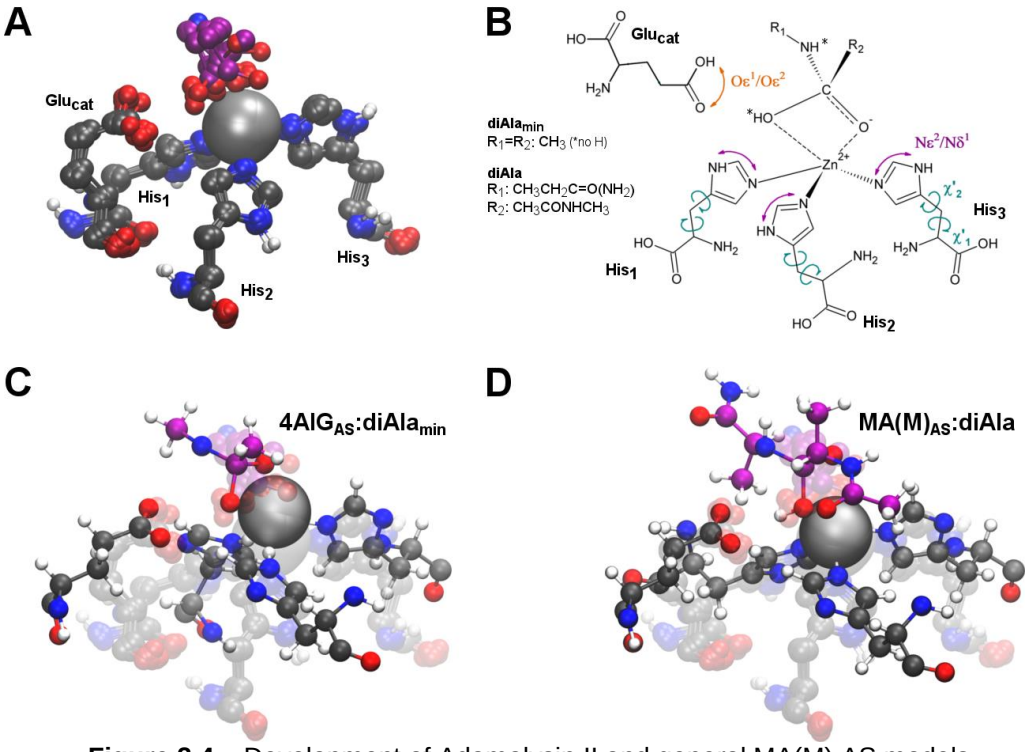



**Figure 2.4 –** Development of Adamalysin II and general MA(M) AS models.
A – Refined set of MA(M) AS, including Glu$_{cat}$ residue and TSA$_{ZBG}$. B – Two-dimensional representation of models Adamalysin II (4AIG$_{AS}$:diAla$_{min}$) and general MA(M) (MA(M)$_{AS}$:diAla). Interactions with more than one possible atom identified as coloured double arrows. Sampled histidine rotamers controlled at the level of $\chi'_1$ and $\chi'_2$ dihedrals. C – three-dimensional representation 4AIG$_{AS}$:diAla$_{min}$ D – three-dimensional representation MA(M)$_{AS}$:diAla. Superimposition of modelled AS with the refined set of MA(M) AS displayed in transparent representation and using arbitrary $\chi'_1$ and $\chi'_2$ sampling values for clarity.

Starting from the characterized geometrical features, a computational version of the Adamalysin II and general MA(M) AS was developed using the *CstfileToTheozymePDB* program. The corresponding models of the Adamalysin II-diAla$_{min}$ (4AIG$_{AS}$:diAla$_{min}$) and general MA(M)-diAla (MA(M)$_{AS}$:diAla) ASs were generated and checked against the refined set.

Overlap with experimental structures shows that both models reproduce proper AS geometries and that sampling of side chain conformers does not cause distortions of interacting atoms. Moreover, the AS maintains the trigonal bipyramidal Zn(II)-coordination geometries found in MP-TSA complexes. This coordination geometry was enforced by attributing pseudocovalent bonds between the Zn(II) ion and the coordinating nitrogen atoms of histidine residues. Both computational models were therefore validated, and their usage in a native enzyme tested in the next section.

### 2.3.4 Control design of astacin

The MA(M)$_{AS}$:diAla model was tested for its capacity to recapitulate the AS of native astacin in a control design. Using the sequence positions of AS residues as input, the *matcher* program successfully generated 4 UMs where MA(M)$_{AS}$:diAla was accomodated in the binding pocket, as shown in Figure 2.5. Each UM corresponded to distinct diAla conformers, with small deviations in residue side chain orientations and substrate position (total > $10^4$). Using the *enzyme design* program to evaluate the UMs, the corresponding DE with best score (repack with no sequence design) presented residue conformations similar to the one found for the native structure. Integrity of catalytic interactions was kept, as exemplified by Y149 orientation which contribute to oxyanion stabilization upon formation of the TS.[9]

The control design shows that the MA(M)$_{AS}$:diAla model used in Rosetta can reproduce AS features of a native MP, and therefore was used in the following sections in the redesign of a metallopeptide and *de novo* design of peptide/small protein scaffolds.

[9] During the design step Glu$_{cat}$ and His$_{1-3}$ residues are kept fixed, therefore conservation of their orientation is enforced.

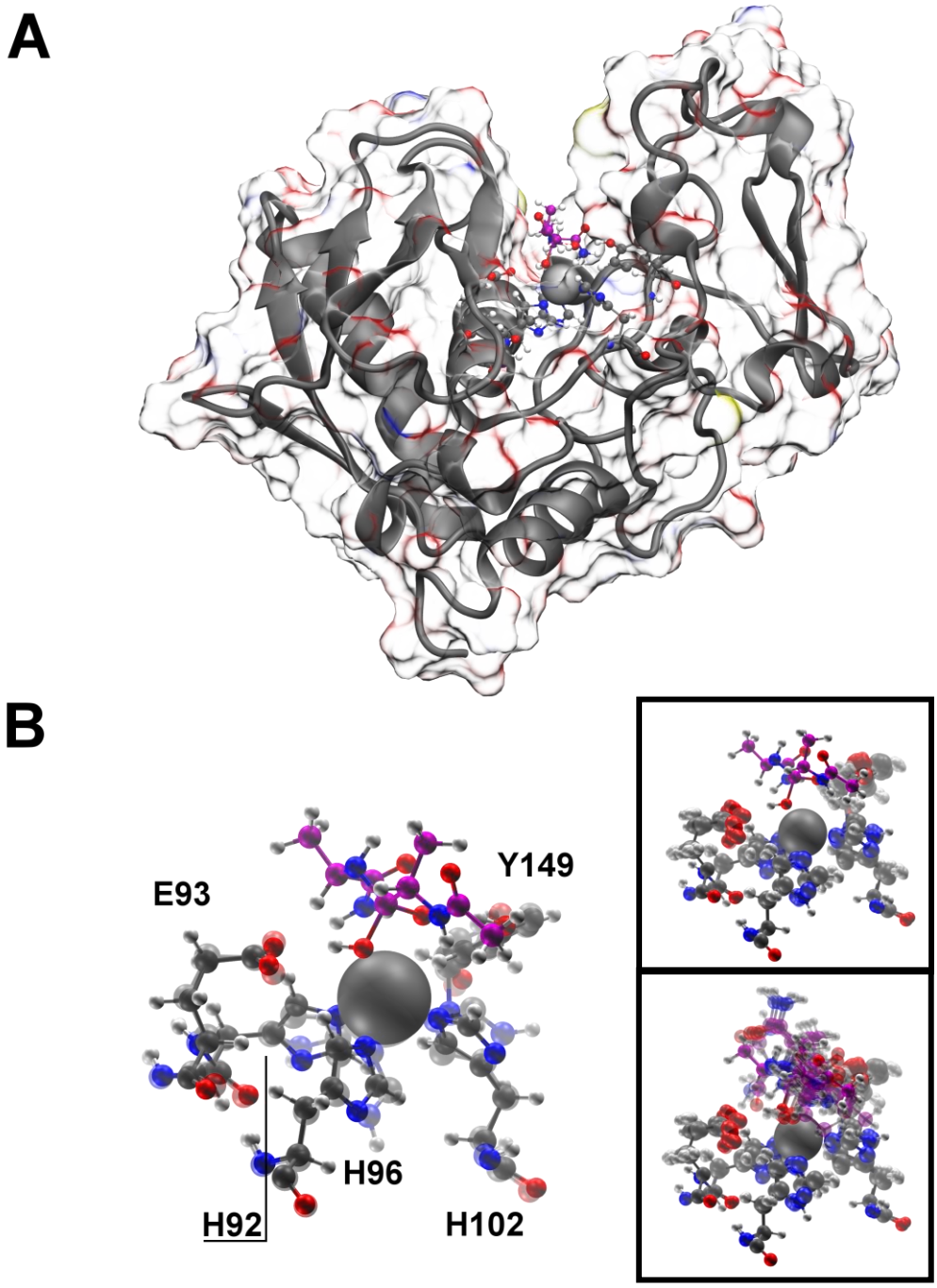
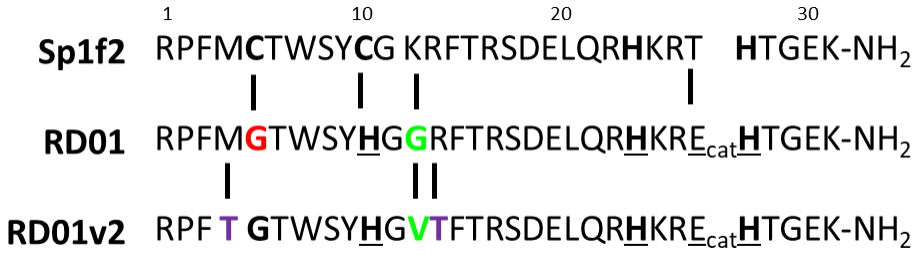



**Figure 2.5 –** Control design of astacin with MA(M)$_{AS}$:diAla model.
A – Scaffold representation of a UM obtained for astacin with modelled MA(M)$_{AS}$:diAla. B – AS detail of the DE with lowest Score$_{total}$ and comparison with native AS in transparent representation (left). Detail of residue fluctuations (right, top) and diAla conformers (right, bottom) designed within the range of sampled geometrical parameters. Residue fluctuations and diAla conformers shown in transparent representation for clarity. AS residues identified in bold. The Y149 residue does not belong to the MA(M)$_{AS}$ but was represented since it was subject to repacking during the design stage and contributes for oxyanion stabilization in the native enzyme.

### 2.3.5 Design of RD01 and RD01v2

The first design was focused on the Sp1f2 ZF, whose sequence is shown in Scheme 2.4. Given that the AS models contain similar Zn(II)-coordination motif as in other active Sp1f2 variants (discussed in section 2.1), redesign of Sp1f2 was expected to also yield catalytically-competent peptides. With this scaffold, it could be addressed if inclusion of second sphere interactions with the metal ($Glu_{cat}$) lead to an increase in catalytic competence of the Sp1f2 scaffold.




**Scheme 2.4 –** Sp1f2 redesign into RD01 and RD01v2.
Modelled residues underlined and Zn(II)-coordinating residues in bold. Sequence changes identified by vertical bars: DEs in green, non-Zn(II)-coordinating residues in red and scaffold stabilization in purple.

Structures of Sp1f2 were used in the Rosetta *matcher* program to test if design of the both AS models was possible. UMs were not found for more compact X-ray structures (PDB ID: 1SP2), which prompted the usage of NMR structures (PDB ID: 1VA2, 31 states), as shown in Figure 2.6.[10] This allowed to test different backbone conformations, therefore allowing for exploration of the flexible features of the peptide in solution. All NMR states were independently screened, using the native $(His)_2(Cys)_2$-Zn(II) residue positions and the $4AIG_{AS}$:$diAla_{min}$ model as input, translating in $>10^6$ conformations sampled for each designed residue. This model was initially used as it corresponds to an empirically-derived AS with geometrical features close to the set average.

Generation of UMs for Sp1f2 was limited, with only 1 $diAla_{min}$ placement with small atomic fluctuations in 1 NMR state, as shown in Figure 2.7. Additional approaches were pursued to generate more UMs, such as combinational variations of tested residue positions, increase of sampling density level, employment of secondary matching algorithm for $His_1$-$His_3$ design. These however did not lead to new UMs, which points to a highly-constrained design. This may be attributed to the C10 sequence change for bulkier H10, corresponding to $His_3$ design in the hydrophobic core.

The designed AS presented a mixed coordination to Zn(II) by $N\varepsilon^2$ (H10, H23) and $N\delta^1$ (H27) atoms, while in native Sp1f2 coordination of the two histidine residues (H23, H27) is made through the $N\varepsilon^2$ atoms. Sampling of both nitrogen coordinating atoms was modelled since it is also found

[10] Additional X-ray structures of finger 1 and finger 3 of Sp1 protein, X-ray structures of Zif268 and the consensus-ZF peptide (CP1) structures were also tested. No positive hits were obtained for these scaffolds with the $4AIG_{AS}$:$diAla_{min}$ model. NMR structures tend to present less packed side chain and backbone conformations.

in native MPs - in MA clan through the $N\varepsilon^2$ atoms and in structurally similar MC clan through the $N\delta^1$ atom (*e.g.* Carboxypeptidase A, PDB ID: 6CPA).

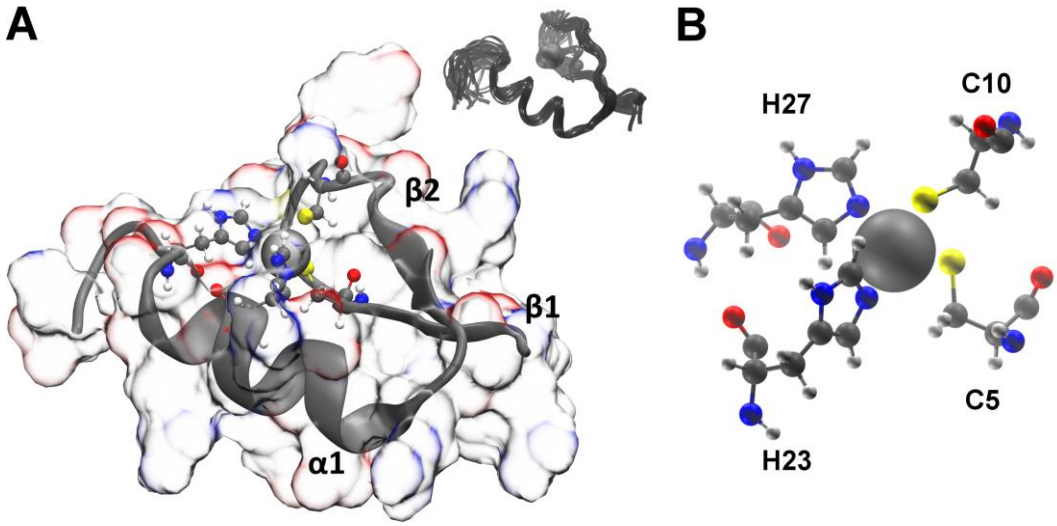



**Figure 2.6 –** Sp1f2 zinc finger metallopeptide.
A – Scaffold cartoon representation. Secondary structure elements identified in bold. NMR states in ribbon representation shown in top right. B – Detail of tetrahedral Zn(II) structural site with coordinating residues identified in bold.

The designed Zn(II) ion was more surface-exposed than in native scaffold, and the $Glu_{cat}$ was designed by a T26E sequence change, thus recapitulating the $HE_{cat}XXHX_nH$ coordination motif found in native MPs. The resulting UMs were edited with an additional C5G sequence change to allow for a bulk solvent molecule to coordinate with the Zn(II) ion.

The Rosetta *enzyme design* program was used to design the vicinity of the modelled AS to optimize interactions of the scaffold with the $diAla_{min}$ substrate. DEs with an additional K12G sequence change resulted from this step, which was accepted given the steric clash between $diAla_{min}$ and Lys side chains observed in UMs. The resulting peptide sequence shown in Scheme 2.4, termed RD01 (Rosetta Design 01), was selected for synthesis and experimental characterization. As it will be described in further detail in the following chapters, this peptide sequence adopted a fold similar to the native structure but presented no catalytic activity towards diAla substrate and low hydrolytic activity towards a general esterase substrate, similar to other reported Sp1f2 variants. Issues related with decreased stability of the scaffold were encountered, therefore it was addressed if improvements at the design stage could lead to a catalytically competent version of RD01.

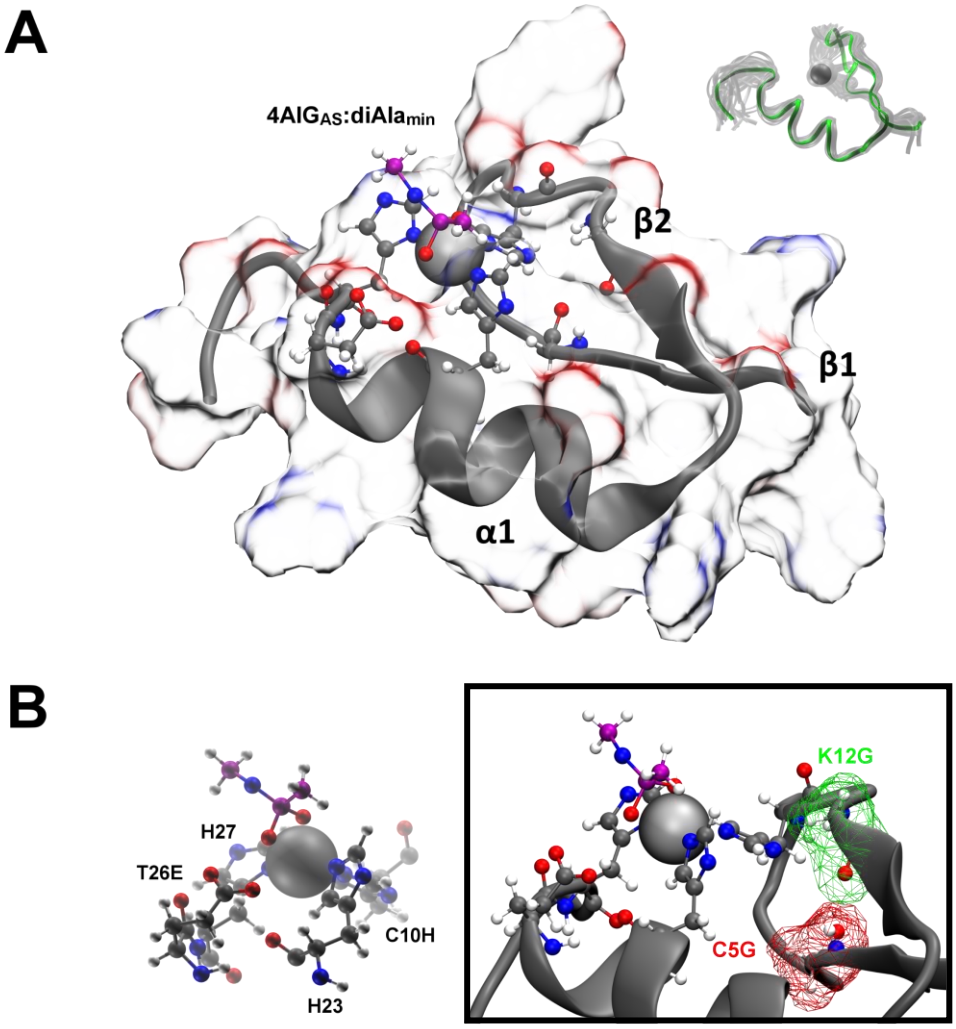



**Figure 2.7 –** Design of RD01.
A – Scaffold representation of obtained UM for Sp1f2 with modelled 4AIG$_{AS}$. Secondary structure elements identified in bold and corresponding NMR state (27) in green ribbon. B – AS detail of the DE with the diAla$_{min}$ substrate and AS sequence changes identified in bold (left). Detail of designed sequence changes identified in bold and shown in coloured wireframe (right).

In the second round of Sp1f2 redesign, the MA(M)$_{AS}$:diAla model was used and all NMR states of the structure were screened indenpendently with the Rosetta *matcher* program, with a higher number of sampled conformations (> $10^7$) due to modelling of diAla conformers. As shown in Figure 2.8, the resulting UMs from 1 NMR state presented the same Zn(II)-coordinating atoms and side chain orientations of AS residues.

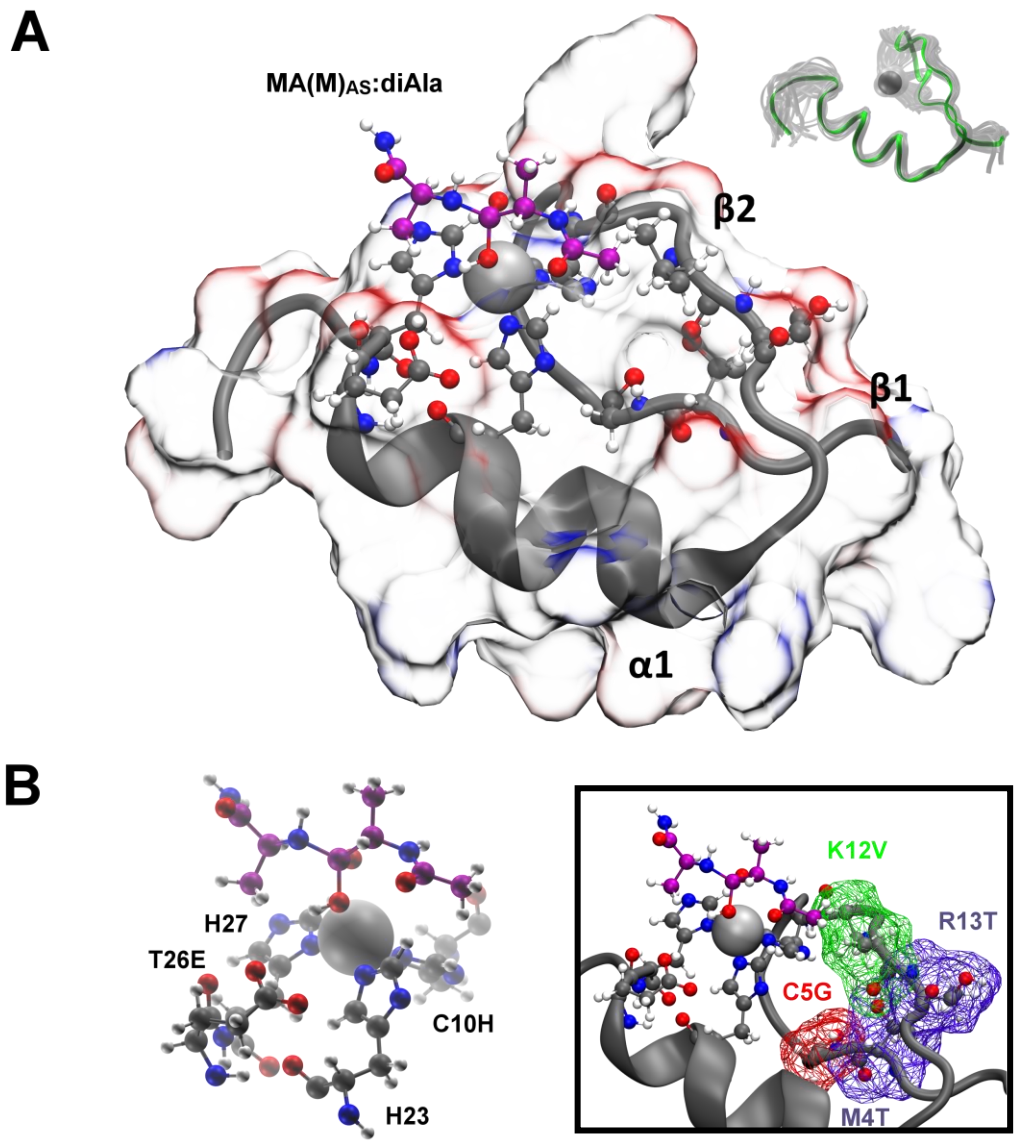



**Figure 2.8 -** Design of RD01v2.
A – Scaffold representation of obtained UM for Sp1f2 with modelled MA(M)$_{AS}$. Secondary structure elements identified in bold and corresponding NMR state (27) in green ribbon. B – AS detail of the DE with the diAla and AS sequence changes identified in bold (left). Detail of designed sequence changes identified in bold and shown in coloured wireframe (right).

Placement of diAla was similar to the one found in previous design for diAla$_{min}$ but only one conformer was obtained. No additional UMs were obtained when employing other redesign approaches, pointing again to a constrained design. Therefore, the usage of both AS models does not lead to differences in produced UMs. An additional C5G or C5A sequence changes were done and scaffold stabilization was pursued by including the M4T (β1) and R13T (β2) sequences changes, as threonine mutations in β-sheets are known to increase the stability of designed ZFs.[103–105]

The modified UMs were used as input for further design with the Rosetta *enzyme design* program. DEs presented a new sequence change, K12V or K12H, the later being discarded in

order to not introduce additional Zn(II)-coordinating residues. The designed valine side chains interacted with the capped N-terminal of the diAla substrate. The two sequence variants, C5G and C5A, presented the K12V sequence change. Differences in the desing of K12G (in RD01) and K12V (RD01v2) were addressed and found to be due to the usage of the *-enz_debug* option flag in the former, which sistematically generates a K12G sequence change for the Sp1f2 scaffold.[11] Evaluation of Rosetta's scoring of these both C5G and C5A variants revealed that they were nealy identical (not shown), and therefore the C5G sequence change was chosen to preserve the non-coordinating residue of RD01. The resulting sequence shown in Scheme 2.4, RD01v2, was synthesized and experimentally charachterized. Similar catalytic and structural features as RD01 were obtained, thus indicating that improvements at the design stage do not lead directly to more profficient catalysts in the case of Sp1f2 scaffold. Wether the AS model or the Sp1f2 scaffold were unsuitable for development of an active MP was addressed in the following section, where a more thorough evaluation of RD01v2 scoring was made in parallel with the screening and design of other peptide/small-protein scaffolds.

### 2.3.6 Screening of peptide and Small protein scaffolds

A set of 42 peptide/small-protein scaffolds (Annex 2) with varying chain lengths and with no structural metal sites or disulphide bridges were selected from the SCOPe database and screened for their capacity to accommodate the catalytic Zn(II)-site of the $MA(M)_{AS}$:diAla model. Two sequential steps of screening were done with the Rosetta *matcher* program: the first – "Secondary Algorithm" - allowed for distorted coordination geometries of $His_{1\text{-}3}$ residues; the second – "Classical Algorithm" - only allowed for both properly defined coordination and catalytic geometries. Whenever possible, NMR structures were used to capture the inherent flexibility of small scaffolds. The number of tested conformations was typically $>10^9$ per modelled residue/substrate/scaffold position. The results from the screening are summarized in Table 2.3 and in full detail in Annex 2. A high number of UMs were produced for the majority of screened scaffolds, although no correlation with chain length or fold type was found. This shows that the $MA(M)_{AS}$ can be easily modelled in smaller systems with reduced number of secondary structure elements, not necessarily restricted to the same architecture of native MPs. Moreover, it indicates that the low number of UMs obtained for RD01 and RD01v2 were due to high structural constraints imposed by the Sp1f2 backbone and not a limitation of the used $MA(M)_{AS}$:diAla model or options controlling the *matcher* executable.

[11] The usage of *enz_debug* option flag was chosen in RD01 redesign, which was made using Rosetta3.4 version. The redesign of RD01v2 and remaining $MA(M)_{AS}$:diAla designs was made using Rosetta3.5 where the *-enz_debug* flag was not used. In a subsequent control design of RD01 using Rosetta3.5 without the *enz_debug* flag, the K12V sequence change was also obtained, thus pointing to the sensitivity of the design methodology to used parameters.

**Table 2.3 –** Summary from the screening and design of peptide and small-protein scaffolds.

| SCOPe | Screened | Scaffolds |
|---|---|---|
| class *g,j,k* | 20-64 residues<br>34 NMR, 8 X-ray | 42 (33 all-α, 6 α+β, 3 all-β) |
| **Matcher** | **Modelled AS** | |
| Secondary algorithm | 279 | 35 |
| Classical Algorithm | 122 (542 UM, $5.5x10^5$ diAla) | 27 |
| **Enzyme Design** | **Designed** | |
| DE | 5494 | 27 (20 all-α, 4 α+β, 3 all-β) |
| $DE_{P10}$ | 555 | 11 (8 all-α, 2 α+β, 1 all-β) |

All UMs obtained were designed with the Rosetta *enzyme design* program, including the UMs from RD01v2 as a negative control (no MP activity) and native astacin as a positive control (MP activity). Given the high number of produced DEs, a detailed evaluation of each one was not feasible. Also, ranking of DEs based on arbitrarily defined cut-offs for parameters was not pursued since the dataset was highly heterogenous. An *ad hoc* evaluation of DEs was therefore developed to select the best candidate for further sequence design. It consisted first in identify which of the 17 selected Rosetta parameters could be used to best discriminate DEs by PCA, as shown in Figure 2.9. [106–108] PCA is a useful orthogonal linear transformation method which allows to reduce an original dataset of variables (17 parameters) and data points (5494 DEs) to a new coordinate system where projection along the major axis (Principal Components, PC) reflects the amount variance encoded by each variable. Thus, parameters which account for higher variance in the dataset, and therefore best discriminate DEs, are those that present higher projections (or loadings) along the subspace of first PCs (typically PC1-PC2). From eigenvalue decomposition of the correlation matrix of 17 parameters, the resulting scree plot presented a steep curve with a bend at PC3 but with no abrupt flattening until PC17. The PC1-PC2 subspace biplot, whose PCs encode for 42% of total variance in the dataset, revealed two clusters corresponding to astacin and remaining peptide/small-protein DEs. The orthogonal pairs of parameters that presented highest loadings along the two clusters were *i)* $Score_{total}$ and sequence length L (*correlation, corr*= -0.52), interpreted as larger scaffolds having more negative scores, *i.e.* being more favourable since lower REU values are considered "good" in Rosetta; *ii)* constraints *k* and $Score_{diAla}$ (*corr*= 0.27), interpreted as DEs with higher constraints tending to present less favourable diAla scores. These orthogonal pairs of parameters were used in the next step of evaluation. The original dataset was projected along the original subspace of pair *i)* and a linear combination of pair *ii)*, as shown in Figure 2.10.[12]

[12] A linear combination of parameters from pair *ii)* was chosen since these two parameters are positively correlated: An increase in one parameter reflects an increase in the other.

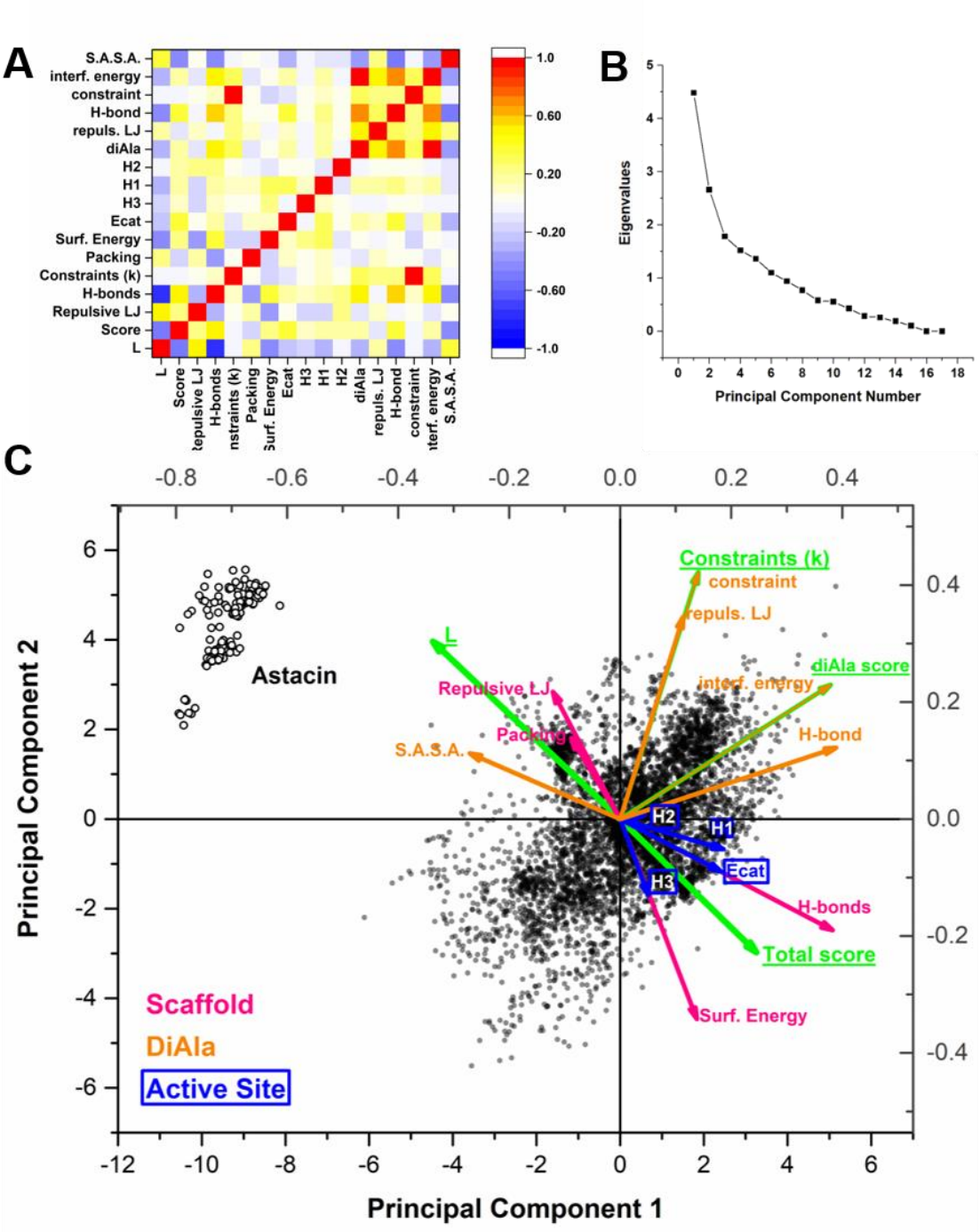



**Figure 2.9 –** PCA of peptide/small-protein scaffolds designed with the MA(M)$_{AS:diAla}$ model. A – Correlation matrix of 17 parameters, where each element was calculated over 5494 DEs. Pairwise correlation values were scaled between -1 for anti-correlated parameters and 1 for totally correlated parameters. Details of used parameters described in Section 2.2. B – Corresponding scree plot obtained from eigenvalue decomposition of the correlation matrix. C – Biplot of parameter loadings (top and right axis) and DE projections (bottom and left axis) along PC1 and PC2. DEs represented as black dots. Parameter loadings represented as vectors, coloured with respect to the corresponding features: scaffold-related in magenta (4), substrate-related in orange (5) and AS residues in blue (4). Parameters selected for further DE analysis in green (4).

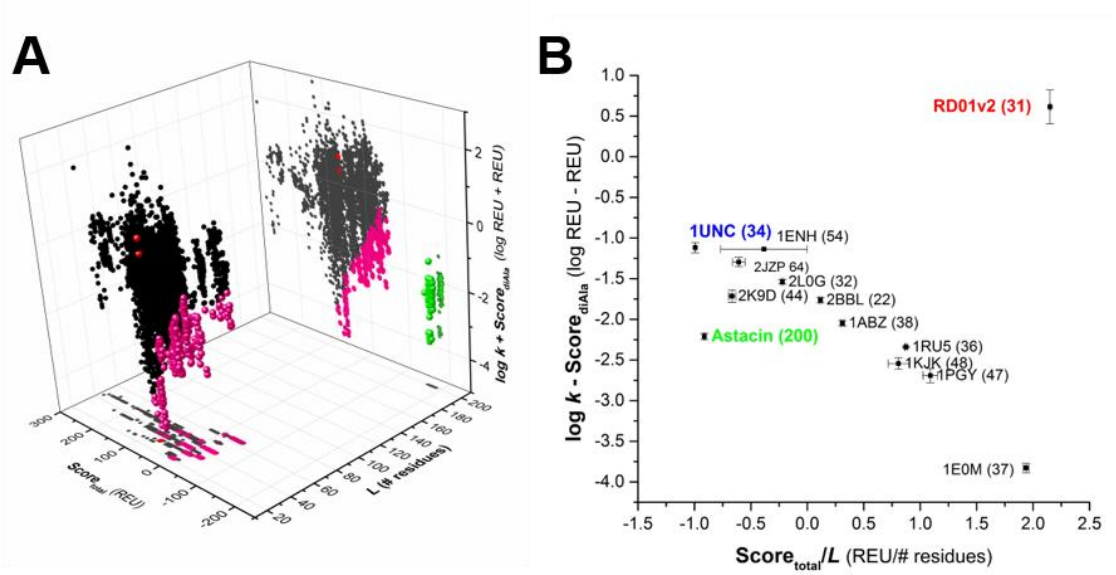


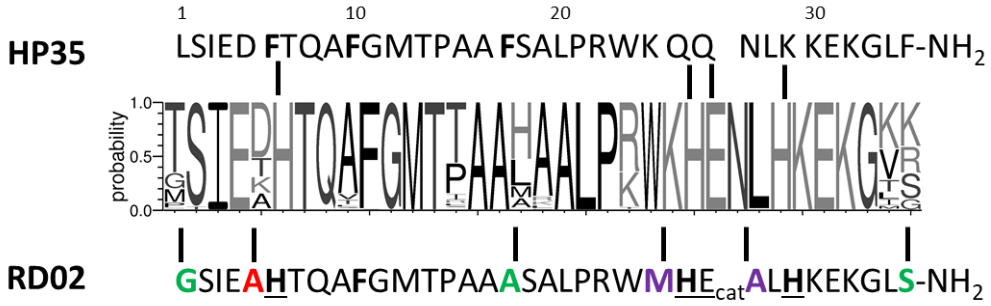

**Figure 2.10 –** Evaluation of peptide and small-protein designs.
A – Projection of DEs along the subspace of parameters $Score_{total}$, chain length (L), and linear combination of constraints (log *k*) and $Score_{diAla}$. DEs from the negative control (RD01v2) represented as red spheres, from positive control (astacin) as green spheres and from the $DE_{P10}$ subset in magenta spheres (see full text for $DE_{P10}$ definition). Remaining DEs represented as black dots with grey projections. B – Hyperplane projection of scaffolds included in the $DE_{P10}$ subset. PDB identifier and chain length in parenthesis. Error bars correspond to variations between the respective DEs. Best candidate scaffold 1UNC identified in blue.

The two clusters of DEs vary in terms of sequence length and scaffold score, although presenting similar *k*/$Score_{diAla}$ combinations to the remaining designs. The later shows large dispersion of $Score_{total}$ but relatively smaller variation of *k*/$Score_{diAla}$ ratio. Evaluation of the designs was focused at this point to the first decile of DEs ($DE_{P10}$) with lower $Score_{total}$ (more negative) and low *k*/$Score_{diAla}$, thus corresponding to the set of top best scaffolds. This set consisted in DEs from 12 scaffolds, including astacin, with varying sequence length and predominantly an all-α fold. Inclusion of RD01v2 results allowed to clearly identify its unfavourable features as a scaffold for the $MA(M)_{AS}$ model, such as positive $Score_{total}$ and high *k*/$Score_{diAla}$ combination. Selection of the best candidate was therefore based on the scaffold that presented scores further from those of RD01v2 and closer to those of astacin. The scaffold which best met these criteria was the human villin headpiece C-terminal subdomain - commonly termed HP35 (PDB 1UNC) - with a $Score_{total}$ more negative than control astacin, although with a relatively high *k*/$Score_{diAla}$. HP35 was therefore selected for further analysis, as described in the next section.

### 2.3.7 Design of RD02

The human HP35 scaffold, whose sequence is shown in Scheme 2.5, is part of the actin-binding protein villin.[109] Its chicken homologue is a model system for protein folding since it is the smallest protein that folds cooperatively, having been extensively characterized with both computational [110–116] and experimental approaches. [117–122] HP35 designs included in the $DE_{P10}$ data set correspond to two sequence variants for the designed $His_1$ residue, F6H and F17H,

as shown in Figure 2.11. The remaining residues $His_2$, $His_3$ and $Glu_{cat}$ were designed at identical positions in the scaffold. The HP35 proved to be quite "designable", since 5 additional sequence variants were also modelled but not present in the $DE_{P10}$ set.[13]




**Scheme 2.5 –** HP35 design into RD02.
Sequence logo with relative amino acid frequencies for each scaffold position. Modelled residues underlined and Zn(II)-coordinating residues in bold, as well as conserved phenylalanine residues. Sequence changes identified by vertical bars: DEs in green, non-Zn(II)-coordinating residues in red and scaffold stabilization residues in purple.

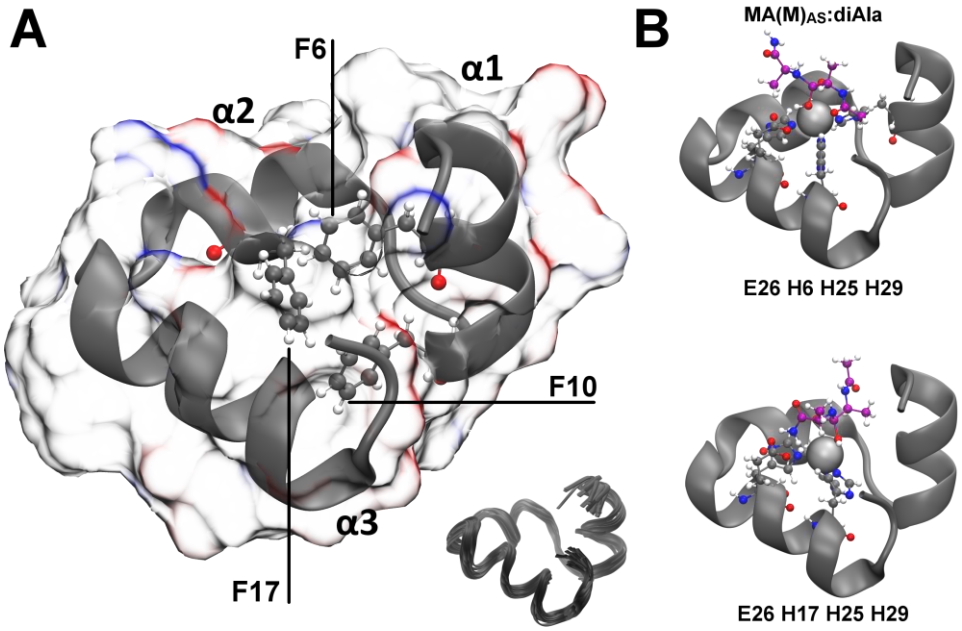



**Figure 2.11 –** HP35 design.
A – Scaffold cartoon representation. Secondary structure elements identified and conserved phenylalanine residues in bold. NMR states in ribbon representation shown in bottom right. B – DEs of sequence variants contained in the $DE_{P10}$ set of designed peptide/small-protein scaffolds, with variations at the level of the $His_1$ residue, H6 (top) or H17 (bottom).

Both variants present at least one sequence change for the highly conserved F6, F10 and F17 residues, whose hydrophobic collapse is attributed to be major driving force of HP35 folding.[118]

[13] Screening of chicken X-ray structure (PDB 1VII) and the closely related human advillin (PDB 1UND) did not yield UMs, which again points to the sensitivity of the screening methodology to small differences in backbone conformations, as in the case of the Sp1f2 scaffold.

Substitution of these residues for leucine is known to lead to significant destabilization of the scaffold, with F17L leading to the most destabilizing variants (substitution for other amino acids leads to unfolded or misfolded variants).[109,118] Analysis of UMs obtained for the F6H variant revealed properly defined Zn(II) coordination geometries through the $N\varepsilon^2$ atoms of $His_{1\text{-}3}$ residues, as shown in Figure 2.12.

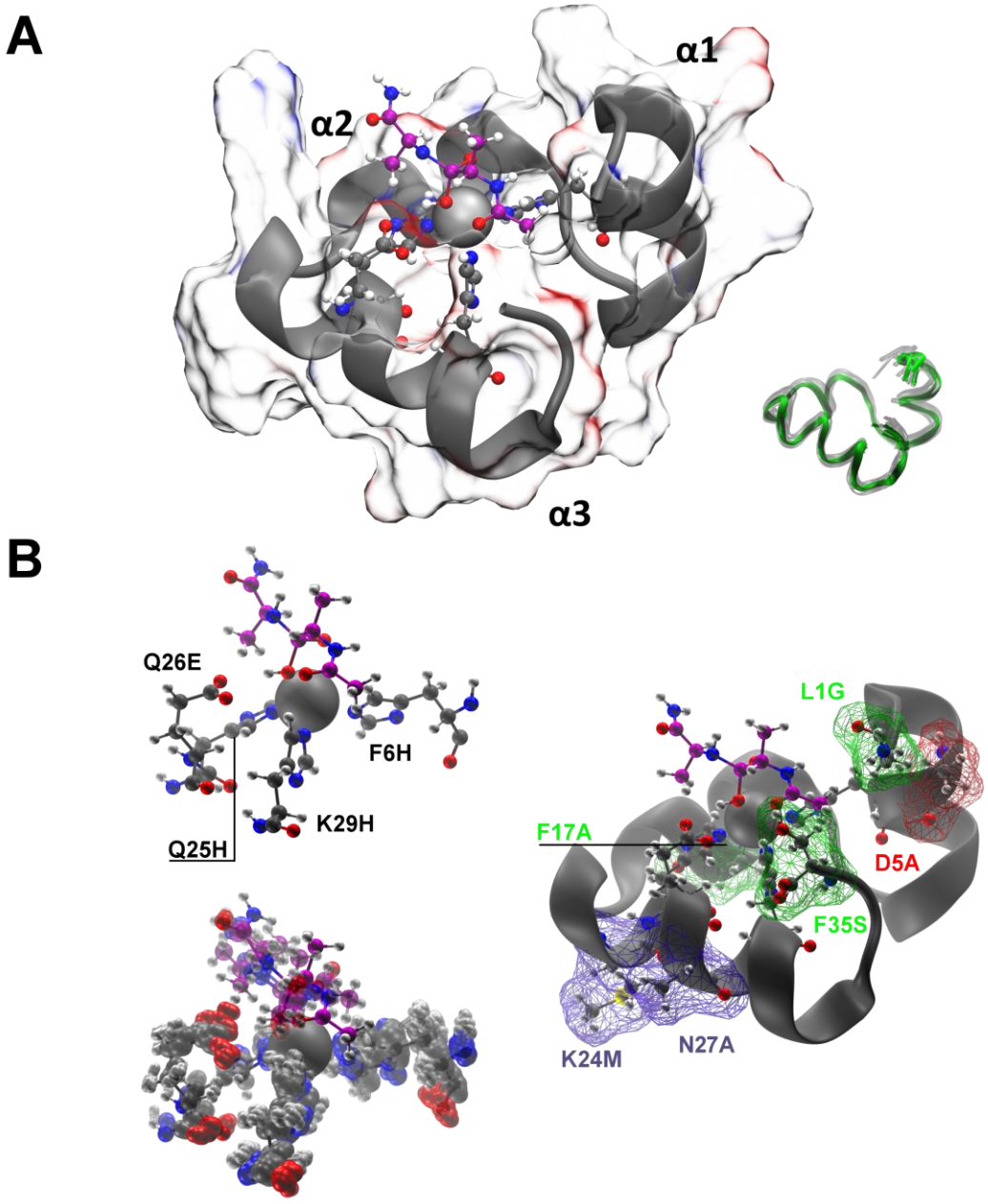



**Figure 2.12 –** Design of RD02.
A – Scaffold representation of obtained UMs for HP35, H6 variant with modelled $MA(M)_{AS}$. Secondary structure elements identified in bold and corresponding NMR states (6, 20, 22, 25) in green ribbon. B – AS detail of the DE with lower $Score_{total}$, diAla and AS sequence changes identified in bold (left, top). Detail of residue fluctuations and diAla conformers designed within the range of sampled geometrical parameters (left, bottom). Residue fluctuations and only one diAla structure per conformer shown in transparent representation for clarity. Detail of designed sequence changes identified in bold and shown in coloured wireframe (right).

A methionine residue in position 12 was positioned below the designed AS (not shown), resembling the conserved “met-turn” of MA(M) members. A higher number of UMs was produced in comparison with RD01v2, corresponding to scaffolds with slight atomic fluctuations of designed residues, four different diAla conformers with small atomic fluctuations (> 4000 placements), and four distinct NMR states.

Regarding interactions with the substrate, the F6H variant presents diAla in extended conformation and placed along the scaffold surface, while in F17H it is positioned closer to the first α-helix and in bent conformation. Given the least destabilizing sequence change and diAla placement resembling those found in protease-substrate complexes, the F6H variant was selected for sequence design.

DEs from the F6H variant yielded 10 sequence changes with more than one proposed amino acid identity, as shown in Scheme 2.5. Native residue identity was kept when it was also proposed by the Rosetta *enzyme design* program, such as F10, P14, R22 and L34. The proposed S18A sequence change was excluded since the native Ser residue interacts through a H-bond with native R22 residue and also to avoid formation of poly-Ala sequence motifs. The F17A sequence change was accepted over the F17L in order to decrease steric clashing with the metal centre.[14] The sequence changes L1G and F35S were accepted to optimize interactions with diAla by reducing side chain size in the termini of α1 and α3. In addition to the sequence changes proposed by the Rosetta *enzyme design* program, other modifications were done to address the stability of the scaffold: D5A (α1) to eliminate competing Zn(II)-binding residues in the vicinity of designed $His_3$, similarly to the C5G sequence change done in RD01, RD01v2 design; K24M and N27A (α3) to increase scaffold stability, since they are known to contribute for the formation of hyper stable HP35 variants.[111] The scores of HP35 remained practically unchanged since the initial screening through the sequence design stage and after the additional D5A, K24M and N27A sequence changes ($Score_{total}$ -33.9 to -33.2 REU), suggesting that no destabilizing amino acids were introduced during sequence design. The final peptide sequence, RD02, is shown in Scheme 2.5 and was selected for synthesis and experimental charachterization together with the RD01 and RD01v2 peptides.

## 2.4 Conclusion

In this chapter, MPs from the MA clan were characterized in terms of their sequence-structure-dynamics relationships in order to capture essential aspects of their enzymatic function. Besides the conserved sequence motifs, the AS were shown to be structurally conserved in regions of the scaffold with low fluctuations. Characterization of conserved geometric features between AS residues allowed to develop two computational AS models, one of Adamalysin II and another of a

[14] The effect of F10 and F17 sequence changes will be further discussed in Chapter 6.

general MA(M). The general MA(M) model was shown to reproduce the AS features of the native MP astacin and was successfully used to screen and design a total of 43 peptides or small-protein scaffolds. One of such scaffolds was the ZF metallopeptide Sp1f2, whose redesign was made in two rounds, RD01 and RD01v2, the latter being used to address scaffold stability by incorporation of ZF-stabilizing sequence changes. The best candidate from the remaining 42 scaffolds in terms of Rosetta scoring function was identified, which corresponded to the model peptide HP35. The scaffold was extensively designed with the general MA(M) AS model, leading to the final sequence RD02. The employment of NMR structures as input allowed to capture the flexible features of small scaffolds and how successful design of the AS model was consequently affected. The employment of PCA allowed to identify the set of Rosetta scoring parameters that best discriminate between native MPs and small scaffold designs.

The three RD peptides were selected for synthesis, as described in the following Chapter 3. Given the relatively high number of sequence changes made for such small scaffolds, potentially destabilizing interactions could have been introduced, which was addressed by physicochemical characterization in Chapter 4 and evaluation of catalytic proficiency in Chapter 5. While the RD peptides closely reproduced AS geometries found in native MPs, the employed MM-based method provided limited clues if such interactions are kept under more realistic conditions (*i.e.* system properties in solution). Therefore, in Chapter 6 the computational design approach was re-evaluated by exploring the structural flexibility of RD peptides through simulation and experiments.

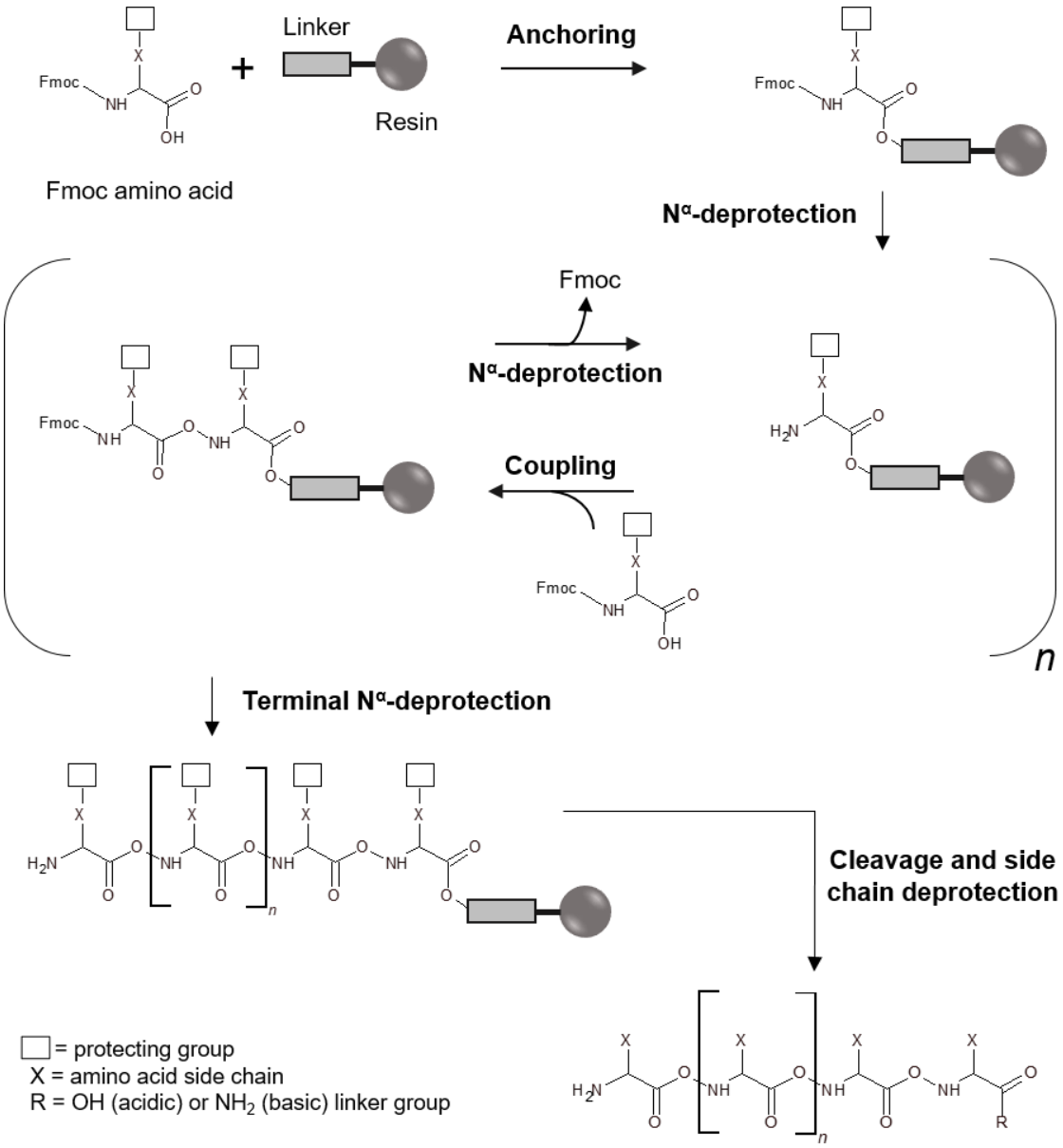
3

# 3. Synthesis and purification of Peptides

## 3.1 Introduction

Production of candidate sequences is a crucial step in any protein design project. Ideally, all designs that are considered for a given target function should be experimentally evaluated to address the robustness of the employed approach, while relying in high-throughput methods of protein purification and characterization.[123,124] However, in practical terms the evaluation of candidate sequences is usually limited to the set that can be obtained in soluble form through recombinant techniques for expression in biological systems. [23,87] While this may reduce the time and effort necessary to find the best catalyst, valuable information from failed designs may be lost merely because proper expression conditions are not met. For enzyme design endeavours, additional problems arise in case the target reaction is also performed by the host's molecular machinery, since even low-level contaminations of native enzymes can mask the observed catalytic activities and lead to false positive results.[125,126] Proteases are an example of this, given that they are ubiquitous and span a wide range of kinetic and substrate specificities. [25] Last, production of metalloproteins in biological systems requires additional treatment of expression crudes with chelators to ensure removal of endogenous metals, which otherwise could interfere with proper characterization of their metal-binding properties.

Peptides and small proteins designs benefit from their reduced size since production can be done through chemical methods and thus avoid some of the caveats of recombinant protein expression. Solid-phase peptide synthesis (SPPS) is a particularly suited technique for rapid and efficient production of peptides. Introduced by Merrifield in 1963 [127] with *t*-butyloxycarbonyl (commonly referred as Boc) based chemistry and later adapted for 9-fluorenylmethyloxycarbonyl (Fmoc)-based chemistry by Carpino in 1970's,[128] SPPS is now routinely used by the majority of peptide laboratories.[129] This iterative method is represented in Scheme 3.1 and further details will be given in Section 3.2. Nascent peptide chains are covalently attached to an insoluble polymeric support. The $N^{\alpha}$ of the nascent chain is protected by the base-labile Fmoc group, which is removed prior to carboxyl activation and coupling with the next amino acid in the sequence. Chain elongation occurs from the C to N terminal by successive cycles of deprotection and coupling, with removal of excess reagents through filtration and washing. After deprotection of the N-terminal amino acid, side chain deprotection concomitant with cleavage from the solid support is achieved though acidic conditions. This orthogonal scheme, *e.g.* removal of $N^{\alpha}$-protecting group under basic conditions and side chain deprotection under acidic ones, has been increasingly adopted for standard synthesis protocols given the less harsh conditions required for cleavage.[130]




**Scheme 3.1 –** Summary of Fmoc-based SPPS. Description of each step is made in main text.

Since its initial developments, Fmoc-based SPPS has been optimized and can now be automatized for production of peptides and small proteins.[131] It is thus a suitable method to produce the RD peptides described in Chapter 2 and their respective native sequences, shown in Scheme 3.2. SPPS also allows the possibility of introducing unnatural amino acids and other specific probes/modifications in posterior design stages. Nonetheless, it still faces limitations regarding longer chain lengths due to accumulation of by-products from incomplete coupling and deprotection reactions that lead to premature termination or truncated sequences. [132] This is usually attributed to steric hindrance or intra- and intermolecular aggregation phenomena, which can be minimized, for example with the aid of microwave heating, mixtures of solvents, special amino acids or employment of different solid supports.[133,134]

**Sp1f2** RPFMCTWSYCGKRFTRSDELQRHKRTHTGEK-$NH_2$
($C_{165}H_{258}N_{56}O_{45}S_3$, 3841.9 Da)

**RD01** RPFMGTWSYHGGRFTRSDELQRHKREHTGEK-$NH_2$
($C_{164}H_{249}N_{57}O_{46}S$, 3786.9 Da)

**RD01v2** RPFTGTWSYHGVTFTRSDELQRHKREHTGEK-$NH_2$
($C_{164}H_{248}N_{54}O_{48}$, 3743.9 Da)

**HP35** LSIEDFTQAFGMTPAAFSALPRWKQQNLKKEKGLF-$NH_2$
($C_{185}H_{286}N_{48}O_{49}S$, 3998.1 Da)

**RD02** GSIEAHTQAFGMTPAAASALPRWMHEALHKEKGLS-$NH_2$
($C_{164}H_{257}N_{49}O_{47}S_2$, 3730.9 Da)

**Scheme 3.2 –** Sequences of natives Sp1f2, HP35 and designed RD01, RD01v2 and RD02 peptides with C-terminal amidation ($NH_2$). Calculation of mass based on chemical formula.

Aggregation is dependent on local sequence motifs that occur during chain elongation. For example, sequences containing amino acids prone to form β-sheet elements (alanine, valine, isoleucine, asparagine, glutamine) may yield crudes with lower purity due to increased hydrogen bonding and hydrophobic interactions. Indeed, as it will be described in Section 3.3.1, synthesis of both native Sp1f2 and RD01 peptides was relatively straightforward in comparison with synthesis of RD01v2 described in Section 3.3.2 due to the increased β-sheet forming propensity of the later. However, difficulties are not restricted to peptides containing β-sheets, as it will be described in Section 3.3.3 for the synthesis of the all-α HP35 and RD02 peptides.

Although issues were encountered regarding coupling and deprotection cycles, the standard Fmoc-based methods employed resulted in complex crudes from which the target peptides were present as major species. This was observed by reversed-phase high-pressure liquid chromatography (HPLC), the commonly used purification technique used in combination with SPPS. [135] Peptides are first adsorbed in the stationary phase via hydrophobic interactions with the column matrix and then eluted differentially by decreasing the polarity of the mobile phase. The identity of peptides was confirmed by mass spectrometry analysis of isolated HPLC peaks, which allows for direct identification of their amino acid sequence. [136] The details of the synthesis and purification steps will be described in the following sections.

## 3.2 Materials and Methods

*Chemicals:* Fmoc-amino acids were purchased from CEM, Novabiochem (now Merk) and Iris Biotech GmbH. Rink Amide 4-methylbenzhydrylamine (MBHA) (100-200 mesh, loading 0.59 mmol/g), Rink Amide MBHA low-loading (LL, 100-200 mesh, loading 0.36mmol/g) resins and 2-(1H-Benzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate (HBTU) were obtained from Novabiochem. Trifluoroacetic acid (TFA), anisole, thioanisole, 1,2-ethanedithiol, triiso-

propylsilane (TIS), 1-Hydroxybenzotriazole (HOBt) and piperidine were acquired from Sigma-Aldrich (now Merck). Acetonitrile (ACN), dimethylformamide (DMF), diethyl ether, dichloromethane (DCM), N-methyl-2-pyrrolidone (NMP), N,N-diisopropylethylamine (DIEA), acetic anhydride and triethylamine (TEA) were obtained from different commercial suppliers. All reagents used were the highest grade available.

*Synthesis:* SPPS were done in an Initiator+ Alstra Automated Microwave Peptide Synthesizer (Biotage) or in a Liberty Microwave-Assisted Automated Peptide Synthesizer (CEM GmbH). The methods employed were based on protocols provided by the manufacturers (Biotage) or optimized for synthesis of long peptides (CEM), further details are given in Table 3.1. The Rink Amide MBHA and Rink Amide MBHA LL resins were used as solid supports and swelling was done with DCM followed by washing with DMF. 1. An excess of 4 equiv. amino acids, 3.9 equiv. HBTU, 4 equiv. HOBt and 8 equiv. DIEA was used to increase reaction yields. After the last coupling cycle resins were washed with DMF and DCM.

**Table 3.1 –** Methods and conditions used in the synthesis of peptides.

| | **Biotage** | **CEM** |
|---|---|---|
| **Reaction vial** | 10 mL | 45 mL |
| **Deprotection** | | |
| Reagents | 20% piperidine in DMF | |
| Conditions | 13 min or 20 min for “long-peptide”<br>room temperature | 5 min or 20 min for “long-peptide”<br>75 °C |
| **Coupling[a]** | | |
| Reagents | Activator: HBTU in DMF<br>Base: DIEA in NMP<br>Additive: HOBt in NMP | Activator: HBTU in DMF<br>Base: DIEA in NMP |
| Conditions | 5 min or 10 min for “long-peptide”<br>75 °C<br>60 min for histidine/aspartate/cysteine<br>room temperature | 5 min or 10 min for “long-peptide”<br>75 °C<br>histidine/cysteine/glutamate/aspartate<br>50 °C |
| **Acetylation** | - | 20% acetic anhydride in DMF |
| **Stirring** | Vortex | $N_2$ current |

a - Coupling cycles were doubled for arginine, threonine, proline, and tryptophan in CEM.

*Deprotection and Cleavage:* side chain deprotection and cleavage from resin was done with TFA/TIS/$H_2O$ 95:2.5:2.5 v/v (RD01v2), or TFA/thioanisole/1,2-ethanedithiol/anisole 90:5:3:2 v/v for peptides containing cysteine or methionine residues (Sp1f2, RD01, RD02 and HP35), $V_t$=10-20 mL for 2h under $N_2$ atmosphere. The resin-cleaved peptides contained an amide group at the C-terminal from the MBHA linker. Afterwards, TFA from the cleavage filtrate was removed under

$N_2$ current and added to cold diethyl ether to form peptide crude precipitates.[15] After filtration and washing with cold diethyl ether, the peptide crudes were dissolved in water, lyophilized and stored at -20 °C until purification.

*Purification:* crudes were purified by preparative reversed-phase HPLC (Agilent 1260 or Waters 2535 Quaternary gradient module with 2489 UV/Vis detector) with a C18 column (Phenomenex Jupiter 250 mm x 21.2 mm, 15 µm 300 Å) and mobile phase of solvent A ($H_2O$/TFA 99.9:0.1 v/v) and solvent B (ACN/$H_2O$/TFA 90:9.9:0.1 v/v). Crudes were dissolved in the minimum volume possible of solvent A to reach full solubilization (> 20 mL) or in a mixture of solvent A/solvent B 95:5 or 90:10 v/v in case of low solubility. Initial 1 mL injections were made as a test and subsequently the injection volume was increased up to 4 mL (5 mL loop) or 9 mL (10 mL loop). Prior to injection, crudes were filtered with 0.22 µm filters to remove aggregates. Chromatograms were obtained by tracking absorbance signal intensity at 220 nm and 280 nm channels with flow rates of 10 or 20 mL/min. Linear gradient methods were employed and optimized for each crude, starting at 0 to 20% and going up to 60% solvent B in slopes ranging between 1.25 to 5% solvent B/min. Peaks of interest were identified and manually collected after elution in one to three fractions (lower slopes were used for multiple fractions). After collection, samples were lyophilized and stored at -20 °C until usage. Approximately 10 mg of purified peptide were typically recovered.[16] Purity and identity of peptides were determined by analytical HPLC (Agilent 1100 Series or Hitachi LaChrom Elite) and mass spectrometry analysis, respectively. Samples were dissolved in MQ water or solvent A for analytical HPLC in C12 (Phenomenex Jupiter Proteo, 250 mm x 4.6 mm, 4 µm, 300 Å) or C18 (Phenomenex Jupiter 250 mm x 4.6 mm, 15 µm, 300 Å and Discovery HS 250 mm x 4.6 mm, 5 µm) columns, with the same gradient methods used in preparative HPLC and 1 mL/min flow rate. The same steps were applied in re-purification of isolated peaks and recovery of RD01.[17]

*Fmoc removal*: In synthesis where a Fmoc group was kept, the removal was carried out after HPLC purification by addition of 5% v/v TEA to solutions containing the Fmoc fractions for 1h (HP35) or extended to 3h (RD01v2 and RD02). TFA was then added to the reaction for neutralization (6 < pH < 8). Afterwards, the samples were filtered with 0.22 µm filters and the soluble fraction stored at -20 °C until purification as described above.

[15] In the first synthesis of RD01v2 cold diethyl ether was added directly to the TFA solution due to the low solubility of cleavage filtrate.

[16] Assuming efficiencies of 90% for each coupling cycle (and 100% for each deprotection cycle), synthesis yields (mg purified/mg calculated x 100) were < 27% for Sp1f2, RD01 and RD01v2 and < 43% for HP35 and RD02. The example of calculations for RD01 (Biotage) is given: 946 mg (product weight) x $0.9^{31}$ (90% coupling over 31 cycles) = 36.1 mg (calculated). Synthesis yield: 10 mg (weighted maximum)/36.1 mg (calculated)= 27%.

[17] RD01 assays were pooled once finished and the peptide recovered. Treatment with 10 mM EDTA and a drop of 1M HCl overnight at room temperature was made to obtain the Zn(II)-free peptide. Solutions were filtered and injected directly into preparative HPLC.

*Mass analysis:* mass characterization of the peptides was done by time-of-flight (TOF) MS with electrospray ionisation (ESI) - Spectropole, Aix-Marseille Université and UniMS (Waters Synapt G2 HDMS), Mass spectrometry Unit ITQB/IBET (Thermofinnigan ESI-LTQ and ESI-LCQ) services or; matrix-assisted laser desorption/ionization (MALDI) - UniMS, Mass spectrometry Unit ITQB/IBET (Applied Biosystems Sciex 4800plus, TOF/TOF) and Laboratório de Análises, REQUIMTE (Voyager-DE PRO, TOF) services. Solids were dissolved in diluted solutions of either formic acid, acetonitrile or methanol/ammonium acetate before ionization (ESI) or directly eluted in CHCA matrixes (MALDI) and acquired in positive ion mode.

## 3.3 Results and Discussion

### 3.3.1 Synthesis and purification of Sp1f2 and RD01

The native peptide Sp1f2 has been extensively studied and described in the literature, including its production through Fmoc-based SPPS.[137,138] No specific details have been raised regarding difficult couplings within the full polypeptide chain, which prompted the usage of standard coupling and deprotection methods for all the 31 cycles (Biotage). It should be noted that synthesis done in Biotage followed the protocols provided by the manufacturer as the instrument had not been tested at the time for synthesis of long peptide sequences, while those made in CEM (see below) had already been optimized in the laboratory. As shown in Figure 3.1, after cleavage from the Rink Amide MBHA resin, the HPLC chromatogram of peptide crude contained multiple peaks, with the major one eluting at ~28% solvent B. This peak was collected in a single fraction, with the corresponding analytical HPLC revealing the presence of a small impurity.

ESI-MS analysis confirmed the identity of the target sequence (3841 Da) with the presence of a truncated sequence (3582 Da) with less two amino acids, corresponding to the observed by-product which could not be removed even after an additional round of HPLC purification (not shown). No further rounds were attempted since a significant portion of the target peptide was lost when discarding fractions containing the by-product. As will be described in Chapter 4, this batch was used only for control folding and stability assays and the results obtained are in agreement with the literature, indicating small interference from the containing impurity.

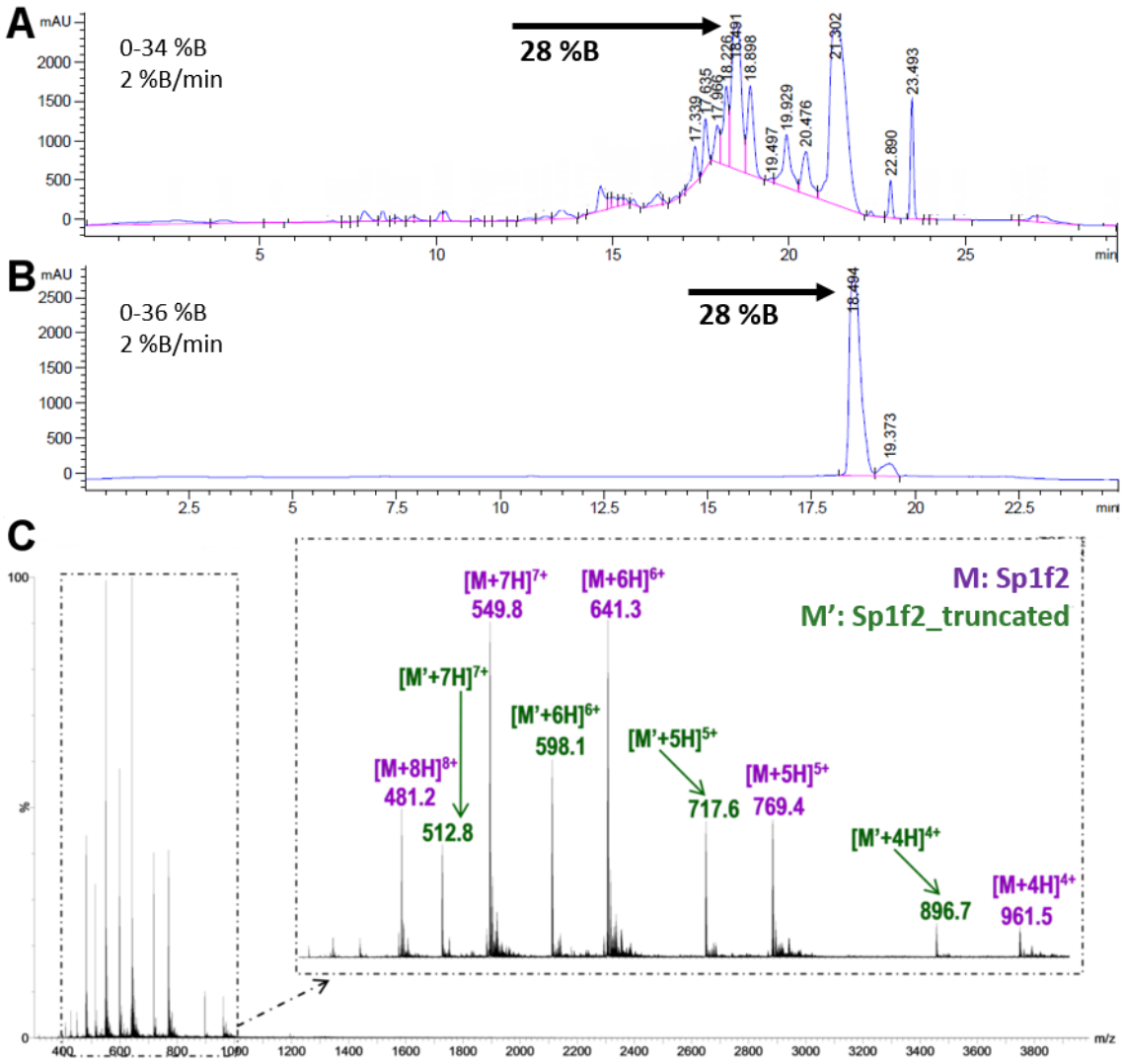



**Figure 3.1 –** HPLC purification and MS identification of Sp1f2 peptide obtained by SPPS. A – Preparative HPLC chromatogram of peptide crude (Biotage). Collected peak and eluting conditions are identified by arrows, details of gradient methods used in top left of chromatograms. B – Corresponding analytical HPLC chromatogram of collected peak. Absorbance signals monitored at 220 nm. C – Mass spectrum of collected peak obtained by ESI-MS. Compounds identified in top right: $[Sp1f2 +4H]^{4+}$ 961.6(calculated)/961.5(measured); $[Sp1f2 +5H]^{5+}$ 769.5/769.4; $[Sp1f2 +6H]^{6+}$ 641.4/641.3; $[Sp1f2 +7H]^{7+}$ 549.9/549.8 and $[Sp1f2 +8H]^{8+}$ 481.3/481.2.

Given the small number of introduced sequence changes (C5G, C10H, K12G), the RD01 peptide was also synthesized with the same methods used for Sp1f2, except for doubled coupling steps in the last three cycles to prevent truncated sequences (Biotage). HPLC of peptide crude shown in Figure 3.2 revealed again multiple peaks, with the major one eluting at ~26% solvent B. After purification, ESI-MS analysis confirmed that this peak corresponded to the target sequence (3786 Da), although containing small impurities (not shown). An additional round of preparative HPLC was required to isolate the target sequence (not shown), as shown by analytical HPLC and ESI-MS analysis (3789 Da).

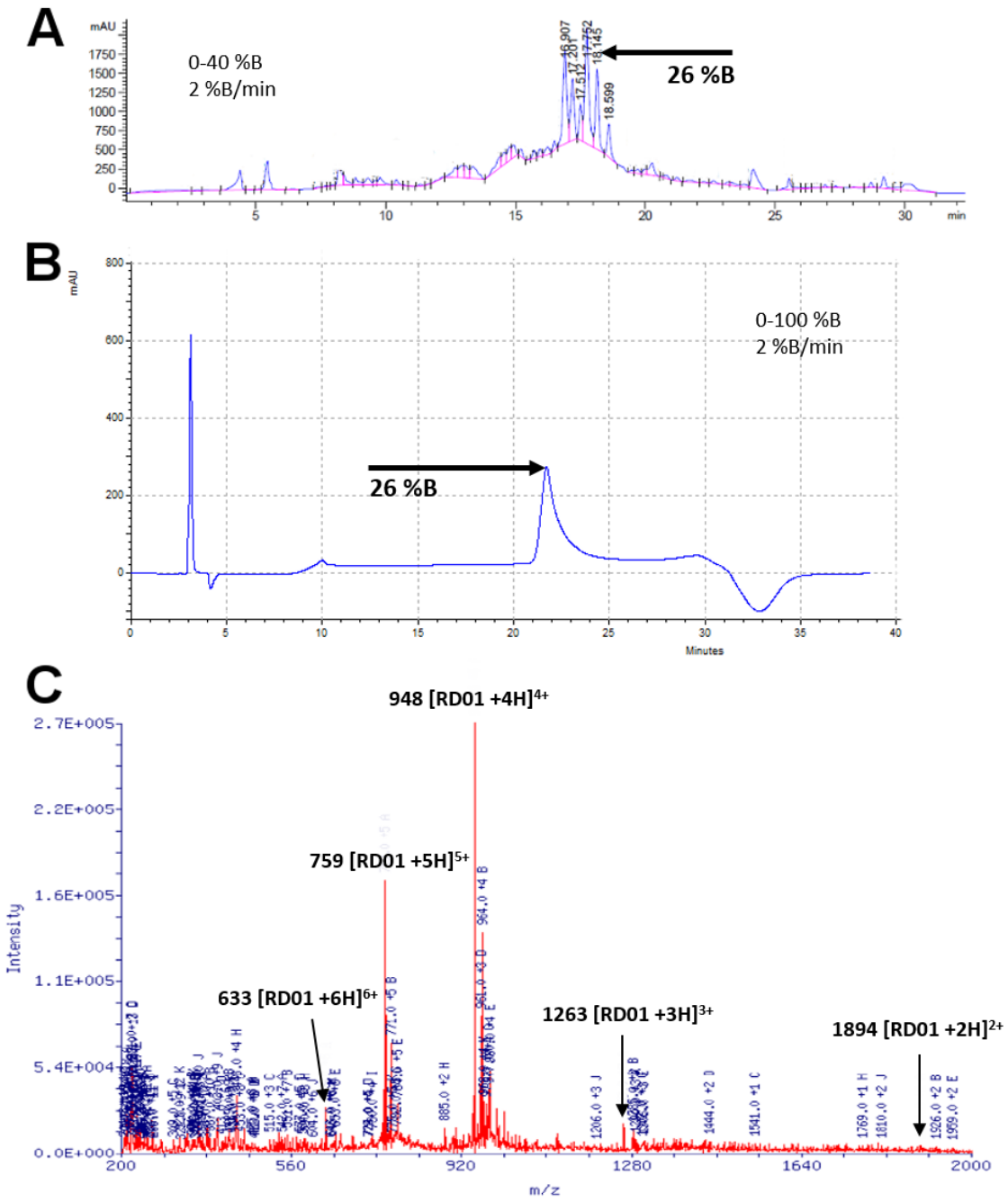



**Figure 3.2 –** HPLC Purification and MS identification of RD01 peptide obtained by SPPS.
A – Preparative HPLC chromatogram of peptide crude (Biotage). Collected peak and eluting conditions are identified by arrows, details of gradient methods used in the chromatograms. B – Corresponding analytical HPLC chromatogram of collected peak. Absorbance signals monitored at 220 nm. C – Mass spectrum of collected peak obtained by ESI-LTQ MS. Compound was identified based on: *m/z* $[RD01 +2H]^{2+}$ 1894.6(calculated)/1894.0(measured); $[RD01 +3H]^{3+}$ 1263.4/1263.0; $[RD01 +4H]^{4+}$ 947.8/948.0; $[RD01 +5H]^{5+}$ 758.4/759.0 and $[RD01 +6H]^{6+}$ 632.2/633.0.

Optimization was approached in the following synthesis of RD01 (CEM). Long-peptide coupling methods were used from the 15th to the 31st cycle to prevent formation of truncated sequences, where increased efficiency is attempted by extending reaction times. Elongation of truncated sequences was also prevented through acetylation of unreacted chains from the 15th to the

30th cycles. The Rink Amide MBHA LL resin was used to decrease interchain interactions on resin beads that may lead to aggregate formation. As shown in Figure 3.3, the HPLC of peptide crude presented a single major peak eluting at ~26% solvent B, which was collected in three fractions.

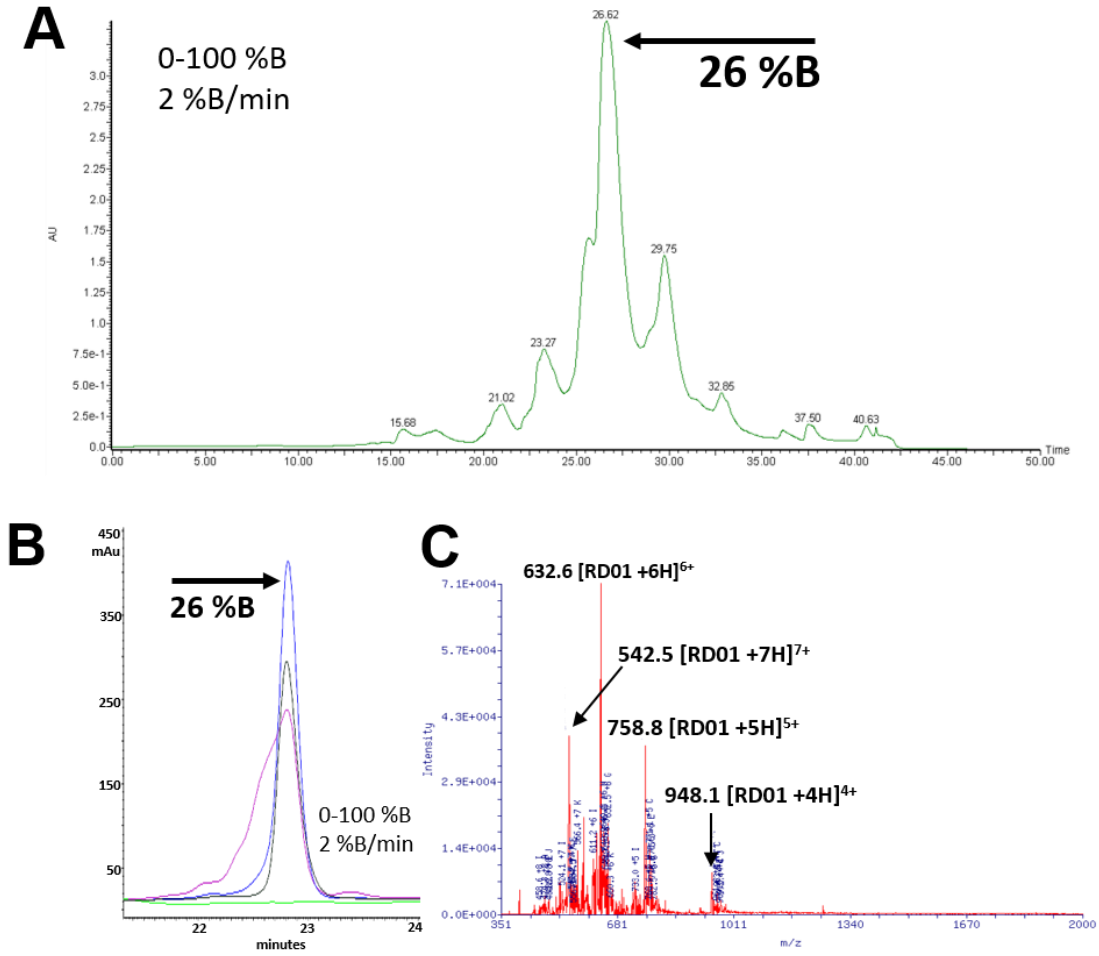



**Figure 3.3 –** HPLC purification and MS identification of RD01 peptide obtained by SPPS with optimized conditions.
A – Preparative HPLC chromatogram of peptide crude (CEM). Collected peak and eluting conditions are identified by arrows, details of gradient methods used in the chromatograms. B – Corresponding analytical HPLC chromatogram of collected peak in three fractions: initial (magenta), central (blue) and final (black). Absorbance signals monitored at 220 nm. C – Mass spectrum of collected peak obtained by ESI-LTQ MS. Compound was identified based on: *m/z* $[RD01 +4H]^{4+}$ 947.8(calculated)/948.1(measured); $[RD01 +5H]^{5+}$ 758.4/758.8, $[RD01 +6H]^{6+}$ 632.2/632.6 and $[RD01 +7H]^{7+}$ 542.0/542.5.

The presence of closely-eluting by-products difficulted the collection of this peak, requiring again two rounds of HPLC purification (not shown) until purity of the central fraction was confirmed by analytical HPLC and ESI-MS analysis (3789 Da). An additional synthesis using the same conditions was done and the obtained crudes were similar, although in that case one round of purification was enough to obtain pure peptide (not shown).

### 3.3.2 Synthesis and purification of RD01v2

In order to tackle eventual formation of aggregates during synthesis of RD01v2 that could originate from the sequence changes made to the previous RD01 version (M4T, K12V and R13T), the Rink Amide MBHA LL resin with lower loading was initially used (Biotage). However, due to high volume increase upon resin swelling this approach was not continued.[18] Instead, the original Rink Amide MBHA resin was used in a second attempt. To prevent decreased coupling efficiencies upon chain elongation, the long-peptide coupling methods were employed from the 15$^{th}$ to the 31$^{st}$ cycles. This approach did not prove to be useful, since HPLC chromatograms of peptide crude presented two major peaks, which eluted at ~28 and 29% solvent B, respectively (not shown). ESI-MS analysis of the first peak purified revealed a 3998 Da compound, which was assigned to the target sequence (3743 Da) with an additional Fmoc group (223 Da) and a $K^+$ ion (39 Da).[19] This was attributed to inefficient Fmoc cleavage during the final deprotection step. Removal was therefore attempted posteriorly on the purified fractions by treatment with 5% TEA. The resulting major peak eluted at ~33% solvent B, although the presence of closely eluting impurities required two additional rounds of HPLC purification, yielding a low amount of recovered peptide.

In the following synthesis of RD01v2 (CEM) additional measures were adopted to obtain a simpler peptide crude. Capping of truncated chains was done only at each 5 cycles (5$^{th}$ to 30$^{th}$) through acetylation. Rink Amide MBHA LL was selected since larger reaction vials were used. Long-peptide coupling methods were used as before. As shown in Figure 3.4, the peptide crude revealed a major peak mixed with closely-eluting small peaks at 29% solvent B, which again prevented efficient separation even for optimized gradients.

A new synthesis was attempted where Fmoc deprotection in the last cycle was not done to help identify the peptide (CEM). Capping of unreacted chain was increased from the 15$^{th}$ to the 30$^{th}$ cycle as initially made for RD01. The resulting crude chromatograms presented a major peak eluting at ~34% solvent B, which was assumed to be the target sequence containing the Fmoc group given the increased hydrophobic character. A coarse purification of this peak was done, followed by a 5% TEA treatment for Fmoc removal. The resulting peak eluted at 29% solvent B and as shown in Figure 3.4, the corresponding analytical HPLC chromatogram of the central fraction was pure, with target sequence confirmed by MALDI-TOF/TOF analysis (3745 Da). The synthesis of RD01v2 proved to be more challenging than previous RD01 and Sp1f2 peptides. Peptide crudes presented a higher amount of closely-eluting impurities suggesting decreased coupling efficiencies.

[18] ESI-MS of major peak eluting at 29% solvent B was inconclusive due to low resolution.
[19] ESI-MS of middle fraction from 28% solvent B, full mass (4006 calculated/3998 measured), series +3 (1335/1333) and +4 (1001/1000).

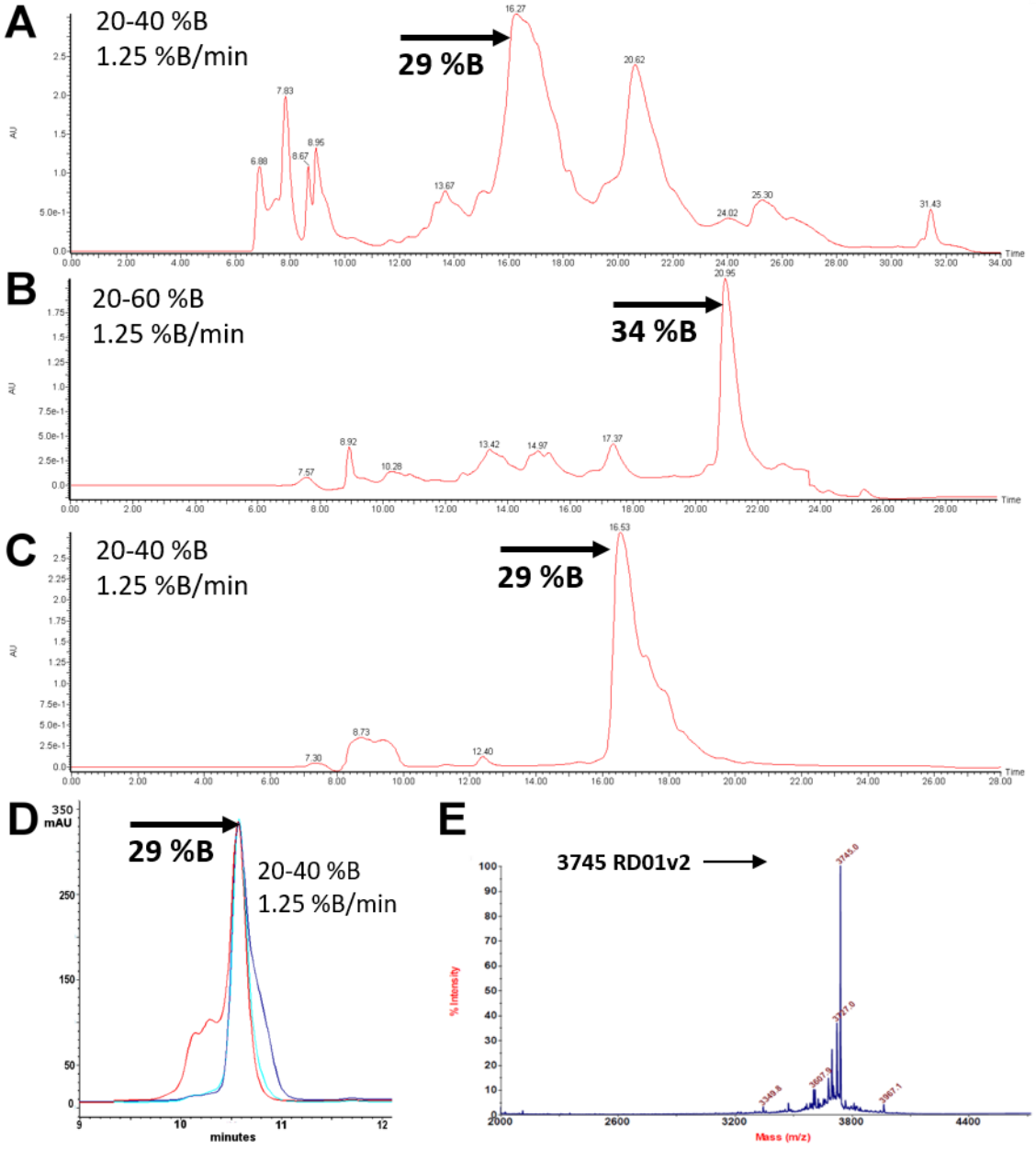



**Figure 3.4 –** HPLC purification and MS identification of RD01v2 peptide obtained by SPPS. A – Preparative HPLC chromatogram of peptide crude (CEM). Collected peak and eluting conditions are identified by arrows, details of gradient methods used in top left of chromatograms. B – Preparative HPLC chromatogram of peptide crude (CEM) with a Fmoc group attached at the N-terminal. C- Preparative HPLC chromatogram of collected peak after TEA treatment for Fmoc group removal. D - Corresponding analytical HPLC chromatogram of collected peak in three fractions, initial (red), central (cyan) and final (blue). Absorbance signals monitored at 220 nm, details of gradient methods used in the chromatogram. E – Mass spectrum of central fraction obtained by MALDI-TOF/TOF MS. Compound was identified based on: *m/z* [RD01v2 + H]$^+$ 3743.9 (calculated)/3745 (measured).

### 3.3.3 Synthesis and purification of HP35 and RD02

Production of HP35 through SPPS has been described in the literature using standard Fmoc-based chemistry.[109] As in the case of Sp1f2, no difficult couplings over the entire sequence were anticipated. The approach followed was therefore similar to RD01v2 synthesis, with long-

peptide methods used from the 15th to the 35th cycles, and a Fmoc group bound to the N-terminal to facilitate purification. No acetylation of unreacted chains was done. After synthesis (Biotage), analytical HPLC chromatograms of peptide crude shown in Figure 3.5 revealed a major peak eluting at ~55% solvent B, which was identified by MS as the peptide containing the Fmoc group.[20]

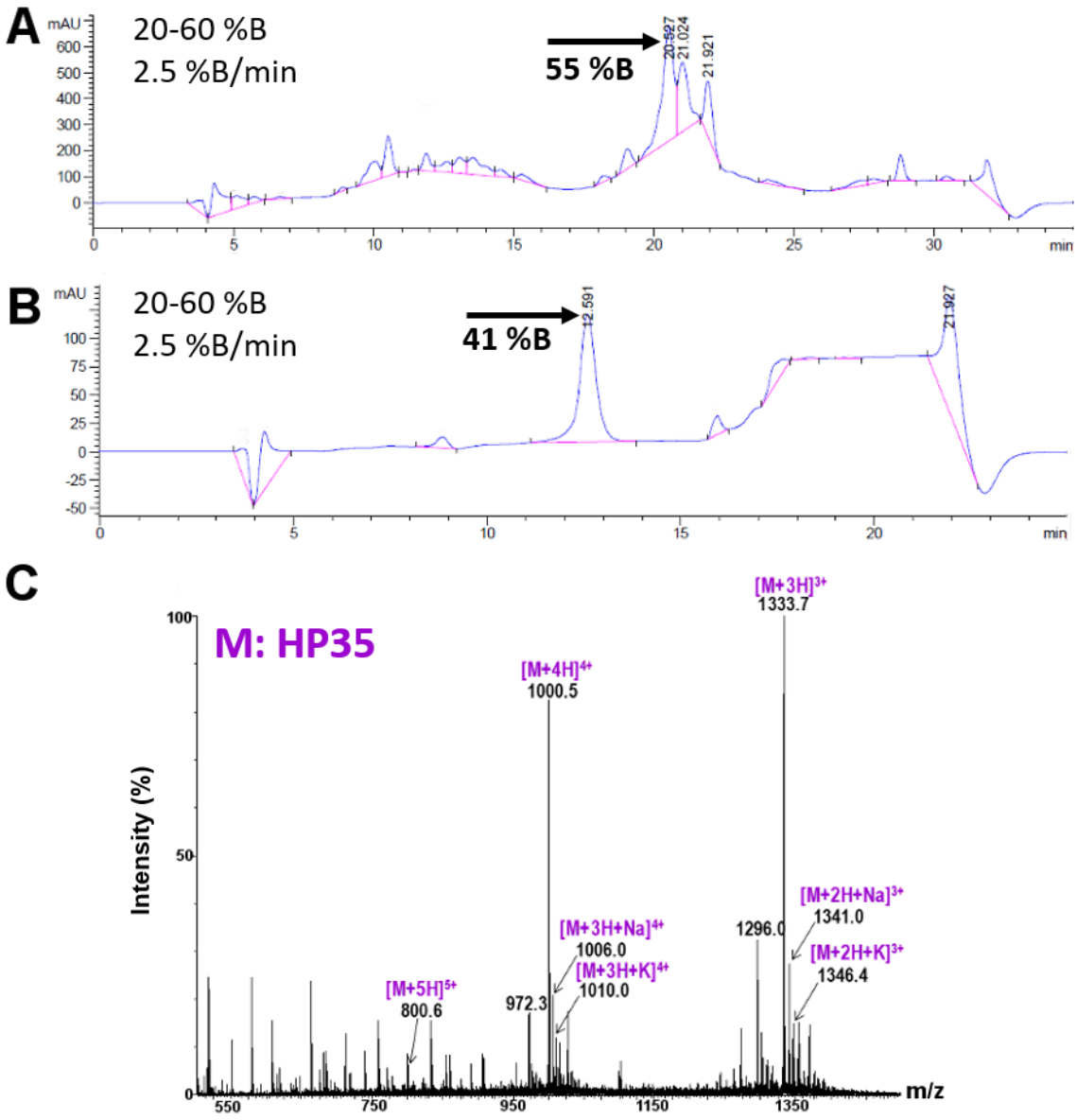



**Figure 3.5 –** HPLC purification and MS identification of HP35 peptide obtained by SPPS. A – Preparative HPLC chromatogram of peptide crude with a Fmoc group attached at the N-terminal (Biotage). Collected peak and eluting conditions are identified by arrows, details of gradient methods used in top left of chromatograms. B – Corresponding analytical HPLC chromatogram of new eluting peak after TEA treatment for Fmoc group removal. Absorbance signals monitored at 220 nm. C – Mass spectrum of collected peak obtained by ESI-MS. Compound was identified based on: *m/z* [HP35 +3H]$^{3+}$ 1333.9 (calculated)/1333.7 (measured); [HP35 +4H]$^{4+}$ 1000.7 /1000.5 and [HP35 +5H]$^{5+}$ 800.7/800.6.

[20] ESI-MS results: [HP35-Fmoc +3H]$^{3+}$ 1407(calculated)/1407 (measured); [HP35-Fmoc +4H]$^{4+}$ 1055/1056 and [HP35-Fmoc +5H]$^{5+}$ +5 844/845.

Removal of the Fmoc group from the crude was done with 5% TEA treatment, after which a new peak eluting at ~41% solvent B appeared. After collection of this peak, the analytical HPLC chromatogram presented no detectable impurities and ESI-MS analysis confirmed the identity of the peptide with no Fmoc group bound (3998 Da). The higher volume percentage at which HP35 elutes in comparison to Sp1f2 reflects the increased hydrophobic character of the former, which can be attributed to the higher number of hydrophobic residues such as the four phenylalanine residues present in its sequence.

Synthesis of RD02 (Biotage) was based on the methods used for HP35, except for the inclusion of a Fmoc group at the N-terminal. Problematic sequence motifs were not considered, although decreased yields were anticipated due of the relatively high number of sequence changes made. As shown in Figure 3.6, preparative HPLC chromatograms revealed a major peak eluting at ~35% solvent B, which was recovered in three fractions. Analytical HPLC chromatograms (not shown) and ESI-MS analysis revealed the presence of three compounds: target sequence (3730 Da), a presumable peptide-ion adduct (3793 Da) and the target sequence with an additional Fmoc group present (3900 Da). This was previously observed for RD01v2, where the unexpected presence of the Fmoc was attributed to inefficient removal in the last coupling cycle. Therefore, the sample was subjected to TEA treatment as described before. The final soluble fraction was confirmed by MALDI-TOF analysis to correspond to the target sequence (3731 Da). The lower solvent B percentage at which RD02 eluted in comparison to HP35 reflects a decreased hydrophobic character, which is attributed to the two F6H and F17A sequence changes made in order to accommodate the designed AS.

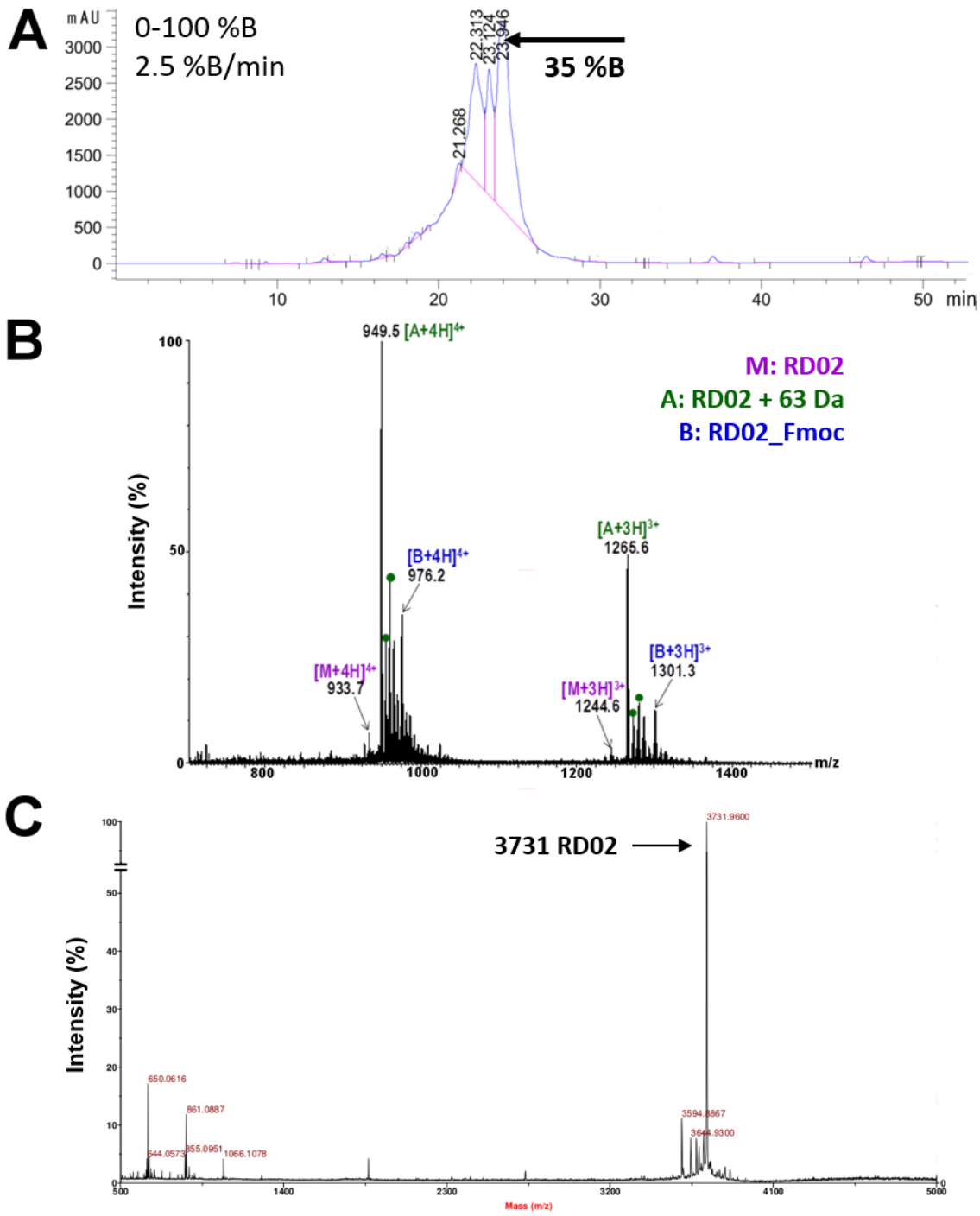



**Figure 3.6 –** HPLC purification and identification of designed RD02 peptide obtained by SPPS. A – Preparative HPLC chromatogram of peptide crude (Biotage) with a Fmoc group attached at the N-terminal. Collected peak and eluting conditions identified by arrows, details of gradient methods used in top left of chromatograms. Absorbance signals monitored at 220 nm, details of gradient methods used in inset. B – Mass spectrum of collected peak obtained by ESI-MS. Compound identified based on: *m/z* $[RD02 +3H]^{3+}$ 1244.8 (calculated)/1244.6 (measured) and $[RD02 +4H]^{4+}$ 933.8/933.7. C - Mass spectrum of collected peak obtained by MALDI-TOF after TEA treatment for Fmoc group removal. Compound identified based on: *m/z* $[RD02 + H]^{+}$ 3732.2 (calculated)/3731.0 (measured).

## 3.4 Conclusion

The synthesis and purification of RD peptides has been described in current chapter. Starting from the target sequences obtained by computational design described in Chapter 2, RD01, RD01v2 and RD02 were synthesized through Fmoc solid-phase methods along their respective native peptides. Although all target sequences could be obtained and typically corresponded to the major species observed in HPLC chromatograms, efficient collection of purified fractions was difficulted due to the high number of closely-eluting reaction by-products. Attempts were made to increase synthesis yields, either by extending reaction times for employed microwave-assisted coupling/deprotection methods or prevention of by-product build up through acetylation of unreacted chains. While this proved to be useful in the case of RD01, for RD01v2 and RD02 additional issues were encountered, namely incomplete Fmoc group removal in the last deprotection cycles. Although additional rounds of purification were required to obtain the target Fmoc-free peptides, their collection was facilitated since the Fmoc-bound form eluted away from the remaining impurities. While the low synthesis yields obtained can be attributed to limitations of solid-phase methods to obtain long polypeptide chains, the influence of introduced sequence changes was considered. Native peptides Sp1f2 and HP35 were relatively more straightforward to obtain than their corresponding designs, and in the case of RD01v2 introduction of residues with β-sheet forming propensity led to more challenging purification steps in comparison to RD01. Alternative coupling strategies were not pursued since reasonable amounts of purified peptide could be obtained. In all cases the identity of peptides (with C-terminal amidation) was confirmed by mass spectrometry analysis, and purified fractions were used for experimental characterization in the following chapters 4 to 6.

4

# 4. Physicochemical Characterization of Designed Metallopeptides

**Oral presentation in conference**

Carvalho HF, Branco RJF, Roque ACA, Iranzo O, 13th European Biologic Inorganic Chemistry Conference (2016 Budapest, Hungary).

## 4.1 Introduction

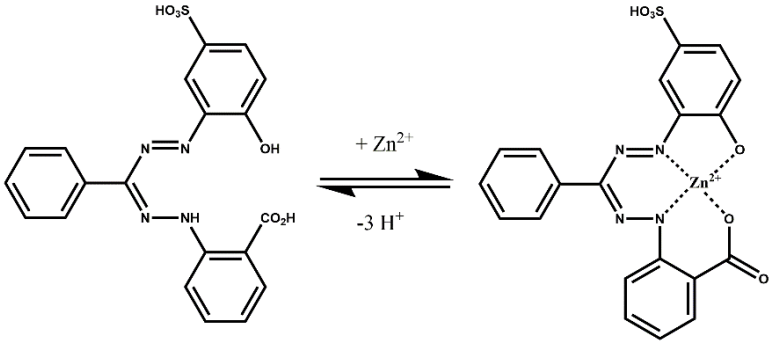
The Zn(II) ion is the most common metal present in human metalloproteins, where it can be found in two types of sites: structural and catalytic. Structural Zn(II) sites are usually buried in the protein matrix and coordinated by four protein ligands, with histidine and cysteine being the most common coordinating residues. These are often high affinity sites and play an important role in fold stabilization, such as in the representative example of the Sp1f2 ZF peptide introduced in previous chapters. Catalytic Zn(II) sites on the other hand are solvent-exposed and coordinated by three protein ligands, typically histidine, aspartate and glutamate residues. The fourth coordinating position is usually occupied by a labile water molecule, such as in representative case of MPs analysed in Chapter 2. As mentioned in Chapter 1, second sphere residues contribute to metal binding affinity by stabilization of first coordination sphere residues, and to catalytic activity by activation of the Zn(II) bound water molecule and positioning/binding of substrates.[139,140]

Metalloprotein design has been successful in generating scaffolds from a wide range of sizes and topologies that effectively bind to target metals. The $(His)_3$-Zn(II) coordination motif found in MPs from the MA(M) subclan and other native Zn(II) metalloenzymes has been in particular the focus of several designs efforts. These include small redesigned ZFs [105] and toxins [141], designed coiled coils [142,143] and helix bundles [144,145], engineered iron-containing proteins [146] and antibodies [147–149]. Zn(II) affinities in these designs can span up to 4 orders of magnitude in the micro- to nano-micromolar range and do not correlate necessarily with scaffold size, putting into evidence the important role of fine-tuned second sphere contributions and surrounding chemical environment in metal binding. Recently, Rosetta has also been used to specifically develop Zn(II)-binding sites in native proteins with micromolar affinities.[85] Nonetheless, such designs still fail to match the sub-nanomolar affinities found in native metalloenzymes such as astacin or Carbonic Anhydrase II, where protein-metal interactions have been optimized throughout the course of evolution.[150,151]

The field of protein design has been paved with known difficulties regarding structural characterization of designs, where elucidation of tertiary structure or metal site geometry is not usually reported or even possible. [126,140] This prevents the precision of employed methods to be evaluated, which is critical when computational modelling of metal sites is part of the process. In other cases, structural elucidation is possible but discrepancies are found between model and experiments.[152] Even in successful Rosetta designs that employ unnatural amino acids for Zn(II) coordination with atomic-level accuracy against crystal structures, a monodentate ligand was found to be in bidentate form and thus precluding access of a modelled water molecule to the metal site.[153] It should be noted however, that flaws in rational design approaches can nonetheless lead to unexpected functional gains. This is the case of a serendipitous discovery of a hydrolytic Zn(II)-binding dimer whose designed fourth ligand did not bound to the metal ion, effectively turning a structural site into an active catalytic one.[154] These findings stress the notion that there

are still considerable gaps in rational design methods employed in the development of artificial metalloenzymes with higher accuracy.[85]

Since Zn(II) is an integral part of the designed $MA(M)_{AS}$ model, binding of the metal is a necessary requirement for validation of the computational approach described in Chapter 2. Therefore, the RD01, RD01v2 and RD02 peptides were evaluated in terms of their affinity for Zn(II) and stability of the respective peptide-Zn(II) complexes. Given that Zn(II) is spectroscopically silent (closed $d^{10}$ shell), the apparent affinity constants of the peptides has to be determined indirectly. Co(II) is a common probe to study ZF metal binding propensities, given that is has almost isostructural coordination to Zn(II) but presents visible absorption bands arising from *d-d* transitions sensitive to coordination geometry and ligand type. However, it has a higher propensity to adopt octahedral geometries, which could give rise to distinct coordination motifs in RD peptides where the $MA(M)_{AS}$ was modelled as a tetrahedral Zn(II) centre.[155] Therefore, Zn(II) affinities were first determined indirectly by a competition assay using the colorimetric metal chelator zincon, Zi (section 4.3.1).[156–158] This approach has been used in the study of other designed protein-Zn(II) complexes with similar histidine-based coordination motifs.[143] As shown in Scheme 4.1, Zi forms a complex with Zn(II) in a 1:1 stoichiometry, with the formation of a distinct absorption band at 620 nm which can be monitored through UV-Vis absorption spectroscopy.




**Scheme 4.1 -** Zi structures in free and 1:1 Zn(II)-bound forms.

Next, folding of peptides upon Zn(II) binding was monitored in section 4.3.2 through direct measurement of backbone conformational changes by far-UV CD spectroscopy. This technique has been commonly used to monitor metal-induced folding of ZF peptides as well as *de novo* designed peptide systems.[159–162] Since the obtained affinity constants fall below the $10^6$ $M^{-1}$ range, direct titrations of the metal could be done without recurring to competitors, as it has been made for other ZF design variants.[163–166] Upon folding, peptides containing α-helices exhibit two characteristic minima at 208 and 222 nm due to $\pi \rightarrow \pi^*$ and $n \rightarrow \pi^*$ transitions of the backbone amide bonds. [167,168] Conversely, thermal unfolding of peptide-Zn(II) complexes can be monitored through spectral changes at different temperatures, as shown in section 4.3.3.[169] In the following sections, the metal binding propensities of the peptides was thus analysed and related with the stability of the corresponding Zn(II) complexes. The results obtained therein were crucial

in setting up the proper assay conditions for the catalytic studies done in Chapter 5 and to identify the structural design flaws done in Chapter 2.

## 4.2 Materials and Methods

*Competition assays:* Assays were done in 1 cm path-length quartz cells, $V_T$=900 µL in 10 mM HEPES 50 mM NaCl, pH 7.5.[21] UV-Vis spectroscopy spectra were obtained at 25 °C with temperature controller in a Cary 100 Bio spectrophotometer (integration time 0.2 s, bandwidth 2 nm, scan speed 300 nm/min), or at room temperature in a Cary 50 (integration time 0.1 s, bandwidth 1 nm, scan speed 600 nm/min) or Thermo Scientific Evolution 201 spectrophotometers(integration time 0.25 s, bandwidth 1 nm, scan speed 240 nm/min). Zincon (Zi) monosodium salt (2-carboxy-2′-hydroxy-5′-sulfoformazylbenzene) was purchased from Fluka and 1 or 2 mM stock solutions prepared by first dissolving the respective amount of solid (molecular weight 462.41 g/mol, 0.924 mg/mL for a 2 mM solution) in 5-10 µL NaOH 5M and then adding MilliQ $H_2O$ until the total volume was reached.[22] $ZnCl_2$ 10.33 or 1.033 mM solutions were prepared from a 103.3 mM stock solution (determined by inductively coupled plasma MS). Peptide stock solutions were prepared by dissolving lyophilized peptide directly in MilliQ $H_2O$ and their concentrations were determined by readings at 280 nm in 6 M guanidium chloride, considering absorbance contributions of tryptophan (ε=5690 $M^{-1}.cm^{-1}$, 1 for RD01, RD01v2 and RD02) and tyrosine (ε=1280 $M^{-1}.cm^{-1}$, 1 for RD01 and RD01v2) residues.[170,171] For Sp1f2, the concentration was determined by the Ellman's test (cysteine content determination).[172] For competition assays, peptide was incubated with $ZnCl_2$ in a 1:1 or 2:1 stoichiometry for 1 or 6h prior to titrations with Zi. In reverse competition assays, 1:1 or 2:1 Zi-Zn(II) complex was titrated with a stock solution of RD01 or RD02. During control and competition titrations $\Delta V_T$ < 3-5%, therefore no corrections to titrant concentration were done. Readings were typically done 5 minutes after titrant addition in order to allow for signal stabilization, except for assays were equilibration times were variable. Plotting and fitting of the data to the models described in Annex 3 were done in QtiPlot. The employed algorithm in fittings was the Scaled Levenberg-Marquardt with no weighting and a tolerance of $1x10^{-4}$ for 1000 iterations.

*Zinc-dependent folding and thermal stability*: Far-UV CD experiments or assays were done in 1 mm path-length cells, $V_T$= 300 µL in either 10 mM HEPES 50 mM NaCl, pH 7.5 or 10 mM TRIS 50 mM NaCl, pH 8.0. Solutions were purged with $N_2$ prior to assays and spectra were baseline corrected to subtract buffer contributions. Spectra were obtained in the "far-UV" region (203-280 nm) in a Jasco J-815 Circular Dichroism spectropolarimeter (Integration time 1 or 2 seg, band-

[21] Preliminary assays in 40 mM HEPES 50 mM NaCl, pH 7.5 were also made but results are not discussed.
[22] Zi is chemically unstable at acidic conditions [157]. Each Zi stock solution was not used for more than 3 days and stored at 4 °C to avoid degradation.

width 1 or 2 nm, 8 accumulations, scan speed 100 nm/min, data interval 0.1 or 0.5 nm). Temperature was monitored and kept constant by the use of an external controller (Jasco CDF-426S/15). Peptide concentrations (Sp1f2, RD01, RD01v2, HP35 and RD02) were 25 μM in all assays. $ZnCl_2$ 10.33 or 1.033 mM solutions were prepared from a 103.3 mM stock solution (determination by ICP). During Zn(II) titration, $\Delta V_T < 5\%$, therefore no corrections to total Zn(II) concentration were done. DiAla was purchased from POP-UP (Peptide Synthesis Facility at University of Porto, Portugal) with a purity of 92% and used without further purification. DiAla 1 mM stock solutions were prepared in $D_2O$, pH 7.5. TFE/$H_2O$ (50% v/v) peptide solutions were added by Hamilton gastight syringe. For variable temperature assays, spectra were obtained after 5 min equilibration, from 5 to 95 °C, in intervals of 10 °C and 2°C/min ramp.[23] A 1:2 peptide-Zn(II) complex was used for Sp1f2 assays and a 1:4 ratio for designed RD01, RD01v2 and RD02 to ensure most of the peptide in solutions were bound to Zn(II). Signal ($\theta_{obs}$) was converted from ellipticity to mean residual weight ($[\theta]_{MRW}$) in units of deg.dmol.cm$^2$ by the equation 4.1:

$$[\theta]_{MRW} = \frac{\theta_{obs}}{10lcm} \quad (4.1)$$

where *l* is the path-length of cell, 10 is the conversion factor from mol to dmol, *c* is the molar concentration of peptide (mol/L) and *m* is the number of peptide bonds in case of Sp1f2, RD01 and RD01v2 (m= 30) and HP35 and RD02 (m= 34). The data points in the plot figures correspond to average values from *n* replicate assays and error bars correspond to the Standard Error (S.E.) calculated as $S.E. = \frac{\sigma}{\sqrt{n}}$, where σ is the corresponding standard deviation. The quality of the fittings was evaluated based on associated $R^2$-value and $\chi^2$-distribution with *k* degrees of freedom. For the thermal unfolding model, fittings done with a fixed $\Delta C_p$=0 yielded lower associated errors.

## 4.3 Results and Discussion

### 4.3.1 Competition assays with Zn(II) chelator

The affinity of RD peptides towards Zn(II) was initially probed by competition assays with the Zi chelator. Zn(II) competes with protons for binding to cysteines and histidines, therefore the affinity of peptides has a strong pH dependency. The p$K_a$ of free histidines in solution is around 6.5, therefore assay conditions were chosen to be made at higher pH values where deprotonated forms are assumed to be the major species.[165] The *apo* form of the Zi ligand has three acidic groups with $pK_a$ values of 4, 7.85 and 15. The binding model of Zi to Zn(II) at pH 7.5 is given by equation 4.2:

$$Zi + Zn \leftrightarrow ZiZn \quad (4.2)$$

[23] For Sp1f2 (iSm2) data were obtained from temperature ramp experiments with no equilibration times.

From which the apparent Zn(II) binding constant corresponds to $K_{ZnZi,app} = \frac{[ZiZn]}{[Zn][Zi]}$. The derivation of the model for the $K_{ZnZi,app}$ determination by UV-Vis spectroscopy titration is described in Annex 3. The Zn(II) binding affinity of Zi was determined at room temperature and at different Zi concentrations ($[Zi]_T$ =10-40 μM). Results are shown in Figure 4.1 and Table 4.1.

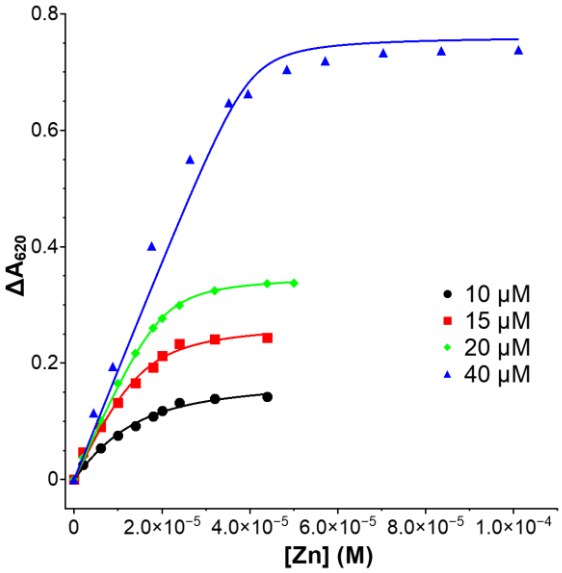



**Figure 4.1 -** Zn(II) binding affinity of Zi at pH 7.5 and at different concentrations.
Binding isotherms obtained in 10 mM HEPES 50 mM NaCl, pH 7.5 at room temperature by monitoring the changes in absorbance at 620 nm upon additions of 0-100 μM $ZnCl_2$ to 10-40 μM Zi concentrations (details in legend). Data corresponds to single assays. Solid lines represent fitted model.

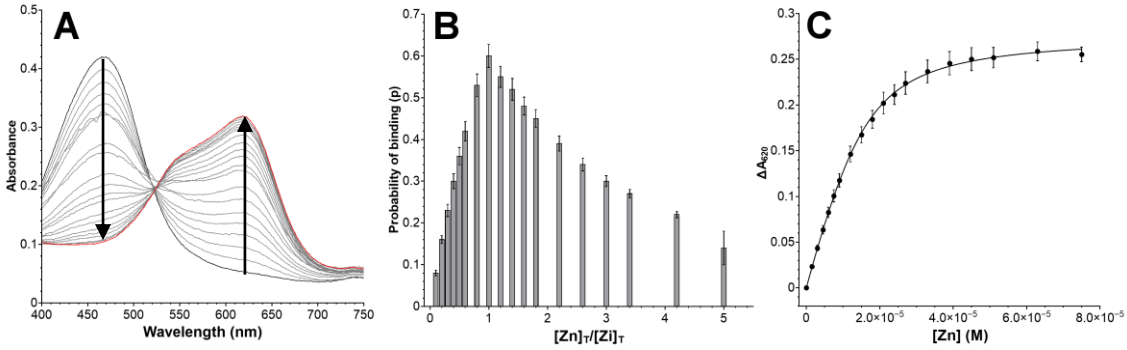



**Figure 4.2 –** Zn(II) binding affinity titrations of Zi (15 μM) at pH 7.5
A - UV-Vis spectra obtained upon additions of 0-87 μM $ZnCl_2$ in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5 with arrows showing decrease in absorbance at 464 nm corresponding to free Zi decrease (black) and absorbance increase at 620 nm corresponding to Zi-Zn(II) complex formation (red). B - Probability of binding (*p*, formula given in main text) as a function of total Zi concentration and added Zn(II). C – Corresponding absorbance difference values at 620 nm upon 0-75 μM $ZnCl_2$ additions, solid line corresponds to fitted model. Data corresponds to n=4 replicates.

Best fittings were obtained for $[Zi]_T$ < 40 μM where binding isotherms were less steep (higher $R^2$, lower $\chi^2$ (*k)* and smaller $K_{ZnZi,app}$ associated error, < 33%). Moreover, the ratio $[Zi]_T/K_{dZnZi,app}$< 100, for which reliable $K_{ZnZi,app}$ values can be obtained, is found only for $[Zi]_T$ < 40 μM.[173] A $[Zi]_T$

of 15 μM was chosen for the remaining assays, as a compromise between the lower amount of Zi used, magnitude of observed signal changes and quality of derived parameters. The results of triplicate assays at 25 °C are shown in Figure 4.2. Upon Zn(II) additions, the spectra show the decrease of the absorption band at 464 nm from the *apo* form with concomitant increase of the band at 620 nm and a isosbestic point at *c.a.* 525 nm, indicating a 1:1 complex formation. For the used $[Zi]_T$ the maximum probability of binding ($p$= [ZiZn(II)]/[Zn(II)]) is 0.6, with most points being collected within $0.2 < p < 0.8$.[174,175] Fitting of the data yields a $K_{ZnZi,app}$ of $2.7 \times 10^5$ $M^{-1}$ and a ε of 18460 $M^{-1}.cm^{-1}$.[24] These values are within the range reported in the literature under similar experimental conditions, and the $K_{ZnZi,app}$ falls within 2 orders of magnitude the range of values obtained for designed systems containing only histidines as coordinating residues (see below).[143,158,176,177]

**Table 4.1 –** Determined $K_{ZnZi,app}$, $K_{dZnZi,app}$ and ε values for Zi-Zn(II) complex in 10 mM HEPES 50 mM NaCl, pH 7.5.

| $[Zi]_T$ | $K_{ZnZi,app}$ ($M^{-1}$) | $K_{dZnZi,app}$ (M) | ε ($M^{-1}.cm^{-1}$) | $R^2$ ($\chi^2(k)$) | $[Zi]_T/K_{dZnZi,app}$ | Temp. |
|---|---|---|---|---|---|---|
| **10 μM** | $1.47 \pm 0.27 \times 10^5$ | $6.80 \pm 1.05 \times 10^{-6}$ | 17500±860 | 0.993 ($1.91 \times 10^{-5}$) | 1.47 | Room T. |
| **15 μM** | $4.01 \pm 1.31 \times 10^5$ | $2.49 \pm 0.61 \times 10^{-6}$ | 18000±887 | 0.992 ($7.01 \times 10^{-5}$) | 6.02 | Room T. |
| **15 μM (n=3)** | $2.70 \pm 0.12 \times 10^5$ | $3.70 \pm 0.16 \times 10^{-6}$ | 18460±127 | 0.999 ($5.12 \times 10^{-6}$) | 4.06 | 25 °C |
| **20 μM** | $8.93 \pm 1.02 \times 10^5$ | $1.12 \pm 0.14 \times 10^{-6}$ | 17600±168 | 0.999 ($9.17 \times 10^{-6}$) | 17.86 | Room T. |
| **40 μM** | $2.65 \pm 3.36 \times 10^6$ | $0.38 \pm 0.04 \times 10^{-6}$ | 18900±641 | 0.983 ($1.38 \times 10^{-3}$) | 106.00 | Room T. |

Note: In 40 mM HEPES 50 mM NaCl, pH 7.5, $K_{ZiZn,app} = 3.22 \times 10^5$ $M^{-1}$ (n=1)

With the knowledge of $K_{ZnZi,app}$, it is possible to calculate the respective Zn(II) binding constant of a given peptide through a competition assay, where Zi is added to a peptide-Zn(II) solution to compete with the peptide for Zn(II) binding. The equilibrium binding of Zn(II) between peptide (P) and Zi is given by equation 4.3:

$$ZnP + Zi \leftrightarrow P + ZiZn \quad (4.3)$$

Details of the derived competition model are given in Annex 3. Starting from a solution containing peptide-Zn(II) complex in a 2:1 stoichiometry to ensure most of Zn(II) is bound to peptide, all designed RD peptides were able to compete for Zn(II) and data could be fitted to the binding competition model, as seen in Figure 4.3-Figure 4.5 and Table 4.2. RD01 and RD01v2 assays

[24] Fitting to individual assays yields similar values but with higher errors (not shown).

present similar spectral features, but with an unexpected baseline increase upon Zi additions and with an apparent shift of the Zi-Zn(II) absorption band from 620 to *c.a.* 640 nm. These spectral features were not observed in the Zi assays and RD02 assays.

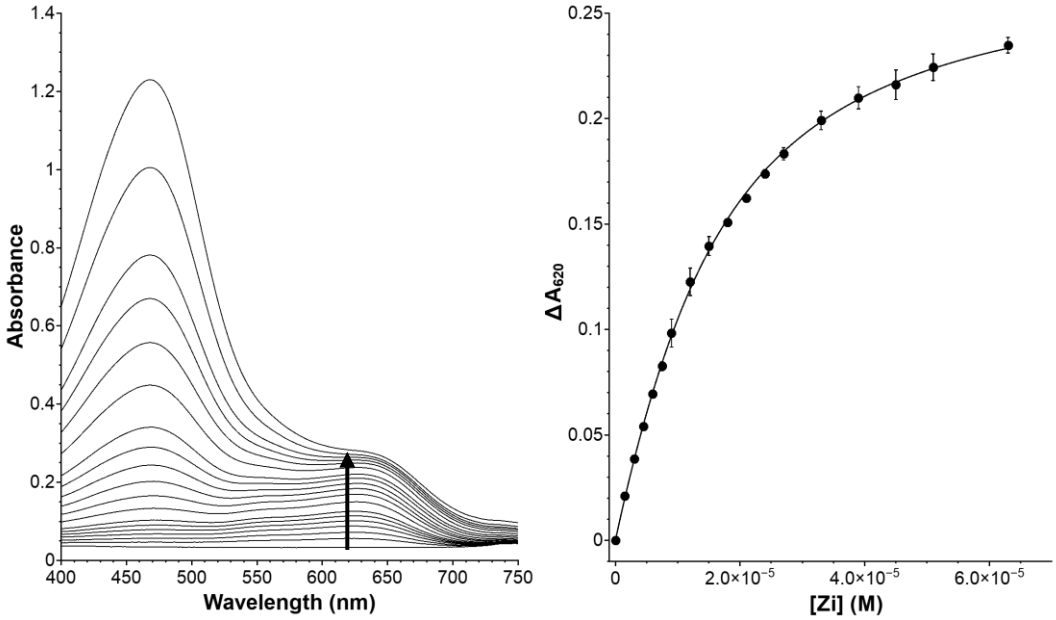



**Figure 4.3 –** Competition assay of Zi with 2:1 RD01-Zn(II) (15 μM).
Left: UV-Vis spectra of Zi titrations, with arrow showing the increase in absorbance at 620 nm corresponding to Zi-Zn(II) complex formation upon 0-63 μM Zi additions in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Right: Corresponding absorbance difference values at 620 nm for each Zi addition, solid line corresponds to fitted competition model. Data corresponds to n=2 replicates.

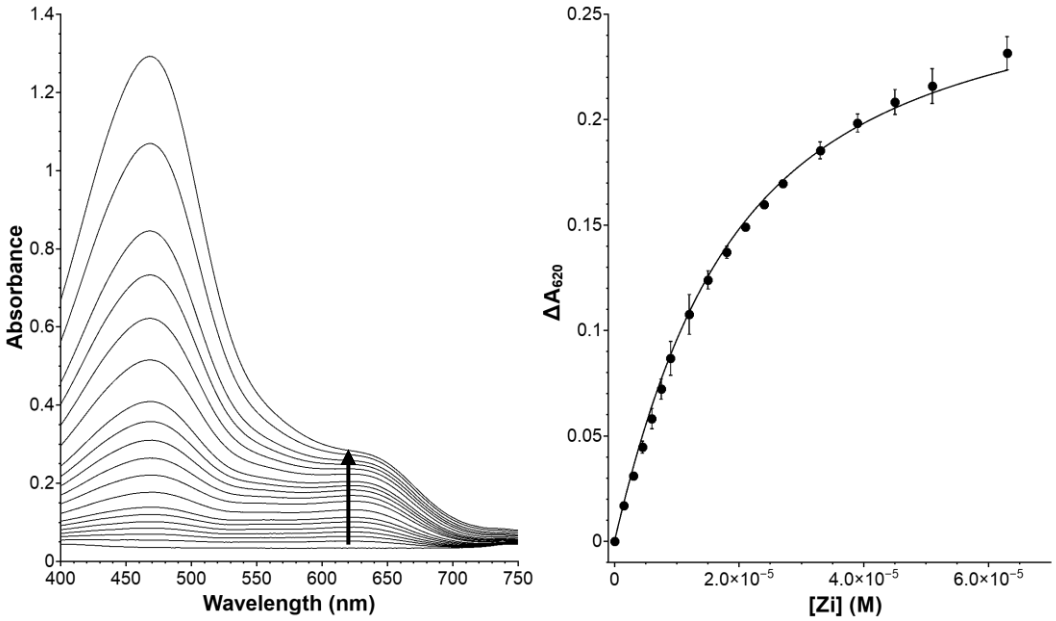



**Figure 4.4 -** Competition assay of Zi with 2:1 RD01v2-Zn(II) (15 μM).
Left: UV-Vis spectra of Zi titrations, with arrow showing the increase in absorbance at 620 nm, corresponding to Zi-Zn(II) complex formation upon 0-63 μM Zi additions in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Right: Corresponding absorbance difference values at 620 nm for each Zi addition, solid line corresponds to fitted competition model. Data corresponds to n=2 replicates.

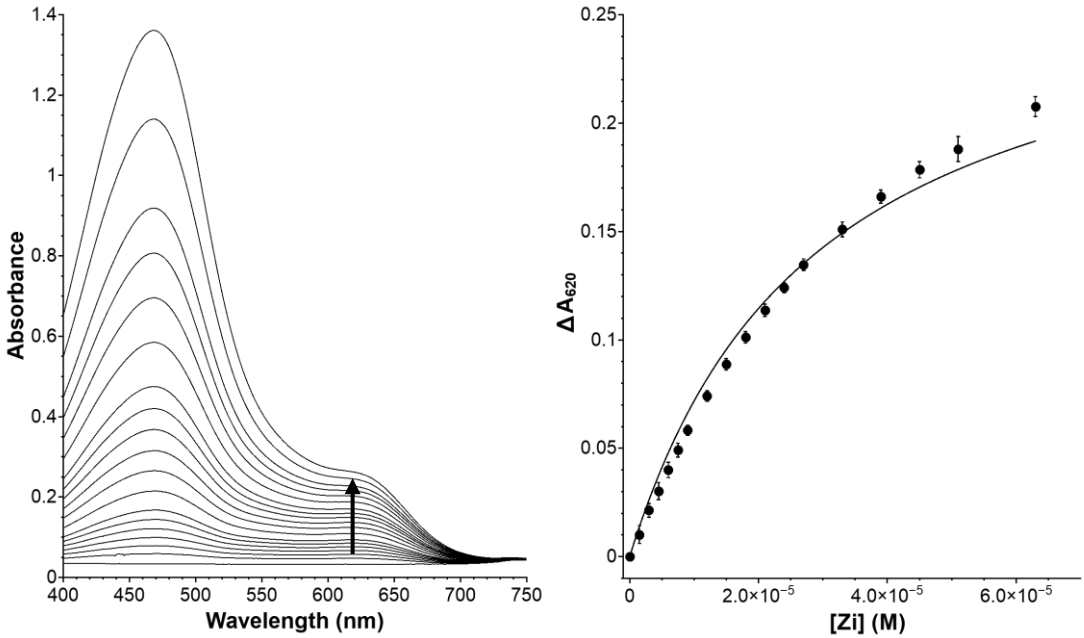



**Figure 4.5 -** Competition assay of Zi with 2:1 RD02-Zn(II) (15 μM).
Left: UV-Vis spectra of Zi titrations, with arrow showing the increase in absorbance at 620 nm, corresponding to Zi-Zn(II) complex formation upon 0-63 μM Zi additions in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Right: Corresponding absorbance difference values at 620 nm for each Zi addition, solid line corresponds to fitted competition model. Data corresponds to n=2 replicates.

**Table 4.2 -** Determined $K_{ZnP,app}$ and $K_{dZnP,app}$ values for RD peptides by UV-Vis spectroscopy in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5.

| Scaffold | $K_{ZnP,app}$ ($M^{-1}$) | $K_{dZnP,app}$ (M) | $R^2$ ($\chi^2$ ($k$)) |
|---|---|---|---|
| **RD01**[a] | $9.30\pm0.10\times10^{4}$ | $10.8\pm0.1\times10^{-6}$ | 1.000 ($2.5\times10^{-6}$) |
| **RD01v2** | $1.23\pm0.03\times10^{5}$ | $8.1\pm0.2\times10^{-6}$ | 0.998 ($1.3\times10^{-5}$) |
| **RD02** | $2.51\pm0.11\times10^{5}$ | $4.0\pm0.2\times10^{-6}$ | 0.988 ($5.3\times10^{-5}$) |

a - In 40 mM HEPES 50 mM NaCl, pH 7.5, RD01 $K_{ZnP,app} = 4.31\times10^{5}$ $M^{-1}$ (n=1)

Overall, $K_{dZnP,app}$ values fall in the micromolar range, with RD01v2 presenting a slightly increased affinity for Zn(II) in comparison with RD01. It was not possible to determine the respective binding constant of the native peptide Sp1f2 under these conditions since its affinity for zinc is > 2 orders of magnitude higher than the one of Zi.[178] RD02 presents a 2- to 3- fold higher affinity towards Zn(II) than RD01 and RD01v2. In the case of RD02, its native peptide HP35 does not appear to bind to zinc, as shown in Figure 4.6. In this case, the observed signal changes are within the error of those observed in the control titrations where Zi was added to a solution of Zn(II).

Although the data could be fitted to the binding competition model in all the cases, the unexcepted spectral features observed for RD01 and RD01v2 were further addressed with additional assays focused on RD01, where changes were apparently more pronounced. The RD01 peptide is presumed to interact in some form with the Zi-Zn(II) complex, since the shifted band at 620 nm

can be attributed only to the latter. It was therefore addressed if formation of a transient species could be the source of the observed spectral features by a competition assay where the system was equilibrated for longer periods of time, as shown in Figure 4.7.

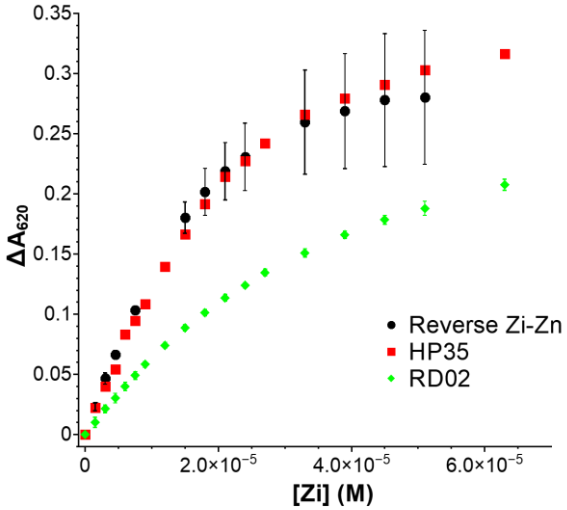



**Figure 4.6 –** Reverse titration of 15 μM $ZnCl_2$ with Zi.
Data from UV-Vis spectra of Zi-Zn(II) complex monitored at 620 nm in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Data corresponds to n=2 replicates. Comparison with single competition assays in the case of 15 μM Zn(II), 30 μM HP35 and RD02 upon 0-75 μM additions of Zi.

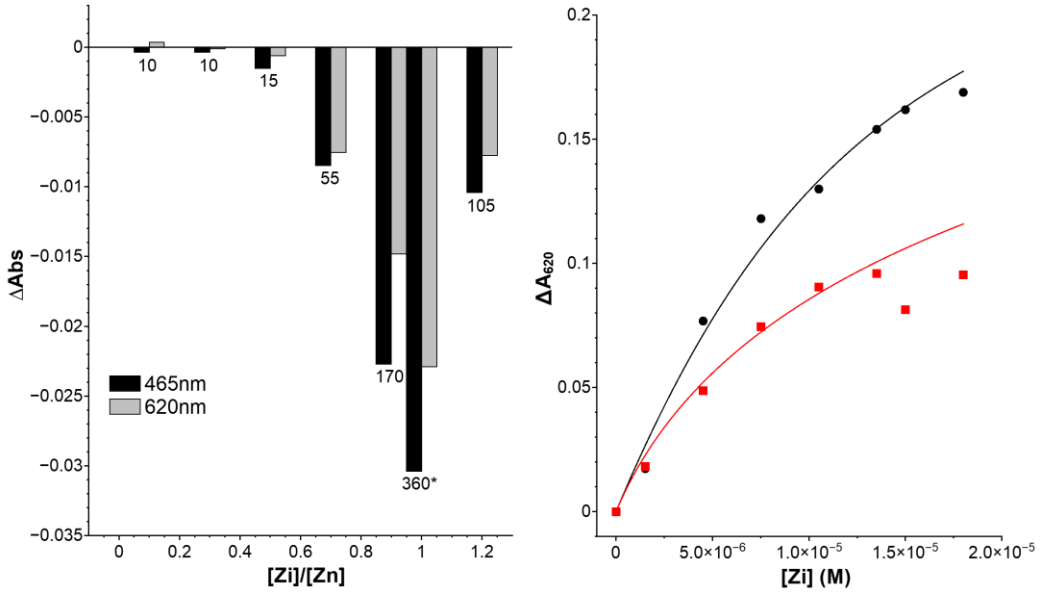



**Figure 4.7 -** Competition assay of Zi with 1:1 RD01-Zn(II) complex (15 μM) with extended equilibration times.
Left: Variations in absorbance monitored at 465 and 620 nm in 10 mM HEPES 50 mM NaCl, pH 7.5 at room temperature between 5 min after Zi additions and until signal stabilization, with time required in labels (*signal did not stabilized over indicated time). Right: corresponding absorbance values at 620 nm for 5 min (black circles) or after signal stabilization (red squares) measurements. Solid lines correspond to the fit competition models for 5 min ($K_{ZnP,app}$ 1.34±0.14x10$^5$ M$^{-1}$, $R^2$ 0.98, $\chi^2$ (k) 4.8x10$^{-5}$) and after signal stabilization ($K_{ZnP,app}$ 7.06±1.28x10$^5$ M$^{-1}$, $R^2$ 0.89, $\chi^2$ (k) 1.5x10$^{-4}$). Data corresponds to single assays.

By increasing the amount of added Zi, the time needed for signal stabilization increases until a Zi:Zn(II) ratio of 1 is reached. A full evaluation at higher ratios was not possible due to the long-time needed to complete the assay. Except for the initial point, variations in signal at 620 nm and 465 nm are negative. Considering that a negative signal variation at 620 nm corresponds to Zi-Zn(II) dissociation, the corresponding value at 465 nm should increase, reflecting higher amount of *apo* Zi. Since this inverse relation was not observed, it suggests that either the surrounding chemical environment of *apo* Zi is changed or the ligand does not become free in solution. Moreover, the data could not be properly fitted to the binding competition model when using $\Delta A_{620}$ values after equilibration time, suggesting that the proposed competition model does not apply under these conditions.[25]

Sénèque et al. reported that for zinc fingers, equilibration times are dependent on peptide sequence.[178] In the case of the consensus zinc finger CP1(CCHH), the equilibration time was around one day and this was mainly attributed to the presence of coordinating histidine residues, since for CP1(CCCC) variants it was in the order of minutes. In the case of a variant with less well-defined hydrophobic core (CP1-Δ8(CCHH)), exchange rates decreased from approximately one hour to milliseconds. Their approach involved a qualitative assessment of zinc exchange kinetics between CP1(CCHH) and EDTA. In an analogous approach to the one reported by Sénèque et al., a set of reverse competition titrations were made where RD01 was added to a solution of Zi-Zn(II) complex, with results shown in Figure 4.8. As the peptide is added, the amount of Zi-Zn(II) complex should decrease due to displacement of Zn(II), thus leading to a decrease of the 620 nm band and a concomitant increase of the 465 nm band corresponding to *apo* Zi. If the measured system is in chemical equilibrium, the data can be fitted to the competition model, thus yielding similar $K_{ZnP,app}$ in both directions. However, a decrease of both bands is observed when RD01 was added to 2:1 Zi-Zn(II) complex. Fitting of the data to the competition model was not possible, with an apparent incomplete dissociation of Zi-Zn(II) occurring at higher concentrations of added RD01.

In order to check if the excess of Zi used precluded displacement of Zn(II) by RD01, one equivalent of RD01 was added directly to 1:1 Zi-Zn(II) complex and the system was allowed to equilibrate for an extended period of time. The kinetics of Zi-Zn(II) decrease could be approximated to an exponential decay with an exchange rate $\tau$ of 125 minutes ($\tau_{1/2} \approx 83$ minutes).[26] Moreover, the previously observed baseline increase only occurred for spectra taken during equilibration. While the results indicate that RD01 can effectively compete with Zi for Zn(II) binding, the kinetics of this process appears to be slow.

[25] Batch titrations with respective equilibration times should have been performed instead.
[26] Data could be fitted to linear decay when the point at 540 minutes was not considered ($R^2$ 0.89).

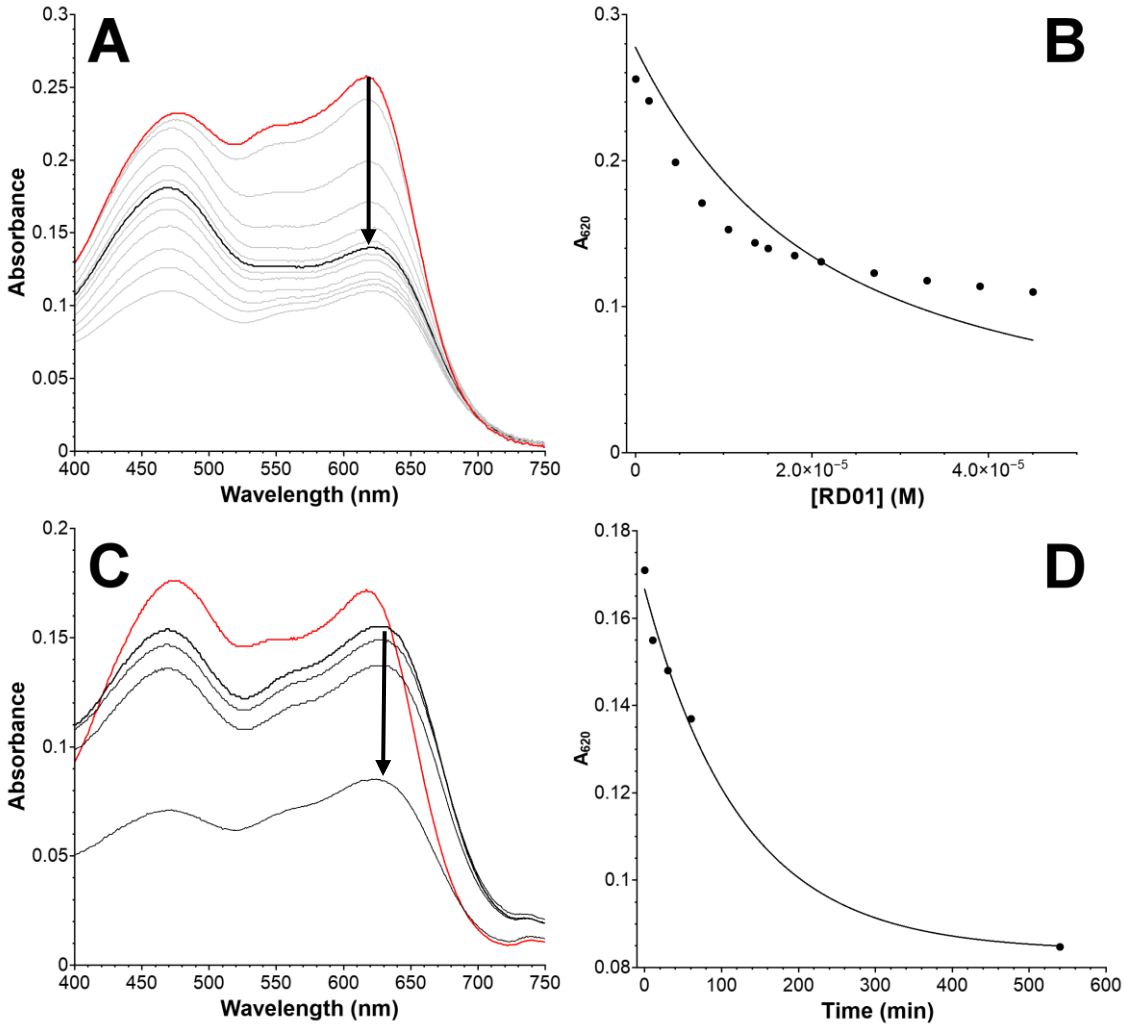



**Figure 4.8 –** Reverse competition assays of RD01.
A - UV-Vis spectra of 2:1 Zi-Zn(II) complex (red) and upon 0-45 μM additions of RD01 (15 μM, black) at room temperature, with arrow showing decrease in absorbance at 620 nm in 10 mM HEPES 50 mM NaCl, pH 7.5. B - Corresponding absorbance values at 620 nm, solid line correspond to fitted competition model ($K_{ZnP,app}$ 5.31±0.91x10$^5$ M$^{-1}$, $R^2$ 0.74, $\chi^2$ (*k*) 5.7x10$^{-4}$). C - UV-Vis spectra of 1:1 Zi-Zn(II) complex (red) and upon addition of 15 μM RD01 (black). Arrow and grey spectra show decrease in absorbance upon equilibration time. D - Absorbance decrease at 620 nm along equilibration time, solid line corresponds to fitted exponential decay ($\tau$ 125±3 min $R^2$ 0.99, $\chi^2$ (*k*) 2.57x10$^{-5}$). Data corresponds to single assays.

In order to test if the long equilibration times were due to slow binding of Zn(II) to RD01, an additional assay was made were Zn(II) was added to a mixture of *apo* Zi and RD01 in equimolar proportions. The results shown in Figure 4.9 indicate that there may be some type of interaction between Zi and the *apo* peptide, given the ≈ 15% intensity decrease in the band at 465 nm and an apparent isosbestic point at 560 nm when RD01 was added.[27] After Zn(II) addition there was the appearance of the band at 620 nm but with ≈ 35% lower intensity as when 1:1 Zi-Zn(II) complex is pre-formed, suggesting that not all Zn(II) binds to Zi.

[27] Dilution effects were considered and corresponding signal variations do not account for the observed amplitudes.

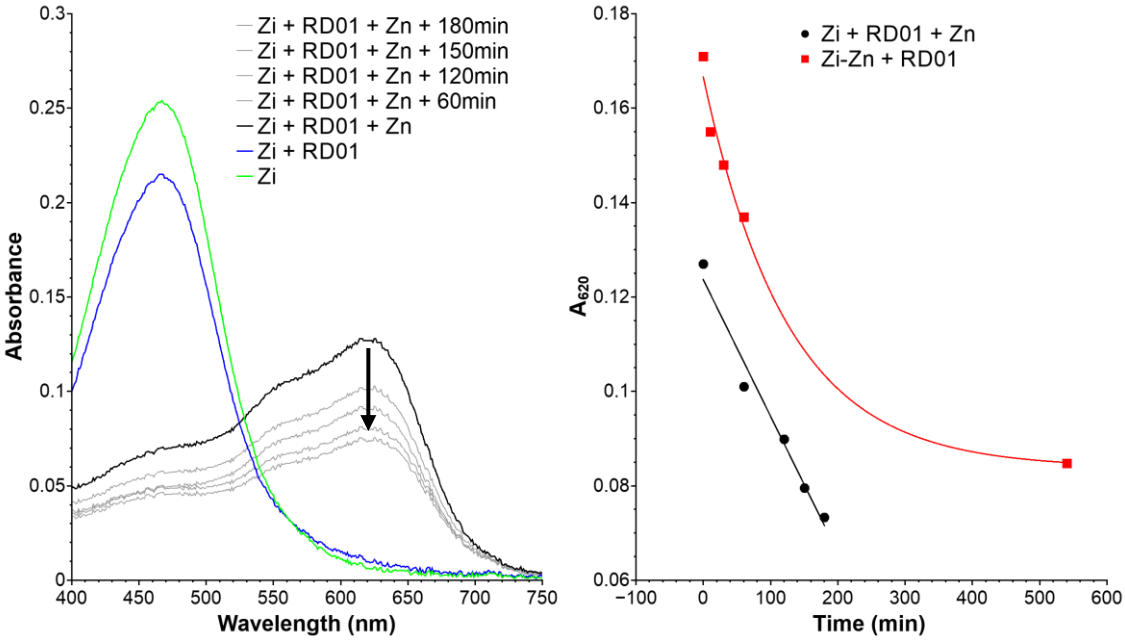



**Figure 4.9 -** Zn(II) binding to RD01 and Zi.
Left: Initial spectra of 15 μM *apo* Zi (green), upon addition of 15 μM *apo* RD01 (blue) and after addition of 15 μM $ZnCl_2$ (black) in 10 mM HEPES 50 mM NaCl, pH 7.5. at room temperature. Arrow and grey spectra show decrease in absorbance at 620 nm over time. Right: corresponding absorbance values at 620 nm (black circles). Solid black line corresponds to fitted linear decay ($\Delta A_{620}$ 0.01 $AU.s^{-1}$, $R^2$ 0.97, $\chi^2$ (*k*) $1.45x10^{-5}$), data in red squares and solid red line taken from Figure 4.8 and used here for comparison. Data corresponds to single assays.

Afterwards, both *apo* Zi and Zi-Zn(II) signals followed a linear decrease throughout the recorded 180 minutes, reaching slightly lower intensities (≈ 15%) than those obtained in the reverse competition assay after equilibration time. This suggests that the Zi-Zn(II) complex then interacts with RD01 without Zi being released.

Altogether, the results suggest that intermolecular exchange of Zn(II) does not occur, as in the proposed binding competition model, but instead an intramolecular mechanism involving the formation of a transient ternary complex of RD01-Zn(II)-Zi may be occurring.[179] Whether the formation of the such ternary complex can be related with incomplete formation of RD01-Zn(II) complex was addressed by extending the incubation time from 1 to 6 hours in a forward competition assay, with results shown in Figure 4.10.

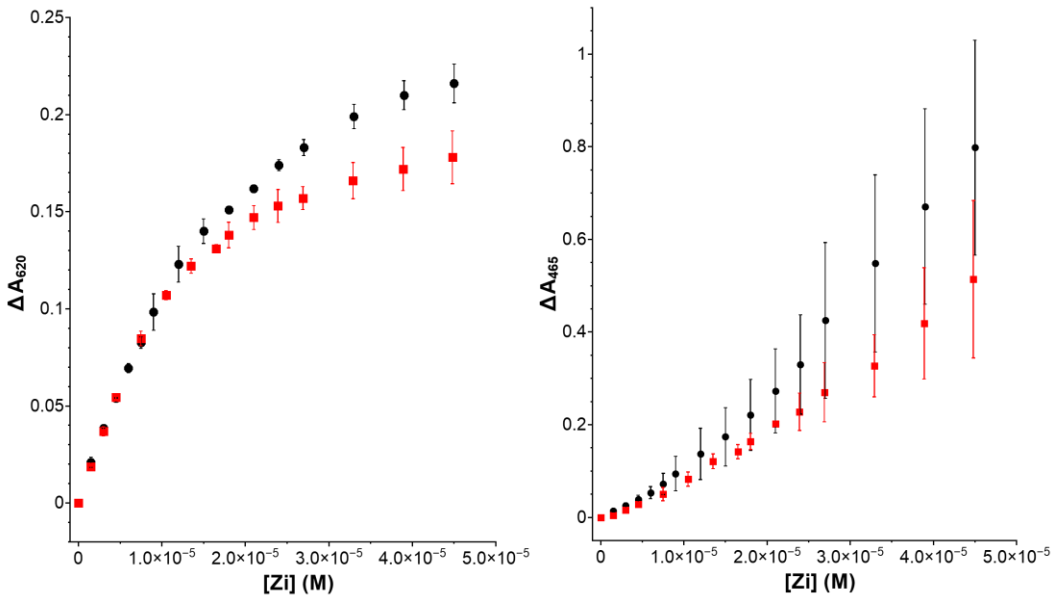



**Figure 4.10** – Effect of extended RD01-Zn(II) incubation times.
Left: data from UV-Vis spectra of Zi-Zn(II) complex formation upon additions 0-45 μM Zi to 15 μM 2:1 RD01-Zn(II), monitored at 620 nm in 10 mM HEPES 50 mM NaCl, pH 7.5 at room temperature. Right: corresponding data monitored at 465 nm. Prior to assays, RD01-Zn(II) were mixed and incubated for either 1h (black circles) or 6h (red squares). Data corresponds to n=2 replicate assays.

With extended incubation times there was less tendency for Zi-Zn(II) complex formation as Zi was added, suggesting a tighter RD01-Zn(II) binding. However, there was also less *apo* Zi, which points to higher tendency for ternary complex formation with increased incubation time. Since Zi states are the only observables in these assays, it is not possible to clarify if the observed slow kinetics are due to the formation of the RD01-Zn(II)-Zi complex.

A reverse competition assay was also made for RD02, as shown in Figure 4.11. Upon peptide additions, there was a decrease in Zi-Zn(II) concomitant with an increase in *apo* Zi and an isosbestic point at *c.a.* 510 nm. Although the corresponding data could not be fitted to the competition model, the spectral changes clearly contrast with those found for RD01, which points to a distinct mechanism of zinc exchange in the case of RD02. A more detailed and quantitative evaluation of the underlying Zn(II) exchange mechanisms would require stopped-flow techniques involving large quantities of peptide and therefore it was not addressed.[180] The kinetics of peptide-Zn(II) complex formation were further addressed in the following section by directly monitoring their formation without the need of a Zn(II) competitor.

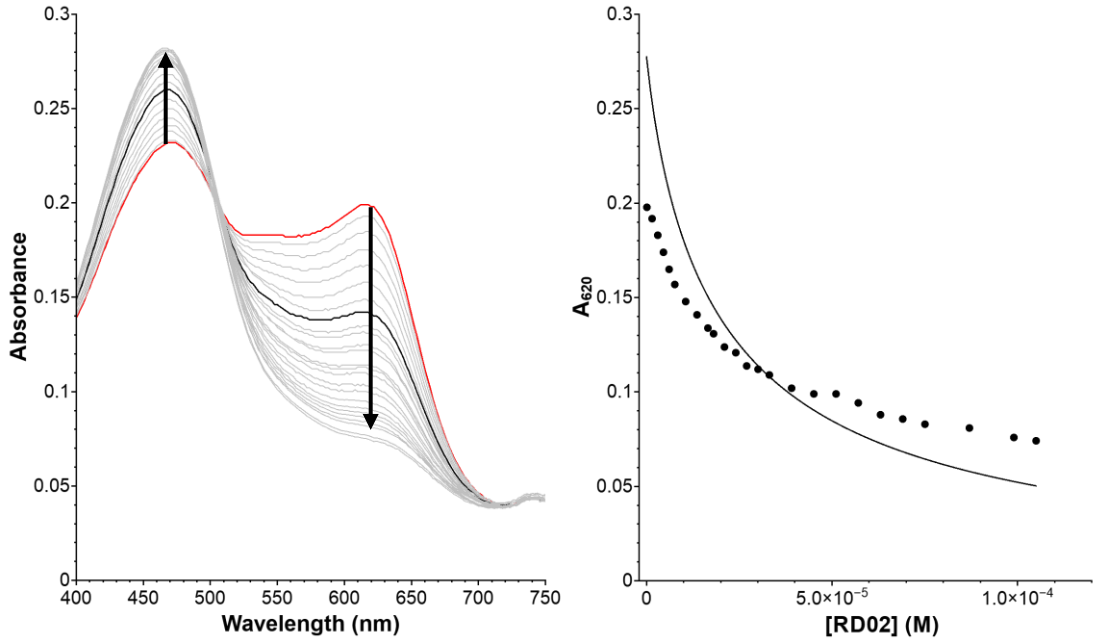



**Figure 4.11 -** Reverse competition assays of RD02 with 15 μM Zi-Zn(II).
Left: UV-Vis spectra of 1:1 Zi-Zn(II) complex (red) and upon 0-45 μM additions of RD02 (15 μM, black) in 10 mM HEPES 50 mM NaCl, pH 7.5 at room temperature, with arrow showing decrease in absorbance at 620 nm and absorbance increase at 465 nm. Right: Corresponding absorbance values at 620 nm, solid line correspond to fitted competition model ($K_{ZnP,app}$ 1.61±0.29x10$^5$ M$^{-1}$, $R^2$ 0.31, $\chi^2$ (*k*) 9.7x10$^{-4}$). Data corresponds to single assay.

### 4.3.2 Zinc-dependent folding

In the previous section, binding of RD peptides to Zn(II) was monitored indirectly through competition assays with Zi. The obtained binding constants are similar between peptides and consistent with a single metal ion binding to three histidines. Nonetheless, the $K_{ZnP,app}$ values determined in competition assays may not be reliable since for the case of RD01 (and presumably for RD01v2), there may be contributions from ternary complex formation between peptide, Zn(II) and the competitor Zi. To avoid this and gain further insight into the Zn(II) binding affinity of these designed peptides, a different approach was taken. Because the *apo* forms of the RD peptides are expected to be unfolded and to adopt a fold similar to the respective native peptides upon Zn(II) addition, the use of far-UV CD spectroscopy was considered. This spectroscopic technique can be used to directly monitor if the designed peptides fold upon Zn(II) binding by tracking changes in the content of secondary structure features, according to equation 4.4:

$$P\ (unfolded) + Zn \leftrightarrow ZnP\ (folded) \quad (4.4)$$

Using this equation, an apparent Zn(II) binding constant can be derived, $K_{ZnP,app} = \frac{[ZnP]}{[P][Zn]}$. The derivation of the model for the direct $K_{ZnP,app}$ determination is given in Annex 3. The results of the Zn(II) titrations done for all the studied peptides are shown in Figure 4.12–4.16 and summarized in Table 4.3.

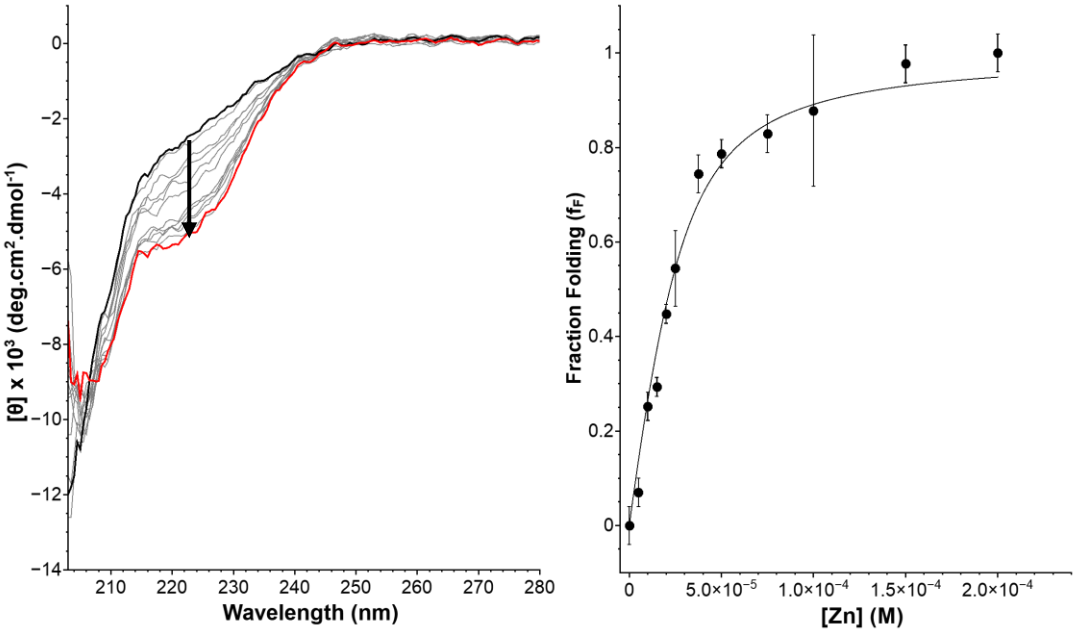



**Figure 4.12 –** Zinc-dependent folding of RD01.
Left: Far-UV CD spectra of the 25 μM RD01 titration with 0-200 μM $ZnCl_2$ in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Black line corresponds to *apo* form and red line to RD01-Zn(II) complex. Arrow show increase in negative peak at 222 nm upon Zn(II) additions. Right: Corresponding fraction of folding upon addition of Zn(II). Solid line corresponds to fitted binding model. Data corresponds to average of two or three replicates (n=2 or n=3).

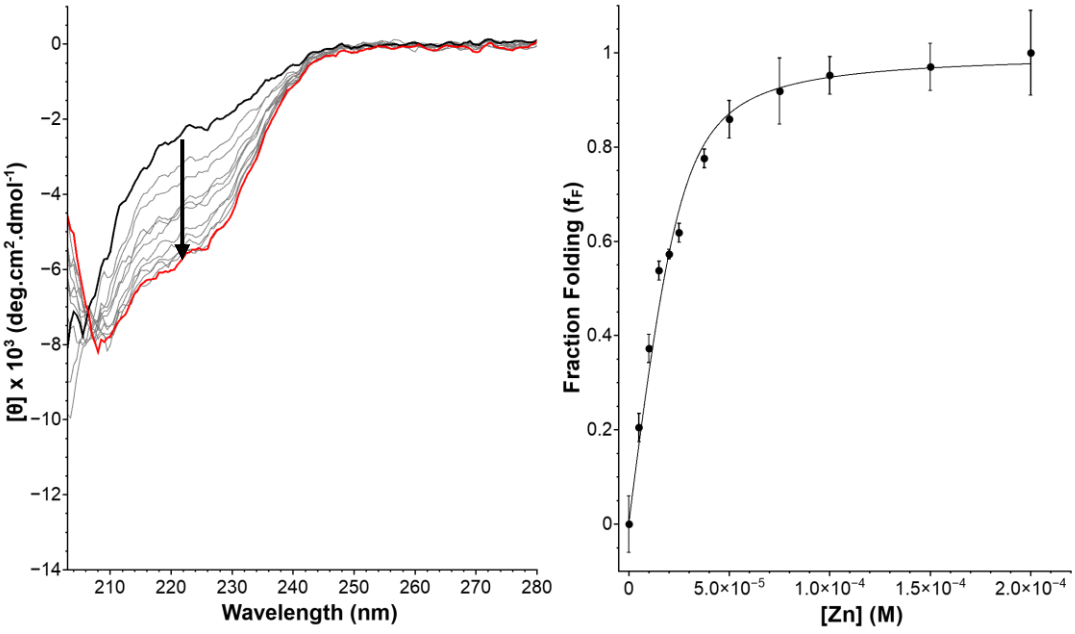



**Figure 4.13 -** Zinc-dependent folding of RD01v2.
Left: Far-UV CD spectra of the 25 μM RD01v2 titration with 0-200 μM $ZnCl_2$ in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Black line corresponds to *apo* form and red line to *holo* form. Arrow shows the increase in negative ellipticity at 222 nm upon Zn(II) additions. Right: Corresponding fraction of folding upon addition of Zn(II). Solid line corresponds to fitted binding model. Data corresponds to average of two or three replicate assays (n=2 or n=3).

Both RD01 and RD01v2 peptides in the *apo* form presented spectral signatures characteristic of random coil conformation, with a large negative ellipticity band at 204 nm, particularly in the case of RD01. The less negative ellipticity observed for RD01v2 at this wavelength could point to a more pre-organized scaffold. Upon addition of Zn(II) there was an increase of negative ellipticity at 222 nm, together with a decrease of the 204 nm band and a isodichroic point at *c.a.* 206 nm pointing to the formation of a 1:1 peptide-Zn(II) complex.[28] At the end-point of the titration the spectra were similar to the ones obtained under similar conditions for Sp1f2 variants containing only three or four histidines as coordinating residues, thus indicating the folding of RD01 and RD01v2 upon complexation with Zn(II) with $K_{ZnP,app}$ values of $1.05 \times 10^5$ $M^{-1}$ and $2.39 \times 10^5$ $M^{-1}$, respectively.[89,181] The Zn(II)-induced folding of native Sp1f2 was also addressed and compared directly with RD01 and RD01v2 at pH 8.0 (Figure 4.14).[29,30] In the *apo* form Sp1f2 also adopts a random coil conformation. In the *holo* form, Sp1f2 showed a large decrease in ellipticity at 206 nm compared to both RD01 and RD01v2, but the latter two presented higher negative ellipticities at 208 and 222 nm, suggesting increased helical content. Increased helical content has also been observed for other Sp1f2 variants lacking one coordinating cysteine residue, presumably resulting from increased conformational degrees of freedom of the backbone which allows for extension of the α-helix.[137]

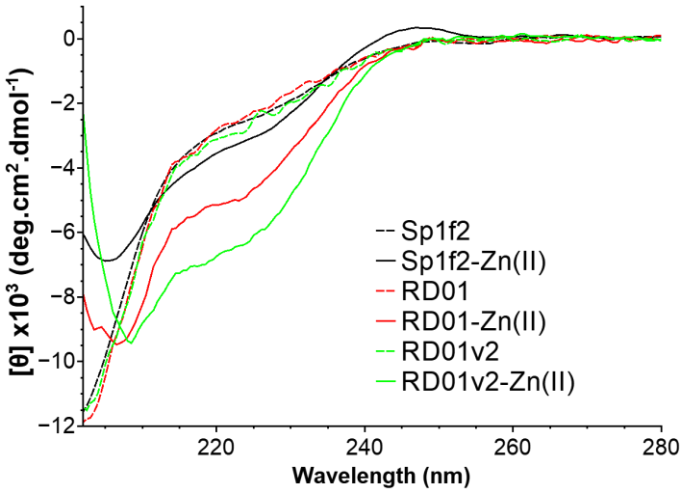



**Figure 4.14 –** Comparison of CD spectra between native Sp1f2, RD01 and RD01v2.
Far-UV CD spectra of 25 μM peptide in *apo* (dashed lines) and *holo* forms (solid lines) in 10 mM TRIS 50 mM NaCl at 25 °C, pH 8.0. Additions of $ZnCl_2$ were 37.5 μM for Sp1f2 (black), 100 μM for RD01 (red) and RD01v2 (green). Spectra of Sp1f2 correspond to average of two assays and were corrected to the average value in 250-280 nm region due to baseline increase upon Zn(II) addition. Spectra of RD01 and RD01v2 correspond to two replicates (n=2).

[28] RD01v2 presented also a more negative ellipticity at 210-216 nm in comparison to RD01, which may suggest an increased β-sheet content.

[29] Spectra were taken at a higher pH 8 in order to ensure similar protonation states of the Zn(II)-coordinating residues, since Sp1f2 contains cysteine residues which tend to have higher pKa values in zinc fingers.[178].

[30] Spectra for Sp1f2 were taken in iSm2 while for RD01 and RD01v2 taken at ITQB. Influence of using different CD spectrophotometers is small.

Regarding RD02, the *apo* form of the peptide also adopted a random coil conformation, with a large negative band at 204 nm (Figure 4.15). However, upon additions of Zn(II) this band decreased, together with an increase in negative ellipticity at 222nm and a isodichroic point at *c.a.* 210 nm. These spectral changes indicate the formation of a 1:1 peptide-Zn(II) complex with a $K_{ZnP,app}$ of 2.54x10$^5$ M$^{-1}$, similar to the values obtained for RD01 and RD01v2. The spectra after Zn(II) additions pointed to a folded α-helical structure, given the two negative bands at 222 and 208 nm with a $[\theta]_{222}/[\theta]_{208}$ of 0.91.

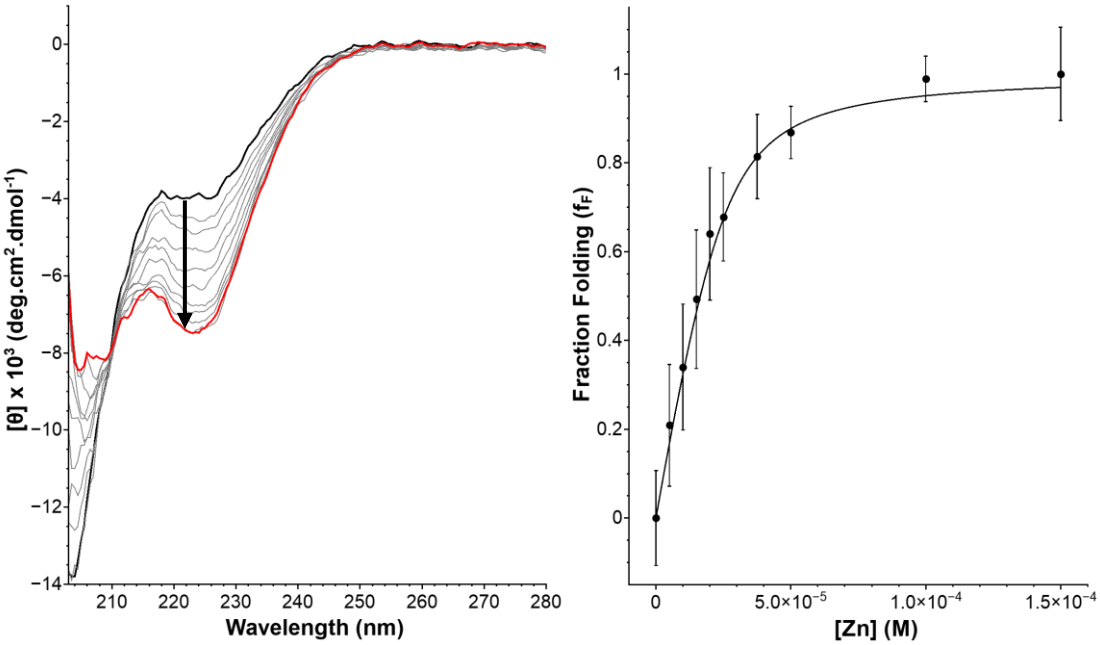



**Figure 4.15 -** Zinc-dependent folding of RD02.
Left: Far-UV CD spectra of 25 μM RD02 titration with 0-150 μM $ZnCl_2$ in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Black line corresponds to *apo* form and red line to *holo* form. Arrow show increase in negative peak at 222 nm upon Zn(II) additions. Right: Corresponding fraction of folding upon addition of Zn(II). Solid line corresponds to fitted binding model. Data corresponds to average of two replicate assays (n=2).

The comparison with the native HP35 is shown in Figure 4.16. Under the experimental conditions used, the HP35 spectra corresponds to the folded state, again with the ratio between band intensities at 222 nm and 208 nm of 0.91.[117] Upon additions of Zn(II) there was only a small general decrease in ellipticity with no observed isodichroic point. This points to a non-specific interaction between HP35 and Zn(II) given that in the competition assays with Zi there was no clear evidence of peptide-Zn(II) complex formation. The *holo* form of RD02 presents similar spectral features as HP35, although with less intense band intensities which may point to deviations from the designed fold. Nonetheless, it is shown that RD02 has a Zn(II)-induced folding in contrast to HP35 where folding is driven by hydrophobic collapse.

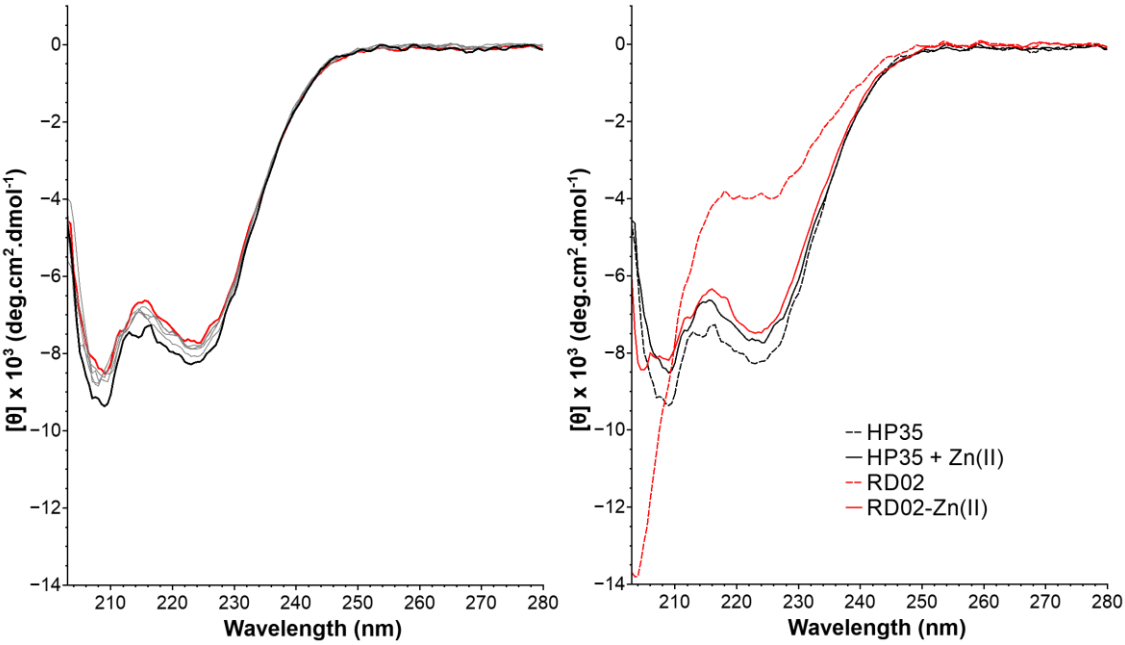



**Figure 4.16 -** Comparison of CD spectra between native HP35 and RD02.
Left: Far-UV CD spectra of 25 μM HP35 in *apo* (black line) and upon addition of 150 μM $ZnCl_2$ (red line) in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Data corresponds to two replicate assays (n=2). Right: Corresponding spectra of *apo* (dashed lines) and *holo* forms (solid lines) of HP35 (red) and RD02 (black). Spectra of RD02 taken from Figure 4.15 and used here for comparison.

The $K_{ZnP,app}$ values obtained by CD spectroscopy are in agreement with the ones obtained previously in competition assays. In the case of RD01, the binding constant is slightly higher while for RD02 the two values are in good agreement and within the associated error. Only for RD01v2 the obtained binding constant increases to approximately double the value determined by competition assay (this will be further addressed in Chapter 6).

**Table 4.3 -** Determined $K_{ZnP,app}$ values for RD peptides by far-UV CD spectroscopy.
Summary of results of zinc-dependent folding assays in 10 mM HEPES 50 mM NaCl, pH 7.5 at 25 °C. Data from competition assays (Table 4.2) included for comparison.

| Scaffold | $K_{ZnP,app}$ ($M^{-1}$) | $R^2$ ($\chi^2$ (*k*)) | $K_{ZnP,app}$ ($M^{-1}$) by competition assays |
|---|---|---|---|
| **RD01** | $1.05\pm0.12\times10^{5}$ | 0.984 ($1.9\times10^{-3}$) | $9.30\pm0.10\times10^{4}$ |
| **RD01v2** | $2.39\pm0.31\times10^{5}$ | 0.988 ($1.3\times10^{-3}$) | $1.23\pm0.03\times10^{5}$ |
| **RD02** | $2.54\pm0.20\times10^{5}$ | 0.996 ($5.4\times10^{-4}$) | $2.51\pm0.11\times10^{5}$ |

All designed peptides show binding constants in the $10^5$ $M^{-1}$ range, which is close to other designed metallopeptides with a $(His)_3$-Zn(II) coordination motif, as shown in Figure 4.17.[182] While for RD02 this represents a large increase in the affinity of the scaffold for Zn(II) in comparison with HP35, for RD01 and RD01v2 the binding constants are more than 4 orders of magnitude lower than for other zinc finger peptides, including Sp1f2 (in the $10^{8.2}$-$10^{14.7}$ $M^{-1}$ range at pH 7.0).[178] This is due to the differences in Zn(II) coordination motifs, since native zinc fingers usually contain structural sites with one or more cysteines as coordinating residues, which tend

to form more strong interactions with the metal ion. Only systems with more complex fold topologies have Zn(II) binding constants generally higher for histidine sites, which may reflect the importance of secondary-sphere interactions in stabilizing the metal centre or scaffold stability (no correlation between system size and $K_{Zn}$ was found). Indeed, given that RD peptides have reduced size and relatively simple topology (three secondary structure elements disposed majorly along a two-dimensional plane axis), there are less chances for establishing second sphere interactions at the solvent-exposed AS. Interestingly, the Zn(II) binding affinities are in the same order of magnitude reported values for ZE2, a TIM barrel fold extensively designed to accommodate the $(His)_3$-Zn(II) coordination motif using Rosetta, although optimization of second sphere interactions was attempted.[85]

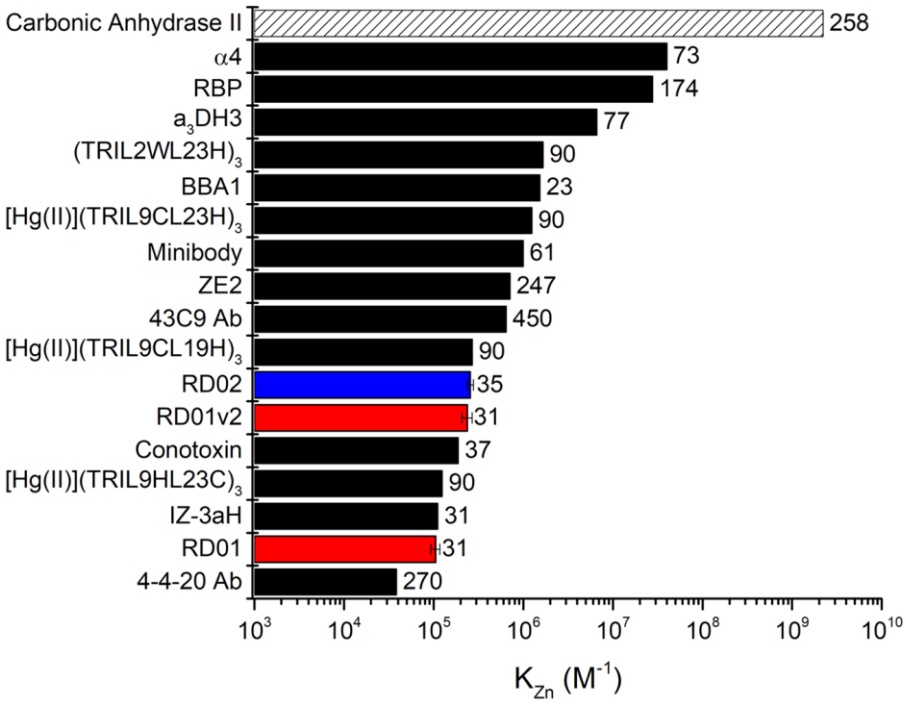



**Figure 4.17 –** Zn(II) affinity of native and designed $(His)_3$-Zn(II) proteins at pH 7.5*.
RD01, RD01v2 coloured in red, RD02 coloured in blue. Native Carbonic anhydrase II in dashed [150]. Sequence size in labels. 4-4-20 (antibody [147]); IZ-3aH (coiled coil [142]); $[Hg(II)](TRIL9HL23C)_3$, $[Hg)II)](TRIL9CL19H)_3$, $[Hg(II)](TRIL9CL23H)_3$, $(TRIL2WL23H)_3$, coiled coil [143]); Conotoxin (toxin [141]); BBA1 (ZF [105]); 43C9 (antibody [148]); ZE2 (TIM barrel [85]); Minibody (antibody [149]); α3DH3 (helix-bundle [144]); RBP (iron protein [146]), α4 (helix bundle [145]). *Except: 4-4-20, pH 6.0; IZ-3aH, pH 7.0; Conotoxin, pH 6.5; ZE2, pH 7.0.

In the case of RD01 and RD02 the $K_{ZnP,app}$ values obtained by direct Zn(II) titrations or by competition with Zi are within the associated error for each method, while for RD01v2 there was a significant increase to approximately double the value obtained previously. This increase in affinity obtained by far-UV CD spectroscopy for RD01v2 may reflect the influence of the presumed ternary complex formation in the determination of binding constants by competition assays. For

RD01v2, an additional Zn(II) titration followed by NMR spectroscopy was also done (Chapter 6) and the obtained $K_{ZnP,app}$ fell between the two values obtained in current chapter.

### 4.3.3 Stability of peptide-Zn(II) complexes

The thermodynamic stability of the designed peptide-Zn(II) complexes was evaluated by thermal unfolding assays, where the two-state model between folded and unfolded conformations described in equation 4.4 was considered:

$$ZnP\ (folded) \leftrightarrow ZnP\ (unfolded) \quad (4.4)$$

with the respective constant of folding $K = \frac{[ZnP\ (folded)]}{[ZnP\ (unfolded)]}$ varying as a function of system temperature. The derivation of the model to obtain the constant of folding (*K*), the related temperature of melting ($T_m$), and the free energy (ΔG) and enthalpy of folding ($\Delta H_{Tm}$) is given in Annex 3. The results obtained for the native and the designed peptides are shown in Figure 4.18-Figure 4.22 and the derived thermodynamic parameters summarized in Table 4.4.

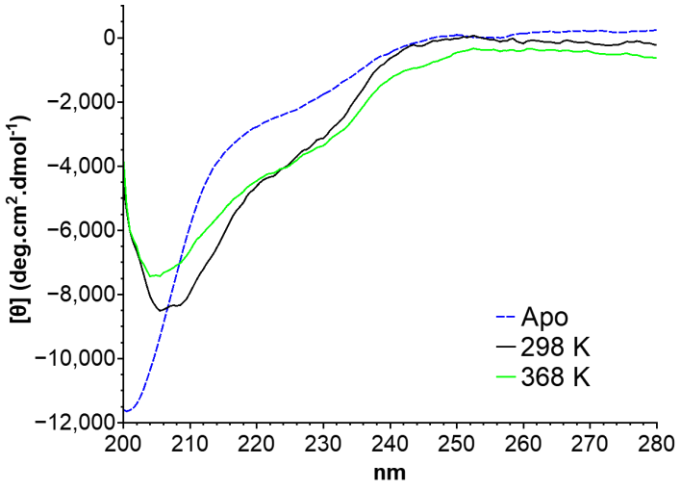



**Figure 4.18 –** Thermal unfolding of native Sp1f2.
Far-UV CD spectra of 25 μM Sp1f2 in *apo* form in 10 mM TRIS 50 mM NaCl, pH 8.0 at 25 °C (dashed blue line) and in *holo* form upon addition of 50 μM $ZnCl_2$ in 10 mM HEPES 50 mM NaCl, pH 8.0 at 25 °C (298K, solid black line) and 95 °C (368K, solid green line).

Initial assays with Sp1f2 at pH 8.0 indicated no unfolding of the peptide-Zn(II) complex upon temperature increase, with almost no signal changes occurring at 222 nm and only with a small decrease of the band at 208 nm (Figure 4.18). Indeed, the spectra of the *holo* form at both 25 °C (298K) and 95 °C (368K) is distinct from the one obtained for the *apo* form at 25 °C (unfolded peptide). Since Sp1f2 was only characterized at pH 8.0, the remaining assays for RD01 and RD01v2 were also made under these conditions.

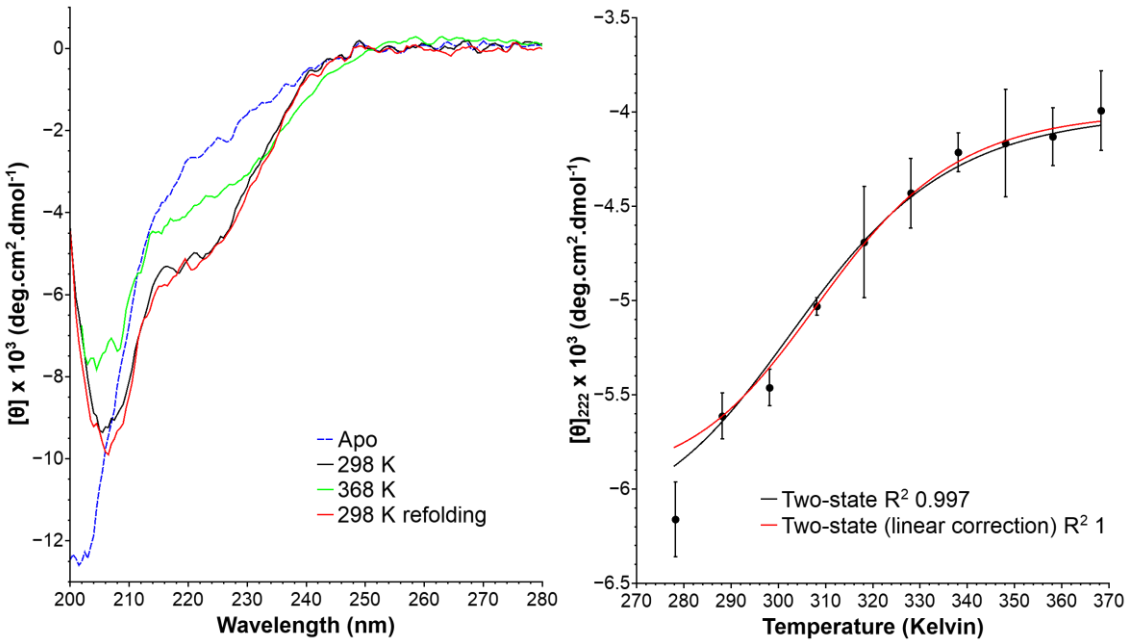



**Figure 4.19 -** Thermal unfolding of RD01.
Left: Far-UV CD spectra of 25 μM RD01 in *apo* form at 25 °C (dashed blue line), in *holo* form upon addition of 100 μM $ZnCl_2$ in 10 mM TRIS 50 mM NaCl, pH 8.0 at 25 °C (298K, solid black line), 95 °C (368K, solid green line), and after refolding at 25 °C (298K refolding, solid red line). Right: Corresponding $[\theta]_{222}$ values as a function of temperature (in Kelvin), solid lines correspond to the fit to the two-state transition models, without (black) and with linear corrections for the pre- or post-transition phase (red). Data corresponds to two replicates (n=2).

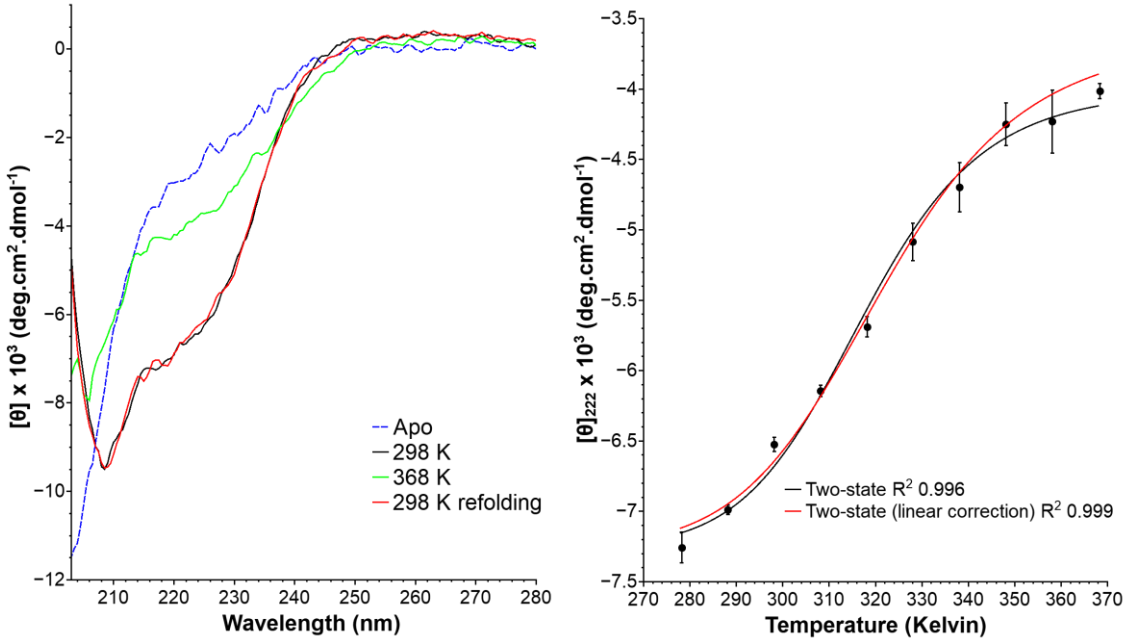



**Figure 4.20 -** Thermal unfolding of RD01v2.
Left: Far-UV CD spectra of 25 μM RD01v2 in *apo* form at 25 °C (dashed blue line), in *holo* form upon addition of 100 μM $ZnCl_2$ in 10 mM TRIS 50 mM NaCl, pH 8.0 at 25 °C (298K, solid black line), 95 °C (368K, solid green line), and after refolding at 25 °C (298K refolding, solid red line). Right: Corresponding $[\theta]_{222}$ values as a function of temperature (in Kelvin), solid lines correspond to the fit to the two-state transition models without (black) and with linear corrections for the pre- or post-transition phase (red). Data corresponds to two replicates (n=2).

In the case of RD01 (Figure 4.19) and RD01v2 (Figure 4.20), there were clear spectral changes occurring as a function of temperature indicating unfolding of the peptides. At the end-point temperature of 95 °C (368K) the two peptides presented similar spectra, although in both cases the *holo* form did not resemble the spectra of the *apo* form. Both RD01 and RD01v2 showed reversible folding, *i.e.* identical spectra at 25 °C before and after temperature being varied up to 95 °C. Although RD01 appears to present a less well defined unfolding curve than RD01v2, the two-state model described above could describe well the experimental data of both peptides. The obtained thermodynamic parameters indicate that RD01 is less thermodynamically stable than RD01v2, since the former has lower $T_m$, although both present similar $\Delta H_{Tm}$. Indeed, the corresponding ΔG of folding of RD01 is almost half the one obtained for RD01v2, suggesting a significant increase in stability between the first and second design rounds. Nonetheless, given that under the same experimental conditions Sp1f2 did not show unfolding, the stability of the two designed peptides is considerably lower than for the native sequence.

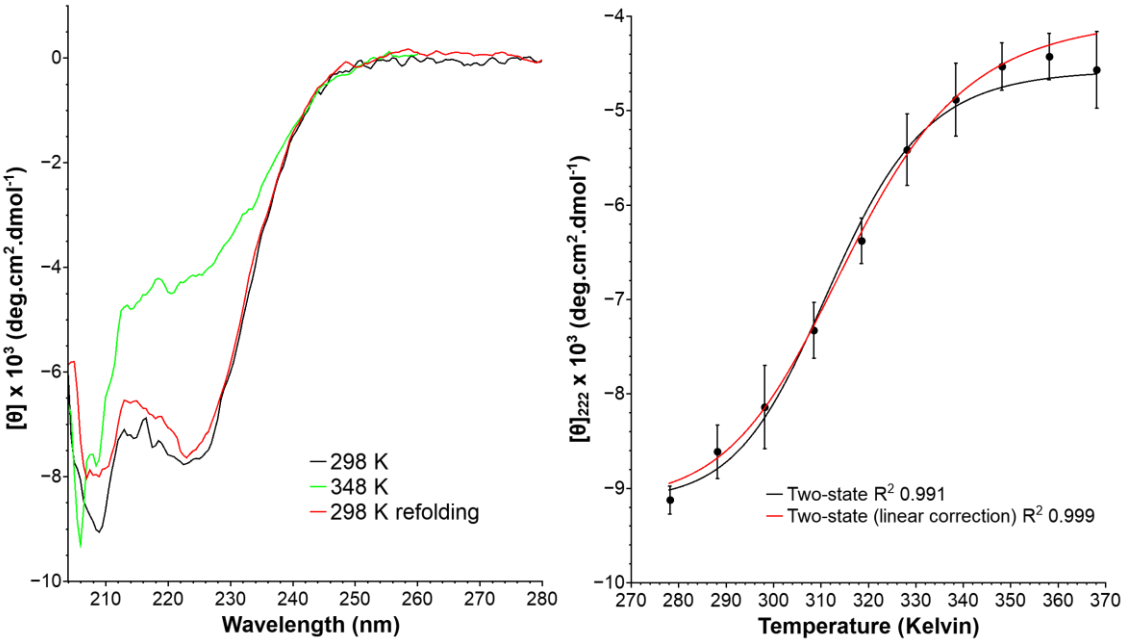



**Figure 4.21 -** Thermal unfolding of native HP35.
Left: Far-UV CD spectra of 25 μM HP35 in 10 mM HEPES 50 mM NaCl, pH 7.5 at 25 °C (298K, black line), 75 °C (348K, green line), and after refolding at 25 °C (298K refolding, red line). Right: Corresponding $[\theta]_{222}$ values as a function of temperature (in Kelvin), solid lines correspond to the fit to the two-state transition models without (black) and with linear corrections for the pre- or post-transition phase (red). Data corresponds to two replicates (n=2).

In contrast to Sp1f2, the native peptide HP35 shows unfolding with increased temperature (Figure 4.21). There are clear differences in the spectra taken at 25 and 75 °C, and the unfolding is reversible. The data could be fitted to the simple two-state model. The corresponding thermodynamic parameters are lower than those reported in the literature. [117] This may be due to the differences in the experimental conditions used. Most studies were done at lower pH values where villin tends to be more stable, while here for comparison purposes a pH of 7.5 was used.

RD02 also shows unfolding at this pH as a function of temperature (Figure 4.22). The spectra of the *holo* RD02 form at 25 and 75 °C are similar to those obtained for HP35 and distinct from the one obtained in the *apo* RD02 form at 25 °C. Data from RD02 could also be fitted to the two-state model, with the derived $T_m$ being higher than the one obtained for HP35 but with similar $\Delta H_{Tm}$. The resulting ΔG of folding for HP35 is within associated error to the one obtained for RD02, although the folding mechanism is distinct between peptides (hydrophobic collapse *vs* metal induced).

There is a correspondence between the derived thermodynamic parameters and the affinity for Zn(II) of the designed peptides: RD01 has lower $K_{ZnP,app}$ and $T_m$ values than RD01v2 and RD02. The difference between RD01v2 and RD02 resides in the $\Delta H_{Tm}$ value, which is higher in the case of RD01v2. This points to a relation between the propensity of the peptides to bind to Zn(II) and adopt a folded conformation and the stability of the resulting Zn(II) complexes. However, it has been recently argued that the cost of folding in ZFs is low compared with the free energy associated with Zn(II) binding.[183] Indeed, the *holo* peptides at 95 °C show distinct spectral features from the *apo* forms at 25 °C, suggesting that once bound, the metal ion is not released upon fold denaturation. A more careful investigation of the thermodynamic properties of RD-Zn(II) complexes would be required to shed light on these issues.[165,184]

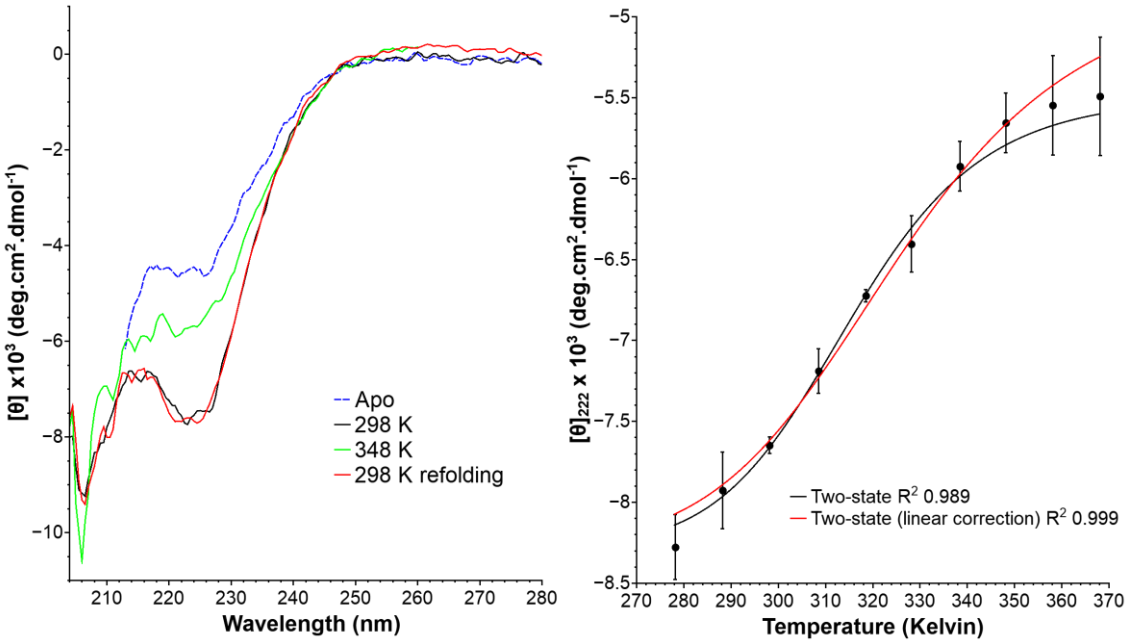



**Figure 4.22 -** Thermal unfolding of RD02.
Left: Far-UV CD spectra of 25 μM RD02 in *apo* form at 25 °C (dashed grey line), in *holo* form upon addition of 100 μM $ZnCl_2$ in 10 mM HEPES 50 mM NaCl, pH 7.5 at 25 °C (298K, solid black line), 95 °C (368K, solid grey line), and after refolding at 25 °C (298K refolding, solid red line). Right: Corresponding $[\theta]_{222}$ values as a function of temperature (in Kelvin), solid lines correspond to the fit to the two-state transition models without (black) and with linear corrections for the pre- or post-transition phase (red). Data corresponds to two replicates (n=2).

**Table 4.4 -** Enthalpies ($\Delta H_{Tm}$), free energies ($\Delta G$) of folding and temperature of melting ($T_m$) determined by far-UV CD variable temperature assays.

| Scaffold | $T_m$ (°C) | $\Delta H_{Tm}$ (kcal/mol) | **ΔG at 25 °C** (kcal/mol) |
|---|---|---|---|
| **Sp1f2**[a, b] | - | - | - |
| **RD01**[a] | 37.6±1.3 | -13.5±0.7 | -0.54±0.03 |
| **RD01v2**[a] | 46.9±2.4 | -13.3±1.3 | -0.90±0.10 |
| **HP35**[c] | 41.5±1.7 | -15.6±1.6 | -0.81±0.08 |
| **RD02**[c] | 49.6±4.5 | -10.8±1.4 | -0.82±0.10 |

a - 10 mM TRIS 50 mM NaCl, pH 8.0.
b – no unfolding at 222 nm observed.
c - 10 mM HEPES 50 mM NaCl, pH 7.5.

The effect of longer equilibration times on peptide-Zn(II) complex formation was also addressed by tracking conformational changes over time, as shown in Figure 4.23. No significant spectral changes were observed after 1 hour of the addition of one equivalent of Zn(II) for none of the peptides, suggesting that the peptide-Zn(II) complexes are readily formed upon mixing. This further suggests that in the case of RD01, the long equilibration times observed in the competition assays with Zi were not related with slow kinetics of formation of the peptide-Zn(II) complex but rather with presumable slow exchange phenomena between the peptide and ternary species.

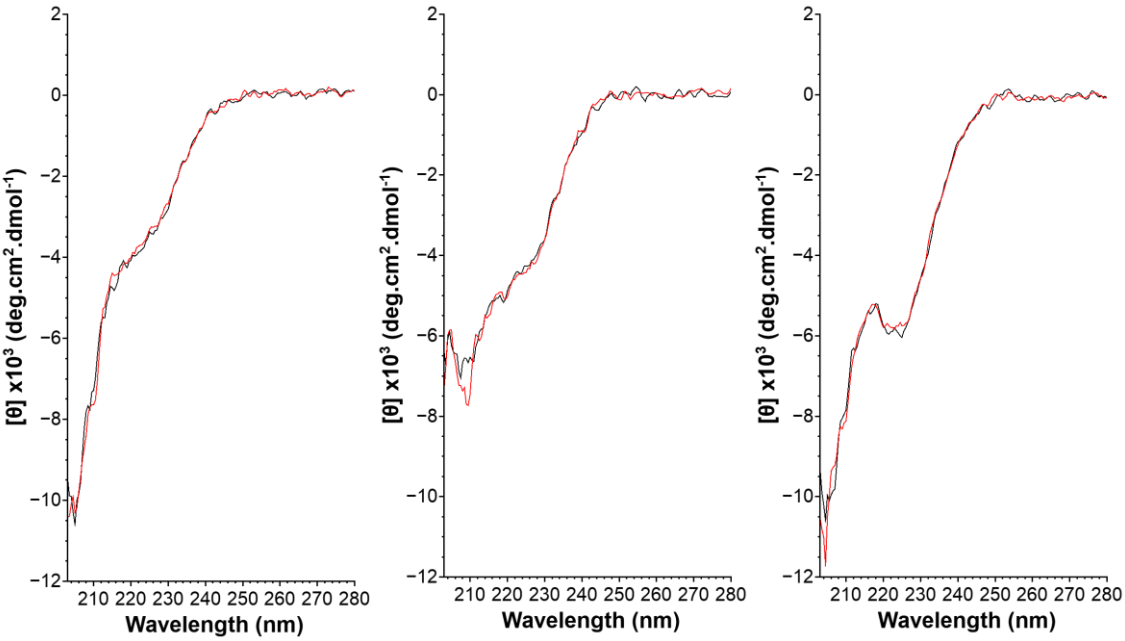



**Figure 4.23 –** Effect of extended equilibration time in the peptide-Zn(II) complex formation. Far-UV CD spectra of 25 μM RD01 (left), RD01v2 (centre) and RD02 (right) obtained upon addition of 25 μM of $ZnCl_2$ (black) and after 1h of incubation (red) in 10 mM HEPES 50 mM NaCl, pH 7.5. Spectra correspond to single assays.

Peptide-Zn(II) complex stability was also addressed in the presence of organic solvents, namely acetonitrile and 2,2,2-triflouroethanol (TFE). Acetonitrile was used as a co-solvent for

some substrates in catalytic assays, and therefore its effect on the folding of the peptide-Zn(II) complexes was analysed. As shown in Figure 4.24, there were no significant changes in the peptide-Zn(II) complex conformation after addition of 5% acetonitrile (amount used in the catalytic studies reported in Chapter 5) with the exception of a small decrease in the 204 nm band for RD01.[31]

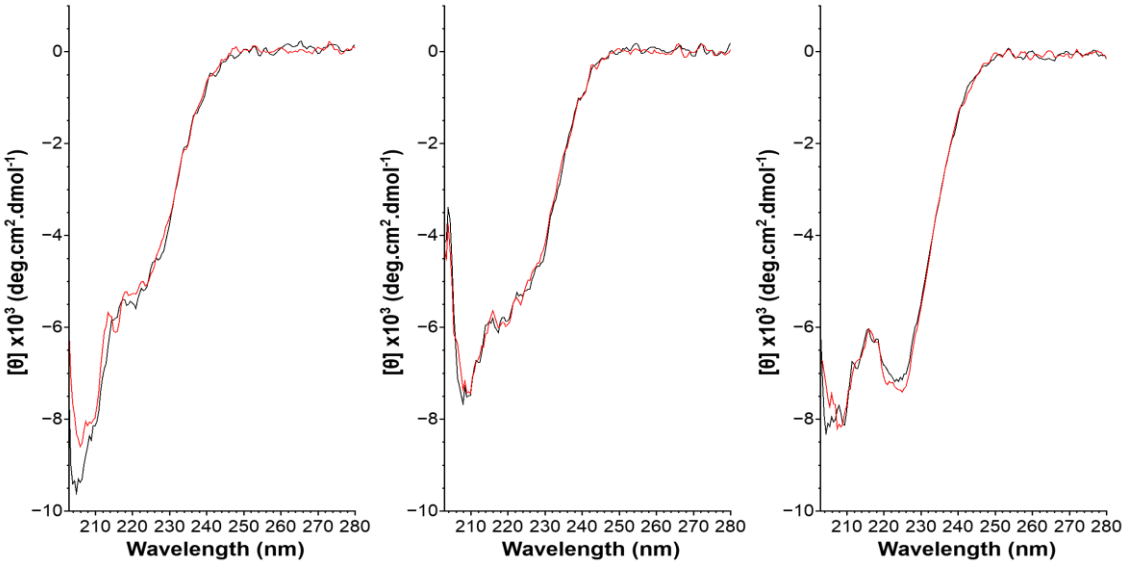



**Figure 4.24 –** Effect of Acetonitrile in peptide-Zn(II) complex stability.
Far-UV CD spectra of 25 μM 1:8 RD01-Zn(II) (left), 1:8 RD01v2-Zn(II) (centre) and 1:6 RD02-Zn(II) (right) obtained before (black) and after (red) addition of 5% (v/v) acetonitrile in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Spectra correspond to single assays.

TFE is known to be an inducer of secondary structure features in peptides.[185,186] Indeed, as shown in Figure 4.25, upon addition of TFE clear spectral changes were observed for all designed peptides. Both RD01 and RD01v2 presented increased negative peaks, with lower amplitude of change in the case of the later. RD02 presented the highest amplitude of ellipticity change of the peptide-Zn(II) complexes, particularly at 222 nm where it approximated an isotherm around 42% TFE. Moreover, with increased volume percentage of TFE the $[\theta]_{222}/[\theta]_{208}$ ratio changed slightly for RD02, while this was not observed in the case of native HP35, suggesting some degree of structural re-organization in the former.

The results suggest that the presence of TFE leads to more structured peptide-Zn(II) complexes, particularly in the case of RD01 and RD02 where the effect is more noticeable. This points to low structural stability of these designs, in line with the small free-energies of folding found previously. As for native HP35, the results indicate again that the peptide has a less stable fold under these conditions. Indeed, tertiary structure elucidation of this peptide in the literature was only possible at a pH values lower than 7.5

[31] Data for HP35 was also obtained but not shown. No significant differences were detected.

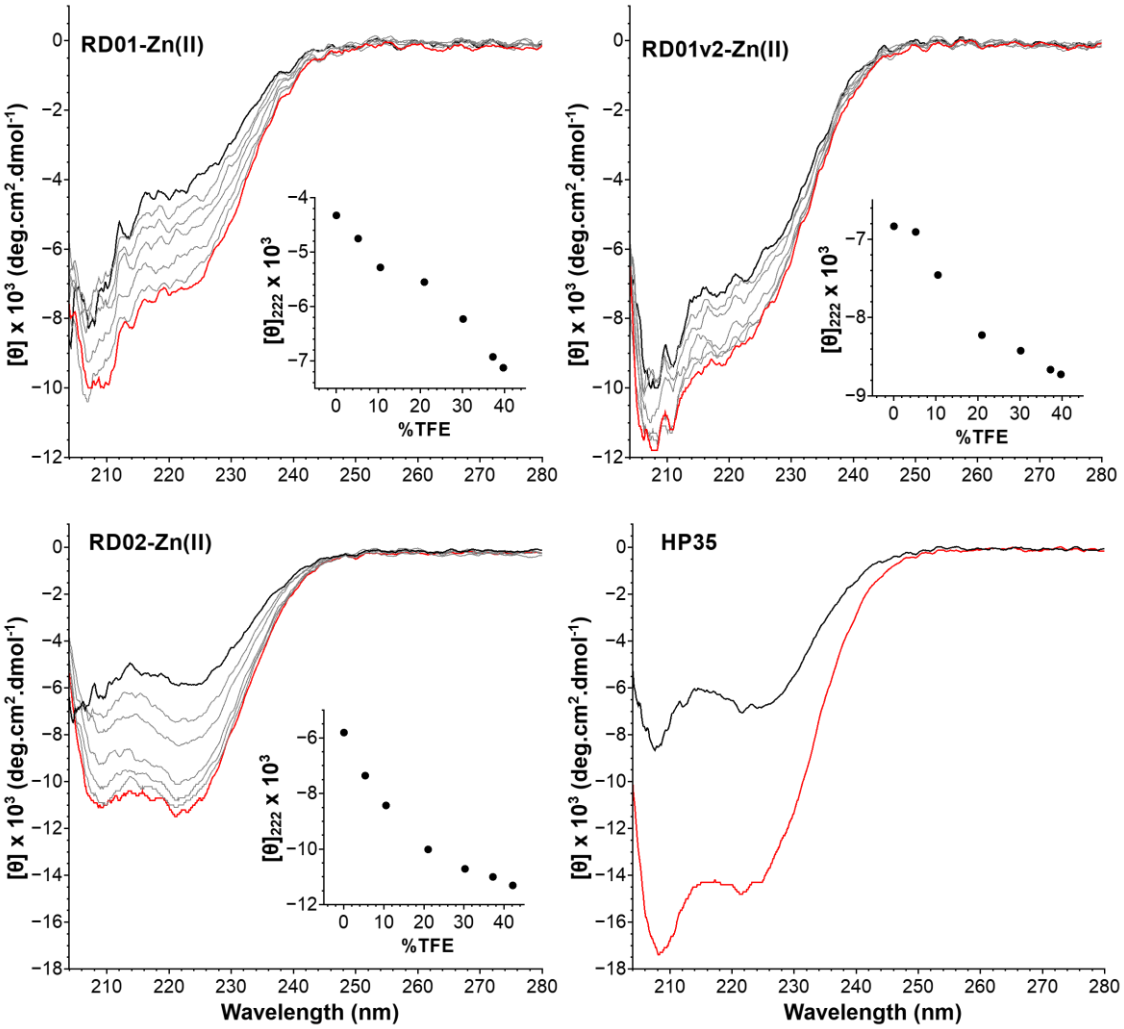



**Figure 4.25 -** Effect of TFE in peptide secondary structure.
Far-UV CD spectra of 25 μM of 1:4 RD01-Zn(II) (top left), 1:4 RD01v2-Zn(II) (top right), 1:4 RD02-Zn(II) (bottom left) and HP35 (bottom right) in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Spectra were obtained before (black) and under the presence of 39.7% TFE for RD01-Zn(II), RD01v2-Zn(II) and 42.2% TFE for RD02-Zn(II), HP35 (red). Inset plots represent the $[\theta]_{222}$ values *vs.* percentage of TFE (grey lines), except for HP35 where da single addition of TFE was made. Data correspond to single assays.

The RD peptides were designed in Chapter 2 under the presence of the model substrate diAla and specific sequence changes were made to accommodate it at the AS. Given the reduced size of the scaffolds, interactions with the substrate could play a significant role in scaffold integrity. This is because in the absence of diAla, the introduced residues could adopt unfavourable interactions with other residues that would render the designs unstable.[32] CD spectroscopy was therefore used to probe if interactions between the peptides and the model substrate diAla could lead to significant secondary structure changes, as observed previously for TFE assays. The results shown in Figure 4.26 indicate that up to a 1:4 excess of substrate there was no consistent conformational changes for none of the designed peptides. Due to limited substrate solubility in water,

[32] As described in Chapter 2, evaluation of designs under the absence of diAla was not considered since the Zn(II) metal ion was treated as part of the substrate.

higher substrate to peptide-Zn(II) ratios could not be tested. Nonetheless, considering that in these assays the peptide-Zn(II) concentration was 25 µM and a total diAla substrate concentration of 100 µM was used, interactions with sub-milimolar constants leading to fold changes appear to be absent.

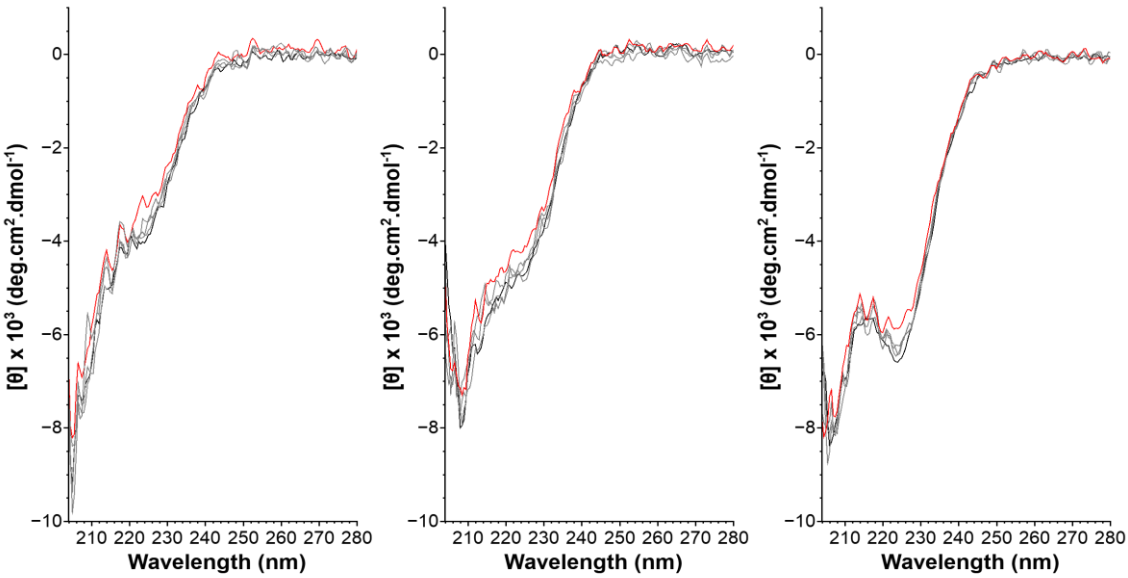



**Figure 4.26 –** Interaction of diAla with the peptide-Zn(II) complexes.
Far-UV CD spectra of 25 µM of 1:4 RD01-Zn(II) (left), 1:4 RD01v2-Zn(II) (centre), 1:4 RD02-Zn(II) (right) before (black) and after the addition of 0-100 µM diAla (red) in 10 mM HEPES 50 mM NaCl at 25 °C, pH 7.5. Data correspond to single assays.

## 4.4 Conclusion

In the current chapter, the characterization of the physicochemical and structural properties of the RD peptides were described. After synthesis and purification in Chapter 3, the designed peptides were shown to bind to Zn(II) with micromolar affinities through competition assays with the chelator Zincon monitored by UV-Vis spectroscopy. The Zn(II) exchange process between peptide and chelator was shown to be distinct for RD01 and RD02, pointing to the formation of transient ternary species in case of the former. Backbone conformational changes induced under the presence of Zn(II) were observed by far-UV CD spectroscopy for all RD peptides and were consistent with micromolar Zn(II) affinities determined in the competition assays. The Zn(II)-induced folding and binding constants obtained point to coordination of the metal ion by the three histidine residues, in agreement with the designed active site in Chapter 2. Although within a narrow range, RD01 presented lower Zn(II) affinity values than RD01v2 and RD02. Similar observations were found in terms of thermal stability of the respective Zn(II) complexes. The iterative approach of design and experiment resulted in improvements of scaffold physicochemical properties. The results regarding Zn(II) binding properties were used to properly set the experimental conditions

used in catalytic assays described in Chapter 5, where both *apo* and *holo* forms of RD peptides were characterized separately. RD-Zn(II) complexes adopted secondary structure features similar to their native counterparts, despite some relevant differences. RD01 and RD01v2 presented increased helical content in relation to Sp1f2 ββα fold. On the other hand, RD02 presented decreased helical content in comparison to HP35 all-α fold. Since drift from native-like fold topologies was accompanied by marginal thermal stability of all RD-Zn(II) complexes, a more detailed structural analysis was issued in Chapter 6.

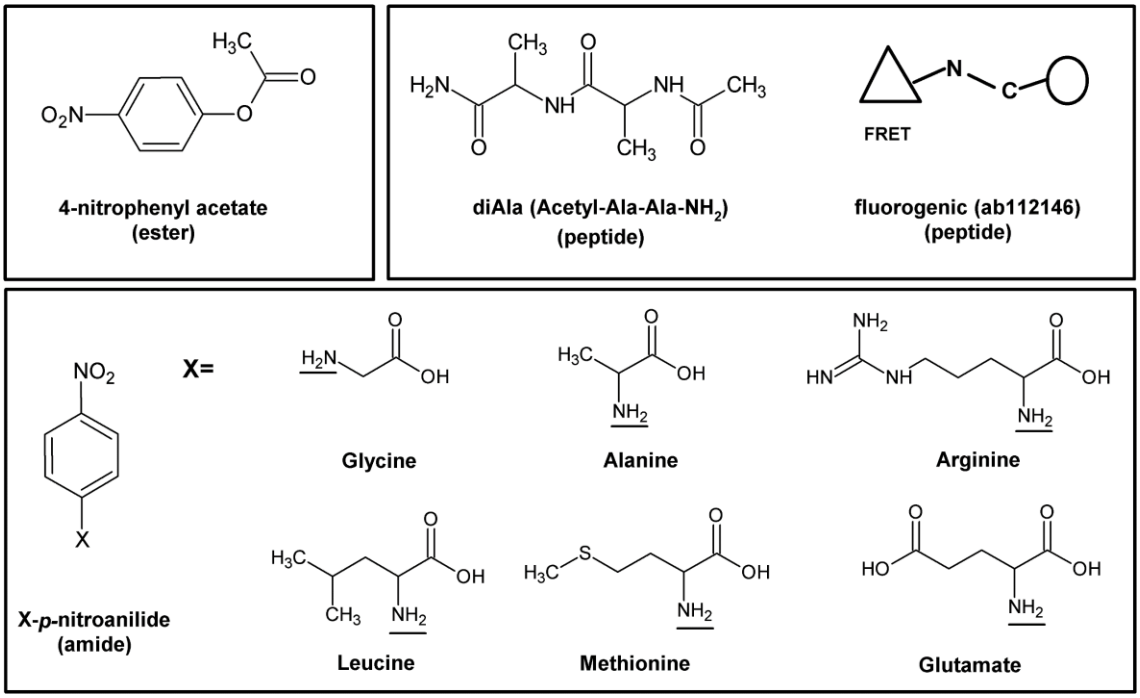
The results obtained in this chapter indicate that successful implementation of computational design is dependent on the properties of input scaffolds. While a significant destabilization of a zinc-finger fold occurred in the case of RD01 and RD01v2, the method was robust enough to allow the redefinition of the villin headpiece folding mechanism in RD02. Interestingly, the redesign of a structural Zn(II) site (Sp1f2) and the design of a catalytic Zn(II) site (HP35) converged to similar physicochemical properties of the corresponding RD peptides, pointing to similar contributions of the designed residues in observed scaffold features. Indeed, similar Zn(II) binding affinities were also reported for native protein scaffolds where a $(His)_3$-Zn(II) coordination motif was extensively designed using Rosetta, suggesting that improvements of peptide-metal interactions through additional rounds of computational design are unlikely.

5

# 5. Hydrolytic Activity of Designed Metallo-peptides

## 5.1 Introduction

The experimental validation of protein designs requires to test if the target function is empirically reproduced. RD peptides were designed in Chapter 2 to act as proteases towards the model diAla peptide (Scheme 5.1), where the designed AS would be able to effectively activate bulk water molecules to perform nucleophilic attack on the N-C peptide bond upon substrate binding. The RD peptides were expected to present hydrolase activity towards substrates amenable for bond cleavage, since strict substrate specificity was not expected in such small and flexible scaffolds. Although high substrate specificity has been a typical feature attributed to the remarkable catalytic efficiency of native enzymes, in recent years a more comprehensive view on enzyme function has been gained through an increased focus on their ability to catalyse other non-native reactions.[187] This so called enzymatic promiscuity is proposed to play an important adaptive role throughout protein evolution and is commonly explored in protein design projects.[188–190] At the molecular level it has been linked to protein dynamics, since the high conformational space of polypeptide chains allows to accommodate different substrate molecules at the AS via induced-fit or conformational selection mechanisms (Annex 1 and references therein). In the particular case of metalloenzymes, enzymatic promiscuity may also originate from the incorporation of non-native metal ions with different chemistries.[191]




**Scheme 5.1 –** Substrates tested for esterase, amidase and peptidase activity.

As it has been described in previous Chapter 4, the RD-Zn(II) complexes present a folded structure in solution that resembles that of their native counterparts, which nonetheless appear to

be highly flexible given their marginal thermal stability. This indirectly points to a depart of the structures from the backbone and side chain conformations modelled in Chapter 2. Such conformational flexibility may lead to catalytic activity towards different substrates, but on the other hand it can be unproductive if it leads to distorted AS geometries that render the RD-Zn(II) complexes inefficient or even inactive catalysts. Structural organization of ASs is not a pre-requisite for function, as it has been increasingly acknowledged in the study of intrinsically disordered proteins and catalytic activity of molten globules.[192] Nonetheless, in the case of RD designs, large deviations from the modelled Zn(II)-$Glu_{cat}$ interactions (which are based on MP conserved structural features) may not be tolerated for efficient activation of bulk water molecules. The general hydrolytic activity of RD-Zn(II) complexes was therefore tested towards different substrates (Scheme 5.1).

In Section 5.3.1 the esterase activity of RD peptides and native HP35 scaffold was screened and characterized towards the chromogenic substrate 4-nitrophenyl acetate (4-nPA, Scheme 5.1). This is a general esterase substrate that has been used as a “benchmark” for designed hydrolases, with most active catalysts being able to hydrolyse this activated substrate several orders of magnitude above the uncatalyzed reaction.[12] This includes redesigned native scaffolds, such as thioredoxin (PDZ2) with Orbit software [18] and the small protein calmodulin (AlleyCatE2) with molecular docking and the Rosetta software.[193] More complex design approaches have also been shown to present esterase activity, such as the *de novo* designed α-helical heptads (CC-hept) using Rosetta [194] or *de novo* designed peptide with ZF fold (BBA-B3) using an adaptation of the Cyana software.[195] Supramolecular fibrils (IHIHIQI) using short peptides designed with Rosetta software were also shown to have high esterase activities.[196] The development of artificial hydrolases is not restricted to computational methods, such as the case of designed small helical dimers (KO-42 and MID1) [154,197], coiled-coils (TRIL9CL23H) [198] or assemblies of helical tetramers ($^{A104}$AB3 and 6HB).[199,200] Smaller designs such ZF (CP-1) scaffolds [90] and the small peptide model of native MPs (mMMA) have also been reported to present hydrolytic activity.[201] Nonetheless, the catalytic efficiencies of artificial hydrolases fail to match those found for native ones by one or more orders of magnitude, thus putting into evidence the current limitations of employed design approaches.

After characterization of the esterase activity of RD peptides, the more relevant amidase and protease activities were screened in section 5.3.2. Amidase activity was probed using single amino acids bound to the chromogenic *p*-nitroanilide group by an amide bond (X-pNA, Scheme 5.1), similar to true peptidic bonds found in polypeptide chains. Finally, peptidase activity was tested towards fluorogenic peptide substrates and the target model substrate diAla. As it will be discussed below, the RD peptides failed to present amidase and peptidase activity under the tested assay conditions. This is to be compared with the few reports on the development of artificial proteases. One serine-carboxyl protease has been computationally redesign to hydrolyse immunogenic α-gliadin oligopeptides using Rosetta.[202] Apart from this, the large majority of reported artificial metalloproteases has been based on small organometallic complexes which

usually cleave peptide bonds through alternative catalytic mechanisms, such as oxidation, formation of reaction intermediates dependent on residue side chain functionality and employment of other metal ions.[35,203–205] A direct comparison with these systems was not approached, since the main focus was to evaluate the computational approach employed in Chapter 2 in light of other enzyme design efforts. Nonetheless, these small organometallic complexes provide important clues on the chemical requirements for efficient peptide bond hydrolysis.[44,206]

The results obtained in this chapter were used to guide the computational redesign approach employed in Chapter 2. Following characterization of RD01 hydrolytic activities and optimization of experimental conditions, the RD01v2 and RD02 peptides were designed with the expectation of developing improved catalytic efficiencies. Comparison of the results obtained in the following sections with computational scores obtained in Chapter 2 did not allow to draw conclusions on the observed hydrolytic activities, which triggered a more detailed structural analysis of the RD-Zn(II) complexes in Chapter 6.

## 5.2 Materials and Methods

UV-Vis spectroscopy assays were done with either 40 mM HEPES, 50 mM NaCl at pH 7.5 or 40 mM CHES, 50 mM NaCl at pH 9.0 in quartz cuvettes ($V_T$= 900 µL, path-length *l*= 1 cm) or in 96-well plates, ($V_T$ = 300 µL, *l* ~ 0.7 cm). Zn(II) additions were done from 1.033 or 10.33 mM $ZnCl_2$ stock solutions (prepared as previously described in Chapter 4). After substrate additions, up to 5% $\Delta V_T$ was achieved and therefore, no corrections of reagent concentrations were considered. Absorbance values were recorded at 25 °C in a Cary 100 Bio spectrophotometer with a Peltier temperature controller (integration time 0.2 s or 0.5 s, bandwidth 2 nm, scan speed 300 nm/min), or in 96-well plates at room temperature in a Tecan Infinite F200 microplate reader (also used for fluorescence assays).

*4-Nitrophenyl acetate*: 4-nPA was purchased from Sigma-Aldrich and prepared by diluting the solid (MW 181.15 g/mol, 18.2 mg/mL) in ACN up to 100 mM concentration (HPLC gradient grade), stored at 4 °C and used up to 2 days.[33] In assays done in the presence of peptide as catalyst (*apo*) the concentration varied between 2.5 and 15 µM. In assays done in the presence of peptide-Zn(II) complex as catalyst (*holo*), $ZnCl_2$ concentrations varied between 2.5 and 60 µM and peptide concentrations between 2.5 µM and 15 µM. For *holo* assays, the peptide was first added and equilibrated in buffer solution prior to addition of Zn(II). Afterwards, at least 1h of incubation was allowed before the beginning of assays to ensure that the peptide-Zn(II) complex formation reached equilibrium. Control assays of the uncatalyzed reaction in buffer were done with the addition of only 4-nPA (background) or with substrate and 10 µM $ZnCl_2$ (Zn(II)). In preliminary assays at 2.5 and 5 µM peptide concentrations, the final volume percentage of ACN varied with

[33]For assays where initial 4-nPA concentration was 6 mM, a 200 mM 4-nPA in ACN stock solution was used instead (36.2mg/mL)

corresponding amounts of 100 mM 4-nPA added (0.25-6% $V_T$). In *apo, holo* and control assays at 15 μM peptide, the final volume percentage of ACN was kept constant (5% $V_T$) by adding variable amounts of 100 mM 4-nPA and ACN according to initial assay conditions, with additions up to 45 μL (5% $V_T$) corresponding to 5 mM 4-nPA. Formation of the product 4-nitrophenol or phenolate (4-nP) was followed by single-wavelength readings of absorbance at 400 nm ($A_{400}$) every 2 seconds (RD02) or 30 seconds (RD01 and RD01v2) in cuvette assays at pH 7.5 (integration 0.5 seconds) and every 0.2 seconds at pH 9.0 (integration 0.2 seconds). In plate-well assays, readings were made at 405 nm in variable 1-10 min intervals. Initial rates recorded until *c.a.* 2-3% total 4-nP formation, 1-2h at pH 7.5 and approximately 10 min at pH 9.0. After addition of 4-nPA and equilibration (1 min in cuvettes, up to 5 min in plate-wells), the $A_{400}$ was used as $A_{400}(0)$. Rates of product formation $k_{obs}$ calculated by equation 5.1:

$$k_{obs}(M.s^{-1}) = \frac{A_{400}}{l\varepsilon} \times \frac{1}{s} = \frac{[4\text{-}nP]}{s} \quad (5.1)$$

where the Beer-Lambert law was used to convert $A_{400}$ to [4-nP] using $\varepsilon_{400}$= 12754±201 $M^{-1}cm^{-1}$ at pH 7.5 and $\varepsilon_{400}$= 18257±216 $M^{-1}cm^{-1}$ at pH 9.0. Linearity considered when $R^2 > 0.99$ over recorded time. The $\varepsilon_{400}$ values were determined under the same experimental conditions (including 5% ACN) by measuring $A_{400}$ values from 0-71.8 μM dilutions of 8.15 mM (pH 7.5) and 12.46 mM (pH 9.0) 4-nitrophenol stock solutions. These values were used as an approximation for plate-well assays, $\varepsilon_{405} \approx \varepsilon_{400} \times 0.7l$, where *l*=1 cm in cuvettes and the 0.7 the correction factor for path-length (approximate sample height). Rates of product formation obtained for control assays were subtracted in the corresponding catalyst assays (*apo* and *holo*). The first order rate constant $V_{cat}$ was calculated according to equation 5.2:

$$V_{cat}(s^{-1}) = \frac{(k_{obs,catalyst} - k_{obs,background})}{[catalyst]} \quad (5.2)$$

for assays with different concentrations of catalyst. Second-order rate constant $k_2$ (or turnover number) calculated by equation 5.3:

$$k_2\ (M^{-1}s^{-1}) = \frac{V_{cat}}{[4-nPA]} \quad (5.3)$$

All cuvette assays were performed at least twice ($n \geq 2$) using 4-nPA stock solutions prepared independently. Data plots correspond to average values and error bars to the S.E. calculated as determined in previous chapter. Data were fitted to linear model in Origin Pro 2016 with Levenberg-Marquardt algorithm using instrumental weighting (more weight for data points with lower S.E.).

*p-Nitroanilide derivatives:* Single amino acid p-nitroanilides (X-pNA) were purchased from Sigma-Aldrich and their solubility up to 100 mM concentration was first tested with different solvents, as summarized in Table 5.1.

**Table 5.1 –** X-pNA substrate solubility tests for 100 mM stock solutions.

| X-pNA | Description | Molecular Weight (g/mol) | Solvent |
|---|---|---|---|
| Gly[a] | Glycine p-nitroanilide | 195.18 | $(CH_3)_2CO$ |
| Ala[b] | L-alanine p-nitroanilide hydrochloride | 245.66 | $H_2O$ |
| Leu[b] | L-leucine p-nitroanilide | 251.28 | ACN |
| Met[b] | L-methionine p-nitroanilide | 269.32 | ACN |
| Glu[c] | L-glutamic acid 1-(4-nitroanilide) | 267.24 | $H_2O$ + NaOH (2 µL 5 M) |
| Arg | L-arginine p-nitroanilide dihydrochloride | 367.20 | $H_2O$ |

a- Not soluble in $H_2O$, ACN, 50:50 $H_2O/(CH_3)_2CO$.
b- Solubility in $H_2O$ not tested.
c- Not soluble in $H_2O$, ACN.

The stability of the respective X-pNA substrates was tested up to a final concentration of 5 mM in assay buffer at pH 7.5, with no precipitate formation or colour development observed overnight at room temperature. Assays were done with 15 µM peptide and 60 µM $ZnCl_2$ (*holo*), with 1 h incubation period prior to the addition of X-pNA substrate. Control assays were done with the addition of only X-pNA (background), X-pNA and 60 µM $ZnCl_2$ (Zn(II)), or X-pNA and 1.5 µM peptide (*apo*). The later corresponds to the amount of free peptide estimated to be present in *holo* assays. X-pNA substrates were added up to 15 µL in microplate or 45 µL in cuvette format (5% $\Delta V_T$). The formation of the product p-nitroaniline (pNA) was monitored by following $A_{405}$ increase at variable time intervals [15 min-12 h], with total recorded time of 3 days. The $A_{405}$ values after 1 min of X-pNA substrate addition were used as $A_{405}(0)$. The product extinction coefficient $\varepsilon_{405}$=5329 $M^{-1}cm^{-1}$ was determined under the assay conditions by measuring $A_{405}$ values of 0-100 µM dilutions from a 3.26 mM pNA stock solution.

*Matrix Metalloprotease (MMP) activity assay:* the ab112146 MMP Activity Assay Kit (fluorometric – green) was purchased from Abcam and assays were performed according to the provided protocol. Samples were 15 µM peptide (*apo*), 15 µM peptide and 60 µM $ZnCl_2$ (*holo*) and 60 µM $ZnCl_2$ only (Zn(II)) dissolved in the provided buffer (pH 7.5). Fluorescence signal was monitored in microplate format at room temperature ($V_T$= 100 µL) with an Ex/Em = 490/525 nm for variable time intervals up to a total duration of 2 days.

*Model diAla:* the diAla peptide (Acetyl-Ala-Ala-$NH_2$) was purchased from POP-UP (Peptide Synthesis Facility at University of Porto, Portugal) with 92% purity. A 1 mM stock solution was prepared in $D_2O$ 50mM NaCl (molecular weight 201.11 g/mol, 0.2 mg/mL) and its pH adjusted to 7.47 with additions of concentrated NaOH and HCl solutions. Assays were done by direct addition of both 25 µM peptide and 75 µM $ZnCl_2$ to diAla 1 mM solutions ($V_T$= 600 µL). $^1$H NMR assays

were performed in a Bruker Avance II+ 800 MHz at 25°C under the supervision of Dr. Manolis Matzapetakis (Biomolecular NMR Lab, ITQB/NOVA). The spectra were recorded immediately after the addition of peptide and metal to 1 mM diAla solution and followed at variable time intervals up to a total duration of at least 2 days.

## 5.3 Results and Discussion

### 5.3.1 Ester hydrolysis

The general esterase activity of the RD peptides was tested by their ability to catalyse the hydrolysis of the 4-nPA substrate, according to equation 5.4:

$$4\text{-}nitrophenyl\ acetate + H_2O \rightarrow 4\text{-}nitrophenolate + acetic\ acid \quad (5.4)$$

The formation of the product 4-nP can be monitored spectrophotometrically at 400 nm. The corresponding extinction coefficient ($\varepsilon_{400}$) is strongly dependent on the protonation state of the molecule due to the equilibrium *4*-nitrophenol ↔ *4*-nitrophenolate, therefore this value was determined for each pH value tested.

Control assays of uncatalyzed 4-nPA hydrolysis in buffer at pH 7.5 were performed and results are shown in Figure 5.1. UV-Vis spectra of 4-nP formation were recorded over time after addition of 100 µM 4-nPA, showing a linear increase of the band at 400 nm concomitant with a decrease of the band at 260 nm and an isosbestic point at *c.a.* 308 nm.

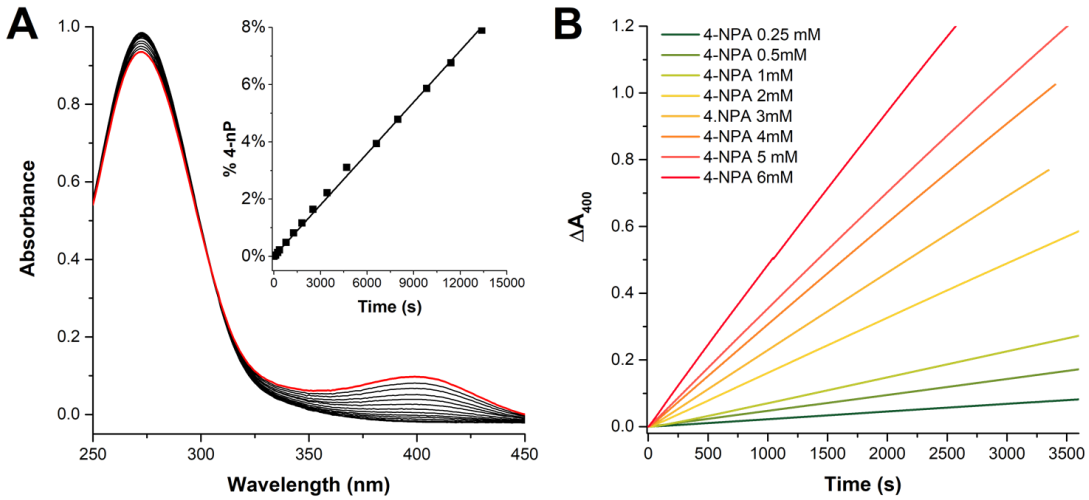



**Figure 5.1 –** Control assays of uncatalyzed 4-nPA hydrolysis in buffer at pH 7.5.
A – UV-Vis spectra obtained over time showing the product 4-nP formation at 400 nm. Red line corresponds to final point of the reaction. Inset graph: corresponding percentage of product formation calculated using the initial concentration of 0.1 mM 4-nPA and using an $\varepsilon_{400}$ of 12754±201 $M^{-1}cm^{-1}$ (details in methods section). B – Absorbance increase at 400 nm ($A_{400}$) followed over time for different initial concentrations of added 4-nPA (from 0.25 to 6 mM). Spectra were obtained in 40 mM HEPES 50 mM NaCl at pH 7.5, 25 °C.

The linear increase was observed until at least 8% product conversion over more than 3 hours, with an approximate formation rate of 2%/h. Linearity of 4-nP formation over time was also observed for assays with higher initial 4-nPA concentrations up to 6 mM (Figure 5.1.B), although rates tended to be lower than 2%/h for higher concentrations of 4-nPA indicating product precipitation. No changes in pH due to acetic acid formation occurred for 1 mM 4-nPA assays over 20h.[34]

Following the characterization of kinetics of the uncatalyzed reaction, the RD peptides and native HP35 were screened in microplate assays for their ability to catalyse the hydrolysis of 4-nPA. Native Sp1f2 was not tested since it has been reported no hydrolytic activity towards 4-nPA under similar experimental conditions.[89] The initial rates of product formation were recorded for the 0.1-6 mM 4-nPA concentration range and results are shown in Figure 5.2. A linear increase of first order rate constant $V_{cat}$ as a function of substrate concentration was obtained for all three RD peptides up to 2 mM.

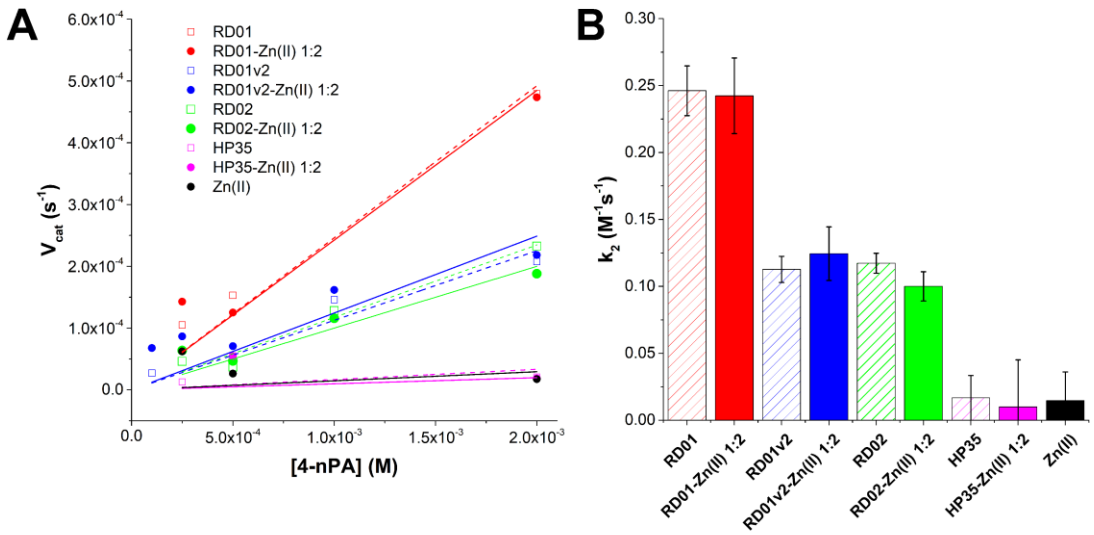



**Figure 5.2 –** Microplate screening assays of 4-nPA hydrolysis by RD peptides and native HP35 at pH 7.5.
A – First-order rate constant $V_{cat}$ values obtained for the 0.25 to 2 mM 4-nPA concentration range tested using 5 μM peptide (*apo*, open symbols) and 5 μM peptide, 10 μM $ZnCl_2$ (1:2 *holo*, closed symbols). Data obtained in 40 mM HEPES 50 mM NaCl, pH 7.5 at room temperature and fitted to linear model for *apo* (dashed lines) and *holo* forms (solid lines). B – Corresponding second-order rate constants $k_2$, with *apo* forms in line filled bars and *holo* in filled bars. Standard error of fitted linear models shown in solid black lines.

[34] Assuming a linear increase throughout the 20h, this corresponds to 0.4 mM 4-nP. This value is above the range of 4-nP concentration build up occurring for *e.g.* 5 mM 4-nPA assays (5 mM x 3% = 0.15 mM 4-nP) and therefore no pH values changes were expected to occur in those assays.

This trend was not kept for higher concentrations, which is attributed to measurement errors due to low solubility of 4-nPA under these conditions (although no precipitate formation was observed by eye). For this reason, all further assays were considered only up to 2 mM initial 4-nPA concentration and second order rate constants $k_2$ determined only for this range. For HP35 and control Zn(II) assays, $V_{cat}$ values did not followed a linear increase as function of substrate concentration, which indicates no catalytic activity towards the target reaction.

Both *apo* and *holo* forms present similar activity in case of RD01, with $k_2$ values (or turnover number) inferior to 1 $s^{-1}$. This prompted the design of RD01v2 and RD02 as described in Chapter 2. However, the RD01v2 peptide presented $k_2$ values lower than those obtained for both *apo* and *holo* forms of RD01. The RD02 peptide presented similar $k_2$ values to RD01v2 and HP35 did not present values above those of the uncatalyzed reaction, indicating that RD02 catalytic activity was due to the designed sequence changes. Control assays of Zn(II) did not present activity above background levels.

The microplate screening assays were made in the presence of 5 μM peptide concentration, and in the case of *holo* assays a 1:2 peptide-Zn(II) ratio was used. Considering that the Zn(II) binding constants obtained for the RD peptides in Chapter 4 are in the $10^5$ $M^{-1}$ range, in these *holo* assays the peptide-Zn(II) complex was not fully formed (the effect of using 10 mM *vs.* 40 mM HEPES buffer does not lead to significant changes in the binding constants of RD01). The unbound histidine residues of the free peptide are responsible for the $k_2$ values obtained in *apo* assays given the Lewis acid character of the imidazole ring and precedent results.[89] Therefore, the effect of using different concentrations and peptide-metal ratios in the determination of $k_2$ values was addressed in more detail for the RD01 peptide in the cuvette format and the results are shown in Figure 5.3.[35]

For both 5 μM and 15 μM RD01 assays the obtained $V_{cat}$ values were similar for the range of tested 4-nPA concentrations.[36] In the case of 5 μM assays using either a 1:1 or 1:2 peptide-Zn(II) ratio (< 50% complex formation) the amount of free peptide in solution is significant and therefore, the determined $k_2$ values contain contributions from both active species (free Zn(II) was not considered as a catalytic species). In the 15 μM *holo* assays with 1:4 peptide-Zn(II) ratio the complex is the major species in solution and the contributions from the free peptide are minimized.[37] Under these conditions the $k_2$ values decreased significantly between *apo* and *holo* forms, from 0.27 $s^{-1}$ to 0.19 $s^{-1}$, respectively.

[35] Only cuvette assays were made from this point on since a tighter control of experimental conditions could be achieved, such as higher sample homogeneity, more data point collection, and 4-nP $\varepsilon_{400}$ determined under this format.

[36] Assays with 2.5 μM RD01 were also performed but data was not considered due inconsistency of results and high $k_2$ associated errors.

[37] Higher peptide concentrations and peptide-Zn(II) ratios were not considered due to a compromise between the lowest amount of used peptide per assay and minimizing the possibility of unspecific interactions with free Zn(II).

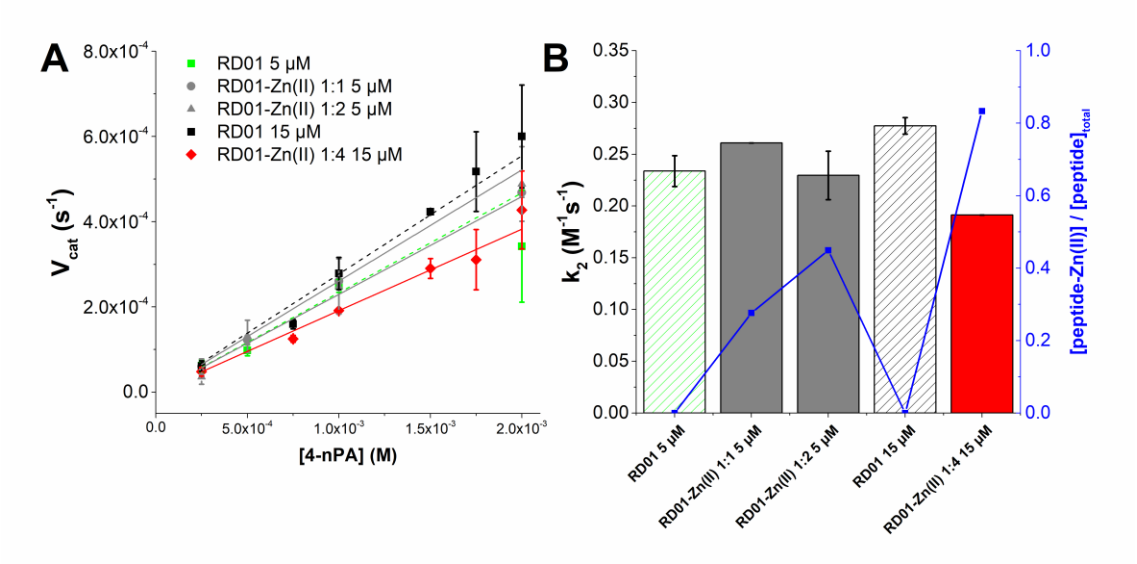



**Figure 5.3 –** 4-nPA hydrolysis by RD01 peptide determined at different peptide-Zn(II) ratios and concentrations at pH 7.5.
A – First-order rate constant $V_{cat}$ values obtained for the 0.25 to 2 mM 4-nPA concentration range tested using 5 (green) and 15 μM (black) RD01 (*apo*) and 5 to 60 μM $ZnCl_2$ (1:1, 1:2 and 1:4 *holo*). Data obtained in 40 mM HEPES 50 mM NaCl, pH 7.5 at 25 °C and fitted to linear model for *apo* (dashed lines) and *holo* forms (solid lines). Values correspond to the average of at least two independent assays and error bars to the corresponding S.E.. B – Corresponding second-order rate constants $k_2$ (left axis), with *apo* forms in dashed bars and *holo* in filled bars. Standard error of fitted linear models shown in solid black lines. Ratio of formed peptide-Zn(II) complex *vs.* total peptide added for each assay in right axis.

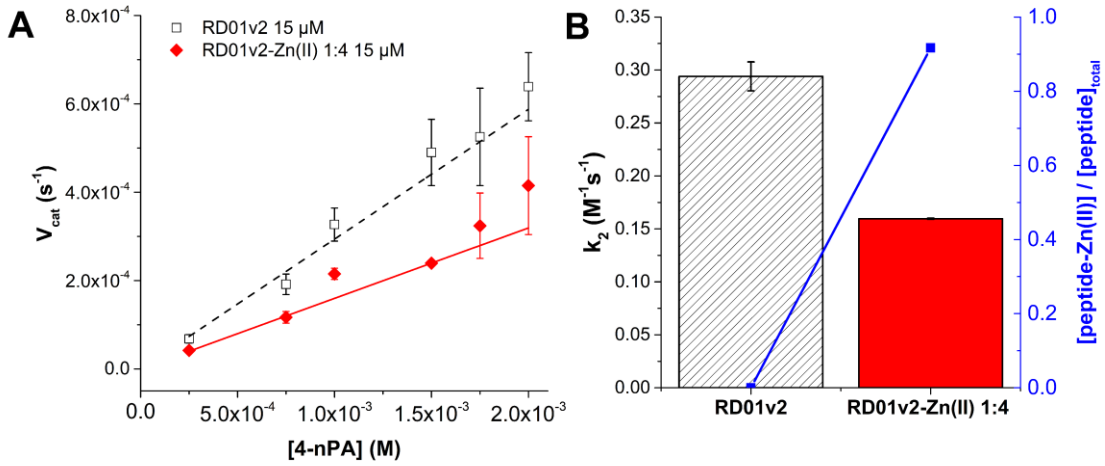



**Figure 5.4 -** 4-nPA hydrolysis by RD01v2 peptide at pH 7.5.
A – First-order rate constant $V_{cat}$ values obtained for the 0.25 to 2 mM 4-nPA concentration range using 15 μM RD01v2 (*apo*, open symbols) and 15 μM, 60 μM $ZnCl_2$ (1:4 *holo*, closed symbols). Data obtained in 40 mM HEPES 50 mM NaCl, pH 7.5 at 25 °C and fitted to linear model for *apo* (dashed lines) and *holo* forms (solid lines). Values correspond to the average of at least two independent assays and error bars to the corresponding S.E.. B – Corresponding second-order rate constants $k_2$ (left axis), with *apo* forms in dashed bars and *holo* in filled bars. Standard error of fitted linear models shown in solid black lines. Ratio of formed peptide-Zn(II) complex *vs*. total peptide added for each assay in right axis.

Assays at 15 μM RD01v2 and 1:4 peptide-Zn(II) ratio were also carried out to re-evaluate the $k_2$ values obtained in the microplate format (Figure 5.4). The *apo* form presents $k_2$ values similar

to those of RD01, while for the *holo* form a lower value is obtained. In the latter case, this may be due to less contributions from the free peptide in solution, since RD01v2 has higher binding affinity for Zn(II) than RD01.

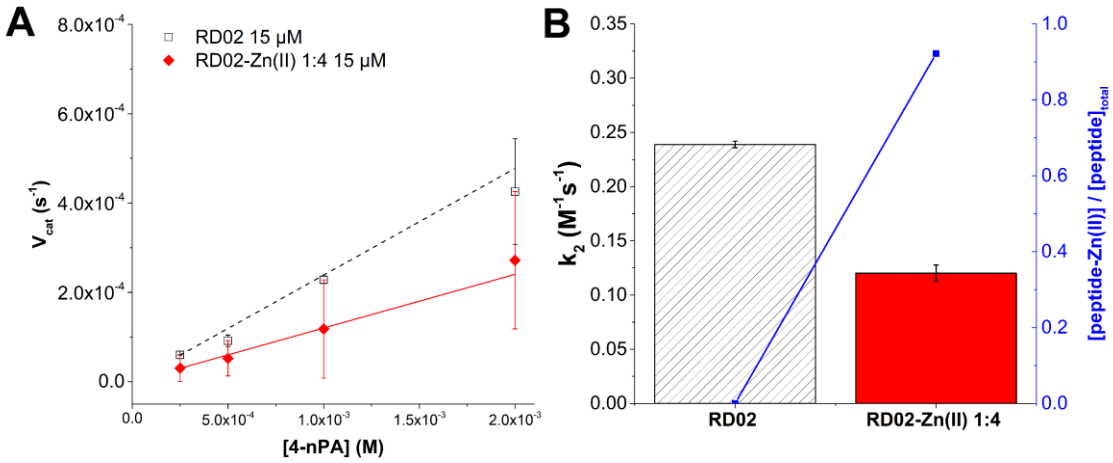



**Figure 5.5 -** 4-nPA hydrolysis by RD02 peptide at pH 7.5.
A – First-order rate constant $V_{cat}$ values obtained for the 0.25 to 2 mM 4-nPA concentration range using 15 µM RD02 (*apo*, open symbols) and 15 µM, 60 µM $ZnCl_2$ (1:4 *holo*, closed symbols). Data obtained in 40 mM HEPES 50 mM NaCl, pH 7.5 at 25 °C and fitted to linear model for *apo* (dashed lines) and *holo* forms (solid lines). Values correspond to the average of at least two independent assays and error bars to the corresponding S.E.. B – Corresponding second-order rate constants $k_2$ (left axis), with *apo* forms in dashed bars and *holo* in filled bars. Standard error of fitted linear models shown in solid black lines. Ratio of formed peptide-Zn(II)complex *vs.* total peptide added for each assay in right axis.

The assays with RD02 were also performed at 15 µM peptide and 1:4 peptide-Zn(II) ratio and results are shown in Figure 5.5. The *apo* form presented $k_2$ values slightly lower than those obtained for RD01 and RD01v2, and this activity is again attributed to the Lewis acid character of the three designed histidine residues. However, the *holo* form presented the lowest $k_2$ with a value of 0.12 $s^{-1}$, pointing to distinct scaffold contributions in the catalytic activities of the peptide-Zn(II) complexes.

A summary of catalytic activities towards 4-nPA hydrolysis of the designed peptides is made in Table 5.2. Interestingly, for all RD peptides the *apo* version was always the better catalyst. As it will be described below, although 4-nPA hydrolysis was not the target modelled reaction, it is the most informative in establishing computational-experimental correlations. Indeed, the hydrolytic activities are in clear opposition to the complexity of designs considered in Chapter 2: while RD01 was the most minimalistic design and RD02 the most complex (with supposedly more favourable score features), RD01-Zn(II) complex is the best catalyst, followed by RD01v2-Zn(II) (with supposedly unfavourable features) and finally by RD02-Zn(II). The influence of having used two distinct AS models in the case of RD01 ($4AIG_{AS}$:$diAla_{min}$) and RD01v2/RD02 ($MA(M)_{AS}$:diAla) is ruled out given the similarity of their hydrolytic activities.

**Table 5.2 –** Summary of $k_2$ values for RD peptides obtained in 40 mM HEPES 50 mM NaCl, pH 7.5 at 25 °C.

| Scaffold | $k_2$ ($s^{-1}$) | |
|---|---|---|
| | *apo* | *holo* |
| **RD01** | 0.27±0.01 | 0.19±0.00 |
| **RD01v2** | 0.29±0.01 | 0.15±0.00 |
| **RD02** | 0.23±0.00 | 0.12±0.01 |
| **HP35**[a] | 0.02±0.02 | 0.01±0.02 |

a – Values obtained in microplate assay, 5 μM peptide and 1:2 peptide-Zn(II) ratio.

The reasons for why the computational approach employed in Chapter 2 did not success in designing peptide-Zn(II) complexes with high catalytic activities will be addressed in further detail in the following chapter. Nonetheless, comparison with other natural and designed systems is shown in Figure 5.6.

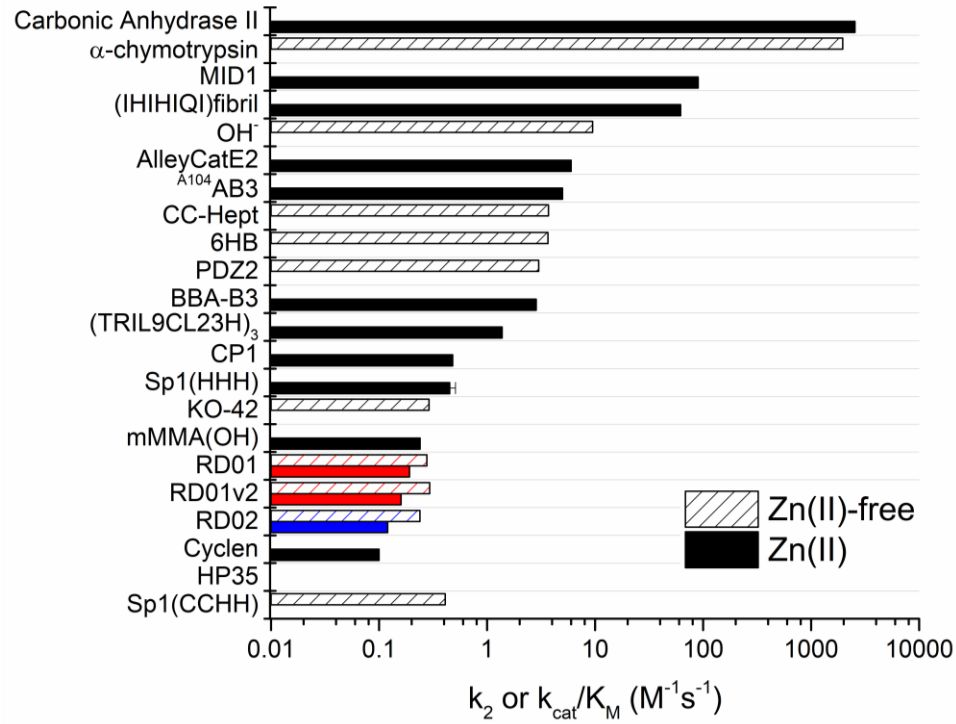



**Figure 5.6 –** Comparison of catalytic efficiency towards the 4-nPA ester between RD peptides and other designs.
Values in log scale correspond to $k_2$ or $k_{cat}/K_M$ obtained at pH 7.5. Designs not dependent on Zn(II) in dashed bars (Zn(II)-free) and in solid bars for Zn(II)-dependent systems (Zn(II)). Native Sp1f2 [89] and HP35 included for comparison. In case of Sp1(HHH), error bars correspond to range of values obtained for different $(His)_3$ Sp1f2 variants. $k_{cat}/K_M$ of native enzymes carbonic anhydrase [207] and α-chymotrypsin [208] also included for comparison. pH dependence of background reaction is shown by the $k_2$ values of the hydroxide ion ($OH^-$).[209] References of designs given in main text.

Values of $k_2$ (or $k_{cat}/K_M$) are comparable with other redesigned zinc fingers, including Sp1f2 variants with three histidine residues (*apo*) and the CP1 consensus peptide (*holo*) not specifically

designed towards 4-nPA hydrolysis.[90] Nonetheless, these values are one order of magnitude lower than those of the BBA-B3 zinc finger (*holo*) designed specifically towards the hydrolysis of 4-nPA.

Peptide-substrate interactions and right positioning of the catalytic residues during the reaction appear to play a key role in catalytic efficiency of designed systems, since the simple inclusion of catalytic residues does not lead necessarily to better catalysts. In the case of RD01 and RD01v2 peptides, inclusion of the $Glu_{cat}$ residue did not lead to improvements when compared to other Sp1f2 variants, where only first sphere Zn(II) ligands were considered.[89] Another example is the mini matrix metalloprotease mMMA with less than 20 residues and a flexible fold, where the monoprotonated *holo* form has comparable $k_2$ values as RD01 and RD01v2. Indeed, these designed systems present similar values as the organic complex cyclen-Zn(II), suggesting only modest contributions of the introduced catalytic residues to the obtained $k_2$ values. However, for designed protein-Zn(II) complexes with higher structural complexity such as the coiled coil TRIL9CL23H, AlleyCatE2, $^{A104}$AB3 tetramer assembly and MID1 dimer, there is one to two orders of magnitude increase in catalytic efficiency with display of Michaelis-Menten kinetics. This is also the case of the heptapeptide IHIHIQI, which exhibits hydrolytic activity only upon Zn(II)-mediated supramolecular assembly into fibrils. For metal-free esterases, both redesigned thioredoxin PDZ2, six-helical bundle 6HB and *de novo* designed heptad CC-hept show similar catalytic efficiencies, and the KO-42 dimer being the exception with lower values comparable to those of RD designs.

Despite several design approaches and different types of scaffolds used, the catalytic efficiency of constructs is still below those found native enzymes, even for promiscuous activities such as the case of 4-nPA hydrolysis. This is the case of native metalloenzyme Carbonic Anhydrase II and the serine protease α-chymotrypsin, which present $k_{cat}/K_M$ values four orders of magnitude higher when compared to RD peptides.[207,208] These, together with the exceptions of MID1 and IHIHIQI, show catalytic rate enhancements higher than the hydroxide ion, which points to specific protein-substrate interactions involved in the catalytic mechanism. With more complex structures the network of interacting residues increases, leading to the possibility of establishing favourable protein-substrate interactions or proper activation of the catalytic species. One extreme case is the 43C9 engineered catalytic antibody, with $k_{cat}/K_M$ value over $10^5$ $M^{-1}s^{-1}$ although presenting strong product inhibition. Resurrected enzymes have also been shown to hydrolyse 4-nPA with $k_{cat}/K_M$ value of 11.7 $M^{-1}s^{-1}$, and it has been argued that the conformational flexibility of these constructs allows to sample more productive substrate and transition-state interactions.[210]

While in metal-free esterases the catalytic activity is usually attributed to cysteine and histidine residues, in zinc enzymes the active species is an activated water molecule bound to the metal ion. First sphere Zn(II)-$OH_2$ interactions and hydrogen bonds with the catalytic glutamate lead to charge stabilization of the oxygen atom containing free electron pairs, thus decreasing the *pKa* of the bound water. Second-sphere interactions also play a role in substrate activation and proper

metal-site geometry.[143] In order to probe the activation of water molecules in RD01 and RD01v2 designs, the effect of pH on $k_2$ values was addressed in preliminary assays at pH 9.0, with results shown in Figure 5.7.

For both *apo* and *holo* forms there is an increase in $k_2$ values going from pH 7.5 to pH 9.0. In the case of *holo* forms the increase is more pronounced, inverting the trend observed at pH 7.5 for higher $k_2$ values of *apo* forms. In the particular case of RD01v2 *holo*, the values are higher than those obtain for RD01

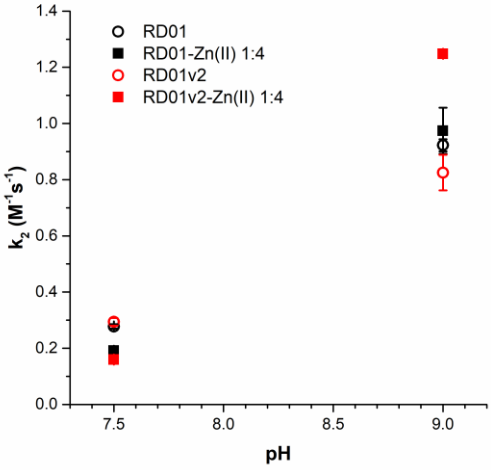



**Figure 5.7 –** Effect of pH on 4-nPA hydrolytic activity of RD01 and RD01v2 peptides. Second-order rate constants $k_2$ values obtained for the 1 to 2 mM 4-nPA concentration range using 15 μM peptide (*apo*, open symbols) or 15 μM peptide, 60 μM $ZnCl_2$ (*holo*, closed symbols) in 40 mM HEPES 50 mM NaCl, pH 7.5 or 40 mM CHES 50 mM NaCl, pH 9.0 at 25 °C. Values correspond to the average of at least two independent assays and error bars to the corresponding S.E.

Although the Zn(II)-complexes have not been structurally characterized at pH 9.0, these results suggest that in RD01 *holo* assays the free (unstructured) peptide is the major species in solution, while for RD01v2 *holo* the Zn(II)-complex remains formed. A more detailed study on the effects of pH in catalytic activity would be required to draw conclusions on whether or not the RD peptides are able to properly modulate the *pKa* of the Zn(II)-bound water molecule.

### 5.3.2 Amide and peptide bond hydrolysis

Given that both esterase and amidase activities can occur by hydrolysis of activated water molecules, the catalytic activity of RD peptides towards amide bonds in pNA-amino acid derivatives was tested by considering the equation 5.5:

$$X\text{-}p\text{-}nitroanilide + H_2O \rightarrow X + p\text{-}nitroaniline \quad (5.5)$$

where X= Ala, Gly, Glu, Arg, Met, Leu. The formation of the product pNA was monitored spectrophotometrically at 405 nm. At pH 7.5 the uncatalyzed reaction is negligible and was considered

as a control assay. No clear activity above control levels was detected for holo assays at 0.5 and 5 mM X-pNA for RD01, RD01v2, RD02 and HP35 in microplate format (not shown). This was also the case for control Zn(II) and apo assays, with no significant $A_{405}$ changes observed over 2 days. Additional tests in cuvette format were made for Met-, Ala-, Glu- and Arg-pNA derivatives in the 0.167-15 mM range, with no detectable product formation.[38]

The RD peptides present no amidase activity over the tested conditions irrespective of the coupled amino acid residue.[39] This is in contrast with native MPs, which show a remarkable range of catalytic activities towards unnatural pNA derivatives. The MA(M) endopeptidase astacin presents very low activity towards single amino acid Ala-pNA ($k_{cat}/K_M$ 0.01 $M^{-1}s^{-1}$), much lower than for longer substrates such as Suc-$(Ala)_3$-pNA (> 22 $M^{-1}s^{-1}$).[211] Meprin A, another MAM(M) endopeptidase presents even higher catalytic efficiency towards alanine tripeptides (231 $M^{-1}s^{-1}$).[212] On the other hand, the leukotriene-A4 hydrolase, an MA(E) arginine aminopeptidase, shows high catalytic activity towards single amino acid derivatives, specially towards Arg-pNA with $k_{cat}/K_M$ of $3.7x10^4$ $M^{-1}s^{-1}$.[213] Indeed, this MP shows high stereospecificity towards D-amino acids and enhanced activity towards various unnatural amino acids, such as the *O*-benzyl ester of aspartic acid with a $k_{cat}/K_M$ of $1.75x10^5$ $M^{-1}s^{-1}$.

Given the low catalytic activity of RD peptides observed towards 4-nPA, it is reasonable to assume that there is no activation of water molecules at pH 7.5 to perform the nucleophilic attack on the amide bond. Whether this was also the case for substrates containing true peptide bonds was addressed by testing the peptidase activity of RD designs towards peptidic substrates. First, the commercial kit (MMP activity assay) of an oligopeptide-fluorogenic probe conjugate was employed to probe low levels of activity but no detectable fluorescence changes were observed above control levels over 2 days (not shown). With this approach, presumable limitations in the detection levels of peptidase activity could be more safely ruled out. Secondly, the model diAla substrate used in Chapter 2 was employed. NMR samples of 1mM substrate incubated at pH 7.5 with either RD01v2 or RD02 peptide-Zn(II) complexes showed no spectral changes over at least 2 days, pointing to the lack of Ala-Ala bond cleavage (not shown). Considering the CD far-UV assays of RD peptide-Zn(II) complexes in the presence of diAla (Chapter 4), this lack of peptidase activity is not attributed to major structural changes caused by some type of catalyst-substrate interactions. As discussed in Chapter 2, the control design of astacin using the diAla substrate model showed favourable Rosetta scores which could indicate potential activity towards this substrate. Nonetheless, favourable scores were also obtained for RD02, which did not present any clear distinction from RD01v2 in terms of diAla hydrolysis, the later showing overall unfavourable design features.

[38] Preliminary assays in CHES buffer pH 8.6 and CAPS buffer pH 10.0 were also performed and no activity was detected.

[39] Preliminary assays with Suc-$(Ala)_3$-pNA as substrate were done but no detectable product formation was found.

## 5.4 Conclusion

The catalytic proficiency of the designed RD peptides in Chapter 2 has been addressed in the present chapter. Following the synthesis and characterization carried out in previous Chapter 3 and 4, respectively, the experimental conditions were set to characterize separately both *apo* and *holo* forms of the RD peptides. Target peptidase activity was not detected for any of the RD peptides, including for the modelled diAla substrate. This was also the case for amidase activity towards pNA amino acid derivatives. Only low esterase activity towards 4-nPA was identified for all RD peptides, with RD01-Zn(II) complex presenting the highest hydrolytic activity among the peptide-Zn(II) complexes. In the case of RD02, successful introduction of a catalytic metal site was confirmed given the lack of esterase activity observed for native HP35. The esterase activities of peptide-Zn(II) complexes are within the same range of the corresponding metal-free peptides due to the contribution of unbound histidine residues and within range with other peptide/small-protein designs containing metal-sites. However, the obtained values are still 2 orders of magnitude lower than more complex designs and up around 4 orders of magnitude below the native enzymes. Given the flexible structural features and marginal stability of the peptide-Zn(II) complexes, it is unclear whether the designs failed to maintain the scaffold integrity or the proper active site pre-organization in order to effectively activate bulk water molecules. Although preservation of structure may not be a requisite for efficient catalysts, the propensity of RD-Zn(II) complexes to maintain proper active site organization will be addressed from the structural point of view in Chapter 6.

6

# 6. Structural Features of Designed Metallo-peptides

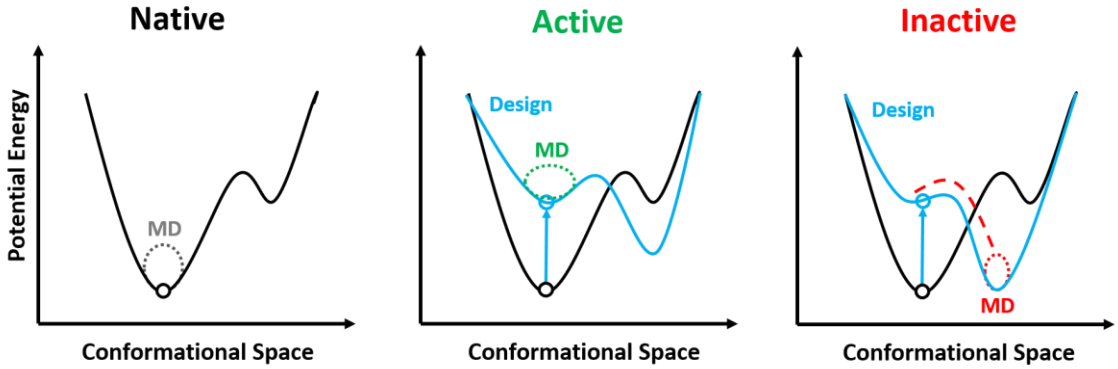
**Book chapter**

Carvalho HF, Barbosa AJM, Roque ACA, Iranzo O, Branco RJF. *Chapter 8 - Integration of Molecular Dynamics Based Predictions into the Optimization of De Novo Protein Designs: Limitations and Benefits.* In: Samish I, editor. Computational Protein Design. New York, NY: Springer New York; 2017. pp. 181–201.

## 6.1 Introduction

In enzyme design projects, the catalytic efficiency of initial candidates are usually rather modest, along with a plethora of produced inactive designs that are usually discarded without further consideration. Experimental solutions to this problem often lie in directed evolution techniques, from which new sequence changes are introduced on active designs and selected based on increased target function and/or stability.[24,214,215] Rationalization on the molecular basis for such low catalytic efficiencies often falls to a secondary role, thus limiting our current understanding of protein chemistry. In this regard, computational methods offer the opportunity to bridge the gap between theory and experimental findings.

Despite their potential in the development of new catalysts, CED tools have nonetheless some inherent limitations. For example, candidate designs are often ranked based on *ad hoc* approaches such as the one described in Chapter 2 for selection of the RD02 scaffold. Moreover, design flaws may be inaccessible by MM-based methods such as the one employed in Rosetta, where static representations of the systems in implicit solvent medium are done, often based on a single crystallographic structure.[40] This is because designs often differ significantly from their native counterparts in terms of their primary sequence and this can affect considerably their potential energy landscape, as represented in Scheme 6.1.




**Scheme 6.1 –** Representation of the potential energy landscape in the conformational space of backbone and side chains for native and enzyme designs.
Native enzymes have potential energy landscapes (black line) with well-defined global minima where active conformation is located (open circle). Sampled subspace during MD simulations may be limited to neighbouring regions (dashed regions). Designs present different potential energy landscapes due to introduced sequence changes (blue lines). Sampled subspace during MD simulations may be close to target active conformation (blue circle) in active designs (dashed green) but might drift considerably towards non-active conformations that correspond to new global minima (dashed red).

Indeed, as described in Chapter 2, new sequence variants are obtained by iterative repacking and energy minimization of native structure upon introduction of the AS of interest. During this

[40] Dynamics of the scaffold were nonetheless approached to some extent by using as input NMR-derived structures.

stage, backbone and AS side chain constrains are typically imposed and as a result the new sequence variant may correspond to shallow, local energy minima in the potential energy landscape.[41] Alternative backbone and side chain conformations corresponding to neighbour energy minima may become more stable and thus being preferentially adopted. This may impact greatly the solvent accessibility of AS residues and substrate binding-interaction energies. Moreover, interactions with solvent molecules may also dictate to a major extent the adoption of alternative conformations, and this is particularly relevant for RD peptides where a water molecule is expected to act as the fourth ligand in the substrate-free forms.

Calculation of energy profiles along the reaction coordinate would require the employment of DFT methods, but these are computationally very demanding and therefore limited to sampling of short time-scales (ps-ns). As a result, a relatively static treatment of protein-substrate interactions needs to be imposed.[216,217] The compromise between a realistic treatment of chemical reactions in fast time-scales and proper sampling of protein conformational space in longer ones has been commonly opted in favour of the latter in comprehensive enzyme design projects, where structural integrity and unfavourable catalytic interactions can be probed in advance by taking into consideration the intrinsic dynamical properties of designs and their evolved variants.[218] These can be addressed either experimentally by nuclear magnetic resonance (NMR) spectroscopy or computationally by the employment of atomistic MM simulation methods, such as Molecular Dynamics (MD). [219] In MD simulations both the biomolecule and solvent are explicitly described by force fields where particle interactions are described by simplified energy potentials. These can be integrated in time according to Newton's second law of motion to obtain forces acting on particles and calculate the resulting accelerations, which are then used to calculate new velocities and positions at each time step. The result is a simulation trajectory of the entire atomic system, which yields the conformational space available for the biomolecule to be explored under the employed simulating conditions of pressure, temperature and solvent composition.

MD simulations in the ns time-scale have been employed to guide several enzyme design projects. For example, MD-derived geometric descriptors have been employed in the study of Kemp eliminases and used to discriminate between active and inactive designs.[216,220] The conformational space explored in such cases is nonetheless restricted to sampled time-scales and may fail to capture major conformational changes occurring in slower regimes (μs-ms).[221] Regarding RD peptides, their inherent structural flexibility characterized in Chapter 4 begs the question whether the designs failed to maintain scaffold integrity or accurate AS pre-organization. Given their reduced size, MD methods are suitable to probe whether major conformational rearrangements occur in μs time-scales. While not a high-throughput method, recent hardware and software development have made investigations of protein dynamics in slow regimes increasingly

[41] This was the case in Chapter 2, where AS geometries were constrained during the design stage to avoid disruption of the metal site. Repacking and energy minimization without constraints led Zn(II) binding residues to depart from the orientations imposed by the pseudo-covalent bonds with the metal ion in a tetrahedral-like fashion, since the later was treated as part of the $diAla_{(min)}$ substrate model.

accessible by simulation.[222,223] This allows to bridge the gap between simulation and experiments when addressing the structural features of protein scaffolds in solution. The current chapter explores this link by employing μs-long MD simulations of RD peptides in explicit solvent and compare the results with experimental findings obtained in previous chapters, together with additional insights obtained by NMR spectroscopy.

Simulation of metal-containing systems such as native MPs and RD designs faces issues regarding the realistic treatment of protein-metal interactions. This is because metal ion chemistry is not adequately captured by all-atom force fields, where charge-transfer and ligand-field stabilization effects often play a crucial role in dictating coordination geometries and binding affinities. Again, QM-based treatment of metal systems would be preferred but the small time-scales sampled preclude major scaffold reorganizations to be probed. Along with QM-based corrections of metal first- and second-coordination sphere interactions, “bonded” models are usually employed where the metal-protein bond is treated as a pseudo-covalent one. [224] This greatly limits sampling of ligand-exchange phenomena and usually non-bonded models are adopted instead. However, such models also present challenges since the metal ion is treated as a charged sphere, therefore not taking into account the spatial orientation of electron orbitals involved in ligand coordination.[42] In the case of RD peptides, unsuitability of these approaches is aggravated since folding is mediated by the Zn(II) binding. Ligand exchange phenomena and significant drift from defined TS coordination geometries is therefore expected in such cases. Attempts to overcome these limitations have been developed recently, such as the employment of polarizable force fields that mimic some of the charge-transfer and ligand stabilization effects associated with metal coordination by biomolecules. An example of this is the Drude oscillator model by introducing an auxiliary charged particle attached to each polarizable atom through a harmonic potential.[225,226] This method has been successfully employed in protein folding simulations, but its currently limited to treatment of monovalent charged species and shorter simulation trajectories.

A robust method for simulation of metal-containing system is the Cationic Dummy Atom (CaDa) approach, where a non-bonded description of *e.g.* Zn(II) is made by the inclusion of charged virtual particles that mimic the orientations of unoccupied $4s4p^3$ orbitals of the closed $3d^{10}$ system.[227,228] This approach has been adopted in several simulations of metalloproteins, including native Zn(II) metalloenzymes, and shown to reasonably capture the structural and electrostatic effects involved in metal-protein interactions, including ligand-exchange events.[229–231] Therefore, in the current chapter the suitability of employing the tetrahedral Zn(II) CaDa variant was addressed for the MD simulation studies of native astacin, Sp1f2 peptide and RD designs.

[42] A non-bonded model was used in Chapter 2 for the study of thermolysin. Discussion of method limitations are addressed in annex 1 therein.

## 6.2 Materials and Methods

*Molecular dynamics simulations:* peptides and astacin were simulated in explicit solvent under periodic boundary conditions following the methods described elsewhere and adapted accordingly.[232] The GROMACS 5.1.2 simulation package with GPU acceleration was used.[222,223] The employed force field was AMBER99SB*-ILDN [233–235] modified to include the CaDa approach [228],[43] where the Zn(II) ion is modelled with four dummy atoms (ZND) in a tetrahedron-shaped geometry and coordinating histidines are modelled in double-deprotonated state (HIZ, charge -1). Input structure files corresponded to native Sp1f2 (PDB ID: 1VA2, NMR state 27), HP35 (PDB ID: 1UNC, NMR state 6) and Astacin (PDB ID: 1AST) or the outputted DEs with best $Score_{total}$ for RD peptides (details in Chapter 2). Residue protonation state was attributed according to the respective *pKa* values; Lys and Arg residues were modelled in protonated state, aspartate and glutamate residues in deprotonated state, Zn(II)-coordinating cysteines in deprotonated state (CYM) and non-coordinating histidines in monoprotonated state at the $N\varepsilon^2$ atom. Molecules were placed in a cubic box with edges at least 12 Å from the solute and solvated with the explicit TIP3P water model [236] (~15900 molecules for astacin, ~5500 for Sp1f2, RD01 and RD01v2, ~4900 for HP35 and RD02).[44] Chloride or sodium counter-ions were added first to neutralize the total charge of the system and then added in equal proportion to achieve a 50 mM NaCl concentration (as used in Chapter 4 and 5). Energy minimization was done in two steps to remove eventual atom clashes in crystallographic, NMR or Rosetta outputted files: steepest descent minimization algorithm (max 2000 steps) followed by a conjugated gradient algorithm (max 1000 steps). Short equilibration in a NPT ensemble was done next (Nosé-Hoover thermostat at 300 K with time-constant of 1.6 ps [237,238]; isotropic Parrinello-Rahman barostat at 1 bar with time-constant 5 ps [239,240]) with positional restrains for all hydrogen bonds for 3 steps of 100 ps each with the LINCS algorithm (order parameter 8, iteration level 2), with a force constant of 1000, 100 and 10 kJ/mol in each consecutive step, respectively.[45] Integration step was 2 fs and coordinates were saved each 25 ps, with long-range electrostatics being treated with the Particle-Mesh Ewald algorithm [241]. Production runs were done 1 μs for peptides and 365 ns for astacin in a NPT ensemble. All steps (solvation, energy minimization, equilibration and production) were made twice (sim1 and sim2), with a total simulated time 2 μs for each peptide and 730 ns for astacin.

*Trajectory analysis and clustering:* Trajectory RMSD and Radius of Gyration plots analysed at each 25 ps frames for sim1 and sim2 independently. Cluster analysis of the total trajectories was

[43] AMBER parameters available in http://www.mayo.edu/research/labs/computer-aided-molecular-design/projects/zinc-protein-simulations-using-cationic-dummy-atom-cada-approach were converted for usage in GROMACS by unit conversion.

[44] TIP3P allows for faster calculations than the more complex TIP4P and TIP5P water molecules often employed.

[45] Constraints to all heavy-atom bonds was not made since Zn(II)-dummy atom distances are smaller than other covalent bonds and this causes software errors. Attempts to overcome this were not approached since the long time-scales employed were expected to allow for proper system equilibration.

done from the RMSD matrix of 20007x20007 (1 μs) or 7389x7389 (365 ns) elements, corresponding to equally spaced 50 ps frames from only one replicate. The "gromos" clustering method described by Piana *et al.* was employed, with a minimum of 10 structures per cluster and a cut-off of 3 Å between backbone atoms. [242] Each cluster was represented by the centroid structure, *i.e.* the one with smallest average distance to the remaining neighbours. Analysis of designed AS ($d_{AB}$, $\theta_B$) done at each 250 ps frames from the aggregate of sim1 and sim2 trajectories. Structure representations and DSSP timeline analysis done as described before with VMD 1.9.2.

*$^1$H nuclear magnetic resonance spectroscopy:* Assays were made in $D_2O$ 50mM NaCl at pH ~7.5, adjusted by additions of concentrated NaOH and HCl solutions in a Bruker Avance II+ 800 MHz or Bruker Avance II+ 500 MHz under supervision of Dr. Manolis Matzapetakis (Biomolecular NMR Lab, ITQB/NOVA). Peptide stock solutions were added up to 150 μM (RD01v2) or 1 mM (RD02), followed by stepwise additions of 10.33 mM $ZnCl_2$ ($V_T$= 550 μL). Spectra recorded in the 10 to 60 °C temperature range (5 °C intervals) using TSP 48 μM added to the solution as reference (0 ppm).

## 6.3 Results and Discussion

### 6.3.1 Molecular dynamics simulations

Microsecond-long MD simulations revealed important dynamical features of designed RD peptides and their native counterparts. As shown in Figure 6.1, all peptides and native astacin depart from the initial conformation as expected, a result of the higher conformational space explored during MD trajectories in comparison with NMR and X-ray structures.

In the case of RD peptides, the displacement implies a divergence from the best candidate structure outputted by Rosetta, with higher RMSD values observed for both backbone and side chain atoms in comparison to native peptides. No significant differences were identified in terms of structure length (as given by the related radius of gyration), indicating that no major unfolding events occurred throughout the two replicate trajectories.[46] Structural sites and astacin AS presented less mobility than the designed AS, with clear disruption events apparent in at least one of the RD01 and RD01v2 trajectories.

[46] Global motions of the astacin fold were not anticipated in 1 μs trajectories since these typically occur in longer time-scales. Trajectories were therefore trimmed at 365 ns to properly sample AS side chain reconfigurations in the substrate-free, "open" conformation. Even for short trajectories (20 ns) the sampled conformational space diverges from the crystallographic structure. Further aspects of MP internal dynamics are described in section 2.3.1 and Annex 1 therein.

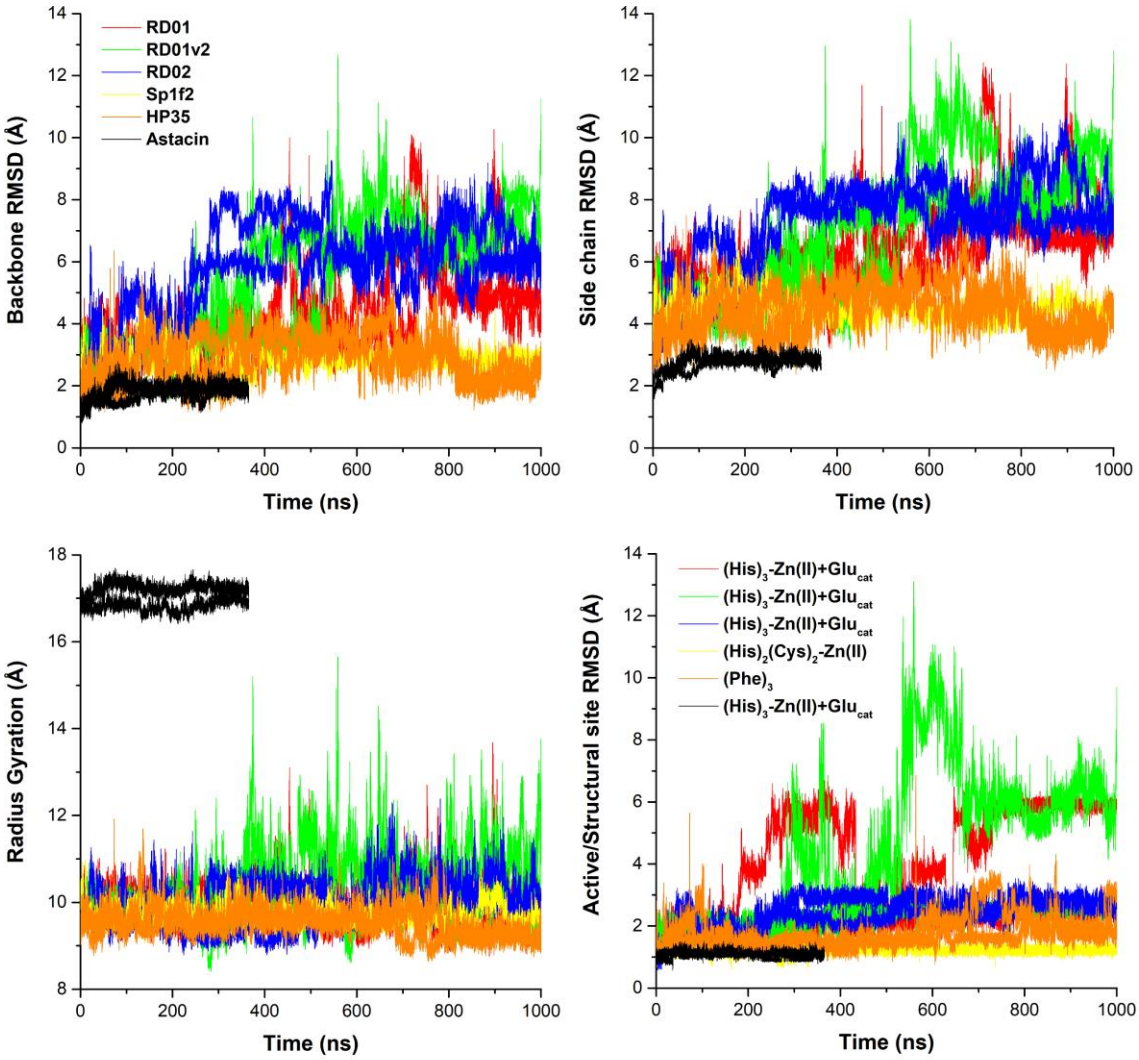



**Figure 6.1 –** General features of MD simulations.
A – RMSD of backbone atoms from each 25 ps trajectory frame in relation to the starting NMR state (SP1f2 and HP35), X-ray crystal structure (astacin) or DE with best $Score_{total}$ (RD01, RD01v2 and RD02). For each system, data from the two replicate trajectories are shown with the same colour coding (details in inset legend). B – Corresponding RMSD of side chain atoms. C- Corresponding radius of gyration. D – RMSD of designed AS residues (RD01, RD01v2, RD02) or native structural (Sp1f2 and HP35) and active (astacin) sites, details in inset legend.

One replicate simulation per peptide scaffold was chosen to probe further details on stability and dynamics, with results of DSSP and cluster analysis shown below and summarized in Table 6.1. DSSP plots allow to track interconversion between secondary structure elements in a given scaffold, which is useful to evaluate qualitatively the local effects of designed sequence changes. [243,244] Clustering based on structural similarity between trajectory frames provides a description of the sampled conformational space, with analysis focused on the set of clusters that populated > 80% of total simulated time.

**Table 6.1 –** Cluster analysis of MD simulations

| Scaffold | RMSD matrix [min, max] | Number clusters (cut-off 3 Å) | # top clusters (> 80% pop. Time) |
|---|---|---|---|
| **Sp1f2** | [0.3, 7.3] | 5 | 1 |
| **RD01** | [0.3, 11.9] | 69 | 8 |
| **RD01v2** | [0.2, 12.7] | 73 | 8 |
| **HP35** | [0.2, 7.9] | 14 | 1 |
| **RD02** | [0.2, 11.6] | 49 | 5 |
| **Astacin** | [0.2, 3.3] | 1 | 1 |

No major changes in the ββα fold of Sp1f2 occurred throughout the trajectory as shown in Figure 6.2, with only one majorly populated cluster corresponding to minor rearrangements at the C-terminal and β-turn (consistent with flexible regions of the NMR structure shown in section 2.3.5, Figure 2.6). With the designed sequence changes made in RD01 these two regions became unstable, which led to partial disruption of the α1 and loss of β1 and β2 secondary structure elements. A higher dispersion of frame pairwise RMSD values and higher number of top populated clusters was obtained, reflecting the increased flexibility of the design.

In RD01v2 similar dynamical features were observed, although with less α-helix disruption towards the C-terminal. However, the introduced M4T and R13T sequence changes led to unexpected reconfiguration of the β2 into mixed turn/helical configurations. This is in sharp contrast with the design objective of increasing β-sheet forming propensity. While this has been successfully achieved in native ZF scaffolds (as discussed in previous chapters), the combination with already destabilizing sequence changes made in RD01 could have led to new interaction networks that favoured helical conformations instead. As a result, the top populated clusters show highly disordered conformations of the original β-sheet in Sp1f2. As discussed in section 4.3.2, both RD01 and RD01v2 peptides present increased helical content in the *holo* form. Indeed, α-helix extension has also been postulated to occur in other Sp1f2 variants lacking one Zn(II)-coordinating residue,[137] which in the case of RD01v2 could have been accentuated due to the introduced threonine residues and the K12V sequence change.

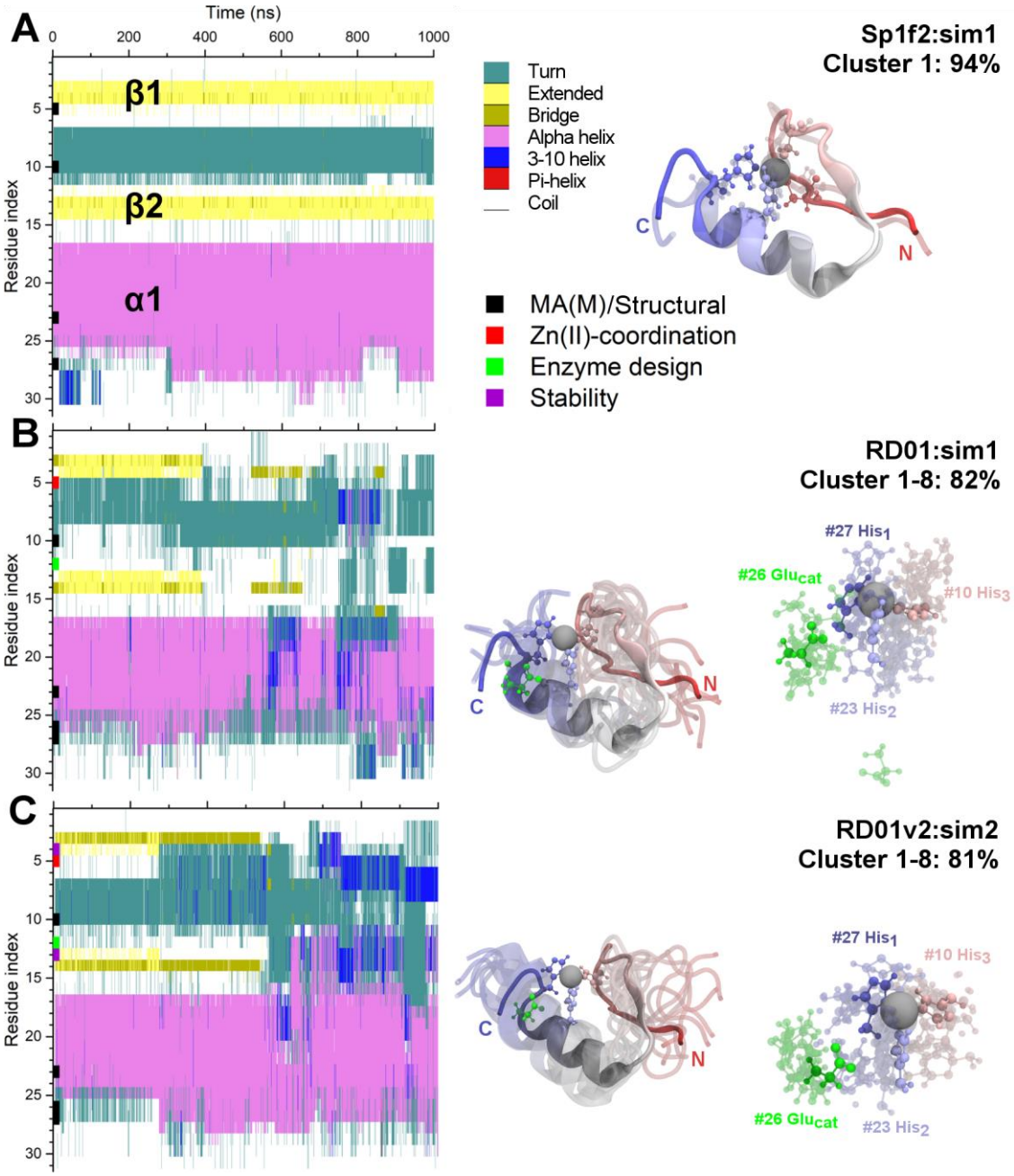



**Figure 6.2 –** Dynamics of Sp1f2, RD01, RD01v2 designs in 1 μs-long MD simulations. A - Timeline analysis of native Sp1f2 secondary structure changes throughout total simulated time. DSSP plots with secondary structure elements identified in bold and backbone conformations in coloured representation (left, legend in inset). Residues of interest labelled by colour-coding (legend in inset). Cluster analysis of simulations (right). Original NMR state (solid) and top populated cluster (transparent) in cartoon representation color-coded by residue index from the N- (red) to the C terminal (blue) and structural site in CPK representation. B and C – corresponding representations for RD01 and RD01v2, respectively. Outputted DE with best $Score_{total}$ (solid) and top populated clusters (transparent). Details of designed AS with Zn(II) coordinating histidines coloured by residue index and catalytic glutamate in green.

For HP35 there were only small changes in the native all-α fold as shown in Figure 6.3, with readjustments occurring mostly at α1 in the top populated cluster (in agreement with the NMR states shown in section 2.3.7, Figure 2.11). On the other hand, RD02 presented major disruptions of α1, helical reconfigurations in α2 and partial disruption of α3 towards the C-terminal, presenting

higher dispersion of pairwise RMSD values and higher number of clusters than the native HP35 structure. The effect of removing the highly conserved phenylalanine residues was clear: the F6H replacement led to disruption of α1 although F10 was kept, F17A led to reconfiguration of α2.[47] In contrast, the L1G and F35S sequence changes located at the N and C termini did not produce obvious fold disruptions or stabilization. The region close to the K24M and N27A sequence changes remained relatively stable throughout the trajectory. The MD simulation results are in correspondence with those reported in section 4.3.2, where RD02 presented slightly less α-helix content than HP35 in folded forms. Also, they are in agreement with findings reported for the HP24stab structure, which lacked α1 but still formed supersecondary structures resembling the native HP35 (chicken) topology.[245] It is therefore acknowledged the small contribution α1 to native fold stability, which in current case of RD02 could not be stabilized although F6H was coordinated to Zn(II).

The MD simulations results can also be compared with the thermal stability of scaffolds described in section 4.3.3: Sp1f2 showed no fold disruption while both RD01 and RD01v2 did. However, HP35 presented more stable features than RD02 although these two peptides had nearly identical thermal stabilities. RD01v2 presented a significant increase in stability compared with RD01, although the peptides behaved similarly in MD simulations. These discrepancies between simulation and experiments may be attributed to the unsuitability of the employed simulation conditions to recapitulate the thermodynamic properties of the modelled systems. More realistic conditions have been successfully employed in the study of HP35 (chicken), with longer simulated times at higher temperatures where fold/unfold conformations are sampled or with replica-exchange MD simulations. [114,115] Nonetheless, decreased thermal stability can be correlated with higher scaffold flexibility, supporting the notion that multiple local minima of the potential energy landscape (corresponding to distinct conformation clusters) are accessible by more flexible scaffolds (RD01, RD01v2, RD02 and HP35) in contrast to the more well-defined global minimum sampled by stable ones (Astacin and Sp1f2).[216] The increased folded content of HP35, RD01, RD01v2 and RD02 under the presence of TFE is also in agreement with their marginal stability associated with flexibility of secondary structure elements. Simulations in mixed water:TFE media would be required for confirmation of this hypothesis.

[47] As referred in section 2.3.7 the F17A sequence change was decided based on preliminary MD simulations. However, these were short, 20 ns trajectories where the Zn(II) ion was treated with the standard non-bonded model (also employed in section 2.3.1 for thermolysin). At these conditions, there is incomplete sampling of the conformational space explored in 1 μs trajectories and therefore the results were not conclusive. Despite this, it was observed that the F17 and F17L sequence variant led to ejection of the metal ion from the $(His)_3$-Zn(II) coordination site during the initial ns of simulation, in contrast to results obtained for the F17A variant (not shown). This was attributed to steric clashing between the phenylalanine and leucine residues and the Q25H sequence change ($His_1$) that led to disruption of AS geometry, which was not apparent in the less bulky F17A variant.

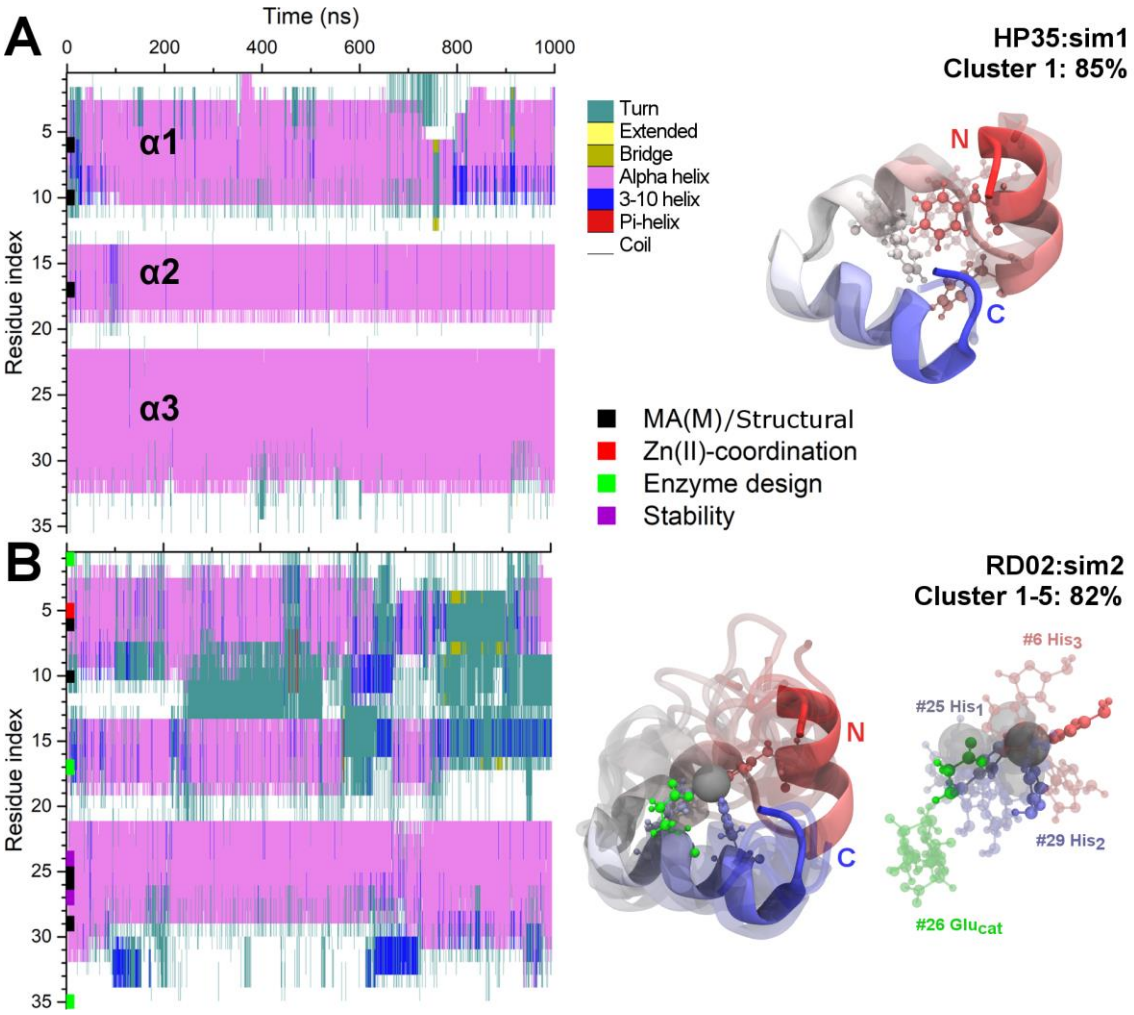



**Figure 6.3 –** Dynamics of HP35 and RD02 designs in 1 μs-long MD simulations.
A - Timeline analysis of native HP35 secondary structure changes throughout total simulated time. DSSP plots with secondary structure elements identified in bold and backbone conformations in coloured representation (left, legend in inset). Residues of interest labelled by colour-coding (legend in inset). Cluster analysis of simulations (right). Original NMR state (solid) and top populated cluster (transparent) in cartoon representation color-coded by residue index from the N- (red) to the C terminal (blue) and structural site in CPK representation. B – corresponding representations for RD02. Outputted DE with best $Score_{total}$ (solid) and top populated clusters (transparent). Details of designed AS with Zn(II) coordinating histidines coloured by residue index and catalytic glutamate in green.

A description of the essential dynamics of the systems by EDA was not approached given the high dimensional subspace of backbone and side chain dihedral angles explored by the top populated clusters.[48] Although a useful method in guiding protein design efforts, essential dynamics is usually performed on native or evolved protein systems where a native-like potential energy surface is typically explored.[232] These are well defined energy minima, where global, large scale motions of protein domains are typically involved in the microsecond to second regime. Despite EDA on shorter trajectories of proteins provide a representative description of global MP

[48] Description of EDA is made in section 2.3.1, Annex 1 therein.

motions, as discussed in Chapter 2, the potential energy landscape topology is expected to be distinct from the ones sampled in the microsecond regime by RD peptides. In respect to this, the higher number of structural clusters found for designs in relation to native folds indicates that the sampled conformational subspace of the former has indeed higher dimensionality than the later. This points to a potential energy surface with multiple local minima in comparison with a more well-defined surface minimum for native scaffolds, which is reflected in major backbone and side chain re-organizations as shown in previous figures. EDA is usually performed on MD simulations assuming non-linear effects (a limitation of PCA-derived methods) and a relatively reduced conformational subspace, therefore being limited to capture secondary structure rearrangements clearly identified by cluster analysis.

Analysis of RD peptides was made taking into consideration that scaffold integrity, although a common feature of active systems, may not be required given increasing examples of highly dynamic or even disordered ones.[192,210] Therefore, a low dimensional description of the catalytically-relevant subspace of AS interactions was issued in light of the findings obtained in previous chapters. The high number of AS side chain conformations observed for all RD peptide clusters prompted a quantitative analysis of geometrical features. The results are shown in Figure 6.4.

Given the absence of diAla in current MD simulations (substrate-free form), a reduced set of geometrical features was selected from the $MA(M)_{AS}$ model developed in section 2.3.3, namely one distance ($d_{AB}$) and one angle ($\theta_B$) between residues and the Zn(II) ion.[49] Inclusion of data derived from analysis of AS from experimental MP-TSA structures proved to be useful in understanding the flexible features of AS residues for both RD peptides and astacin. Catalytic interactions between the $Glu_{cat}$ and Zn(II) were lost for all RD peptides, since the residue moved away from the $MA(M)_{AS}$ defined distances in contrast to astacin, where it kept slightly shorter distances. Side chain fluctuations on the other hand where quite disperse for all systems including astacin, suggesting that AS configurations are only geometrically conserved under the presence of TSA molecules. In astacin the $Glu_{cat}$ is constrained by nearby secondary structure elements which prevents adoption of other side chain conformations, while in RD peptides this residue is solvent-exposed and therefore can freely explore a higher number of conformers even if major backbone rearrangements do not occur in this region.

[49] Dihedral $\chi_B$ values were not included to keep the analysed set to a minimum without losing information on side-chain fluctuations.

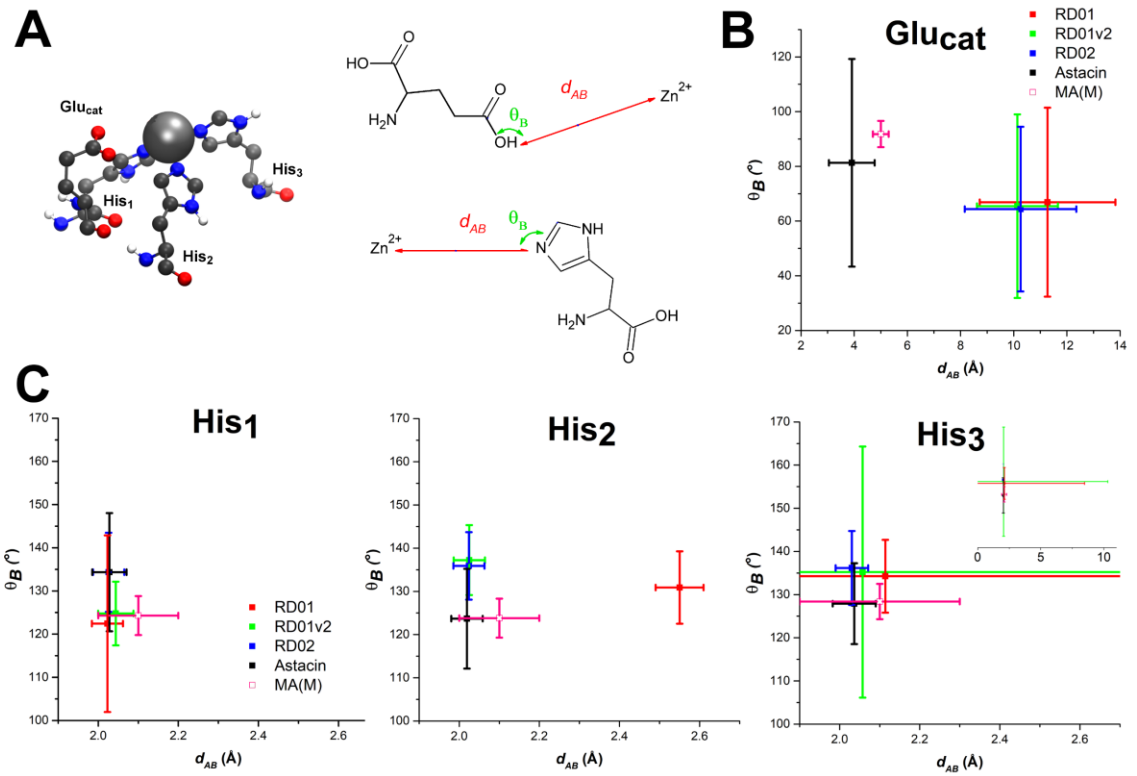



**Figure 6.4 –** Conservation of AS geometrical features of designs by MD simulations. A – Three-dimensional representations of the $MA(M)_{AS}$ model, with Zn(II) coordinating histidines ($His_1$-$His_3$), catalytic glutamate ($Glu_{cat}$) and Zn(II) ion (left). Geometrical parameters defined for catalytic interactions (top) and Zn(II) coordination (bottom), corresponding to one distance ($d_{AB}$) and one angle ($\theta_B$) between residues and the metal ion (right). B – Scatter plot of parameter value distributions for RD peptides and astacin $Glu_{cat}$ interactions (legend in inset). Values correspond to medians and error bars to standard deviations for a total of 2 µs (RD peptides) or 730 ns (astacin) from two replicate simulations. Data from $MA(M)_{AS}$ model included for comparison (details in section 2.3.3). Active site representations for one replicate per scaffold except astacin shown in previous Figure 6.2 and 6.3. C – corresponding scatter plots for Zn(II)-$His_{1-3}$ coordination interactions.

Simulations under the presence of diAla would be required to probe if induce-fit interactions at the AS occur when the substrate is bound. [224,246] Nonetheless, this provides a rationale for the low hydrolytic activity observed for RD peptides in Chapter 5 given that these fail to maintain pre-organization of the AS. Substrate-induced scaffold re-organization is not expected, given the apparent lack of RD fold changes when diAla was added to the peptide-Zn(II) complexes in section 4.3.3.

The $His_{1-3}$-Zn(II) coordinating interactions were more conserved than the catalytic one. In astacin distances kept very close to those of the $MA(M)_{AS}$ for all three histidines, although side chain fluctuations gave rise to a higher dispersion of angles. The geometrical features of $His_1$ were also kept for all RD peptides. However, this was not the case for $His_2$ and $His_3$ residues. In RD01 a drift of the residue to longer distances was observed in contrast to the remaining peptides and astacin. $His_1$ and $His_2$ are located in portions of a α-helices that tend to be preserved for all systems, resulting in less deviations from the $MA(M)_{AS}$ geometry (except for RD01). Conservation of $His_3$ positioning was more critical, given that this residue is located in β2 of RD01 and RD01v2

and in α1 of RD02, regions where major backbone rearrangements occurred. In RD02 the geometrical interactions were nonetheless close to the ones obtained for astacin, while in RD01 and RD01v2 there was a complete drift of the residue away from the metal centre. This again points to the destabilizing interactions promoted by the introduced sequence changes in RD01 and RD01v2 β-sheets, whose backbone rearrangements led to transient breaking of the $His_3$-Zn(II) bond.

These findings relate to some degree with important aspects found during the design stage. In section 2.3.5 and 2.3.7, proper AS geometry was reproduced in only 1/31 NMR states for RD01 and RD01v2, while in RD02 it was in 4/25. Upon dynamical treatment of the systems under MD simulation conditions, AS geometry was lost to a higher degree for RD01 and RD01v2 than for RD02, supporting the argument that the latter was a less restrained designed than the former two. This can also be correlated with ranking of designs in Rosetta, where RD02 presented the most favourable scaffold scores. There is also a correspondence between the dynamical properties of the systems and the experimental $K_{ZnP,app}$ values obtained in section 4.3.1 and 4.3.2. RD01 is the design with lowest affinity for Zn(II), followed by RD01v2 and RD02. The latter retains more native-like interactions with Zn(II) and was correspondingly the design with slightly higher affinity for Zn(II).

In cases where His-Zn(II) were kept, the distances were lower than those of the $MA(M)_{AS}$. This may reflect accommodations of the first-coordination sphere to the presence of TSA molecules or some computational bias introduced by the employed CaDa model. Employment of this model proved to be quite adequate since it could reproduce realistic aspects of first-sphere interactions. For instance, internal rotations of the ZND molecule occurred, where the dummy atom-residue pairs switched throughout the trajectory replicates. In the case of astacin, the $His_3$ switched from coordination through the $N\varepsilon^2$ to the $N\delta^1$ by rotation of the imidazole ring in one of the replicates. The free coordinating position was occupied by bulk water molecules and exchange phenomena were observed in the ns time-scale. Moreover, for systems with sub-micromolar Zn(II) affinities (Sp1f2 and astacin) the $1^{st}$-coordination sphere was kept stable. Second-sphere interactions appeared to have also contributed for the close $Glu_{cat}$-Zn(II) distances observed in astacin. The validation of the employed CaDa approach for the study of microsecond-long peptide systems would require further investigation and therefore the interpretations made in this section should be made with caution, despite being coherent with most of the experimental findings. Since Zn(II) is closed-shell ($d^{10}$ series), different coordination geometries are degenerate and therefore can interchange considerably when simulating microsecond time-scales. It is therefore not guaranteed that the tetrahedral coordination scheme employed would hold for all conformational states explored by the peptides. Simulations with octahedral CaDa variants would be required to probe such aspects.

Although MD simulations and Rosetta modelling are in considerable agreement with general scaffold features, $MA(M)_{AS}$ geometry conservation or favourable catalytic scores do not correlate

with activity towards amino acid-pNA substrates, as discussed for astacin in Chapter 5. A dynamical treatment of substrate-free designs may not necessarily capture the complexity of peptide-substrate interactions seen in chapter 2, and explorations of the potential energy surface of simulated systems under the presence of substrate molecules could be helpful in establishing additional computational-experimental correlates of catalytic activity. This approach has been considered but was not pursued for the moment. Instead, an attempt to further validate the observation made here was attempted by experimental characterization of peptide structures in the following section.

### 6.3.2 $^{1}$H nuclear magnetic resonance

The dynamics of RD01v2 and RD02 peptides was addressed experimentally in the current section by $^{1}$H-NMR spectroscopy. This aimed at validation of the findings obtained by MD simulations and also on the rationalization of results discussed in previous chapters. However, structural elucidation of the peptide-Zn(II) complexes of RD01v2 and RD02 was not possible due to unanticipated problems related with signal duplication and broadening. This was attributed to both nuclear relaxation phenomena and the Zn(II) binding/release processes occurring in the millisecond regime, leading to overlap of signals originating from the two *apo* and *holo* states (or additional intermediate/transient states).[50] General features of the systems were therefore considered only, with focus on specific probes of complex formation under conditions similar to those characterized in Chapter 4.

In the case of RD01v2, as shown in Figure 6.5 upon Zn(II) addition there was the appearance of a new peak in the aliphatic region at 1.05 ppm, which corresponds to methyl side chain groups of L20 residue. The data could be fitted to a monomer model used in Chapter 4. The $K_{Zn,app}$ of 1.41x $10^5$ $M^{-1}$ obtained is in range of values obtained therein, indicating folding of the peptide and formation of α1. Signal broadening was quite evident in the NH region, suggesting major backbone readjustments upon metal binding. The W7 signal at 10.0 ppm presents broadening and splitting upon complex formation, suggesting that the β1/β-turn interface where the residue is located exists in multiple states, consistent with the flexibility of this region found in previous section. Temperature effects were monitored for this signal and show the transition from a single state only at low temperatures where scaffold flexibility is restrained. New signals associated with residues in β-sheet conformation are observed for higher temperatures in the Hα region concomitant with new signals in the aliphatic region, pointing to major fold readjustments close to the determined $T_m$ values in section 4.3.3.

[50] Considering $K_{ZnP}$ values of $10^5$ $M^{-1}$ and an upper limit $K_{ON}$=$10^9$ $s^{-1}$ for zinc binding (diffusion limited), the $K_{OFF}$ process of zinc release are estimated to occur in the $K_{Zn}$=$K_{ON}$/$K_{OFF}$ ⇔ $K_{OFF}$=$10^9$/$10^5$=$10^4$ $s^{-1}$ time-scale.

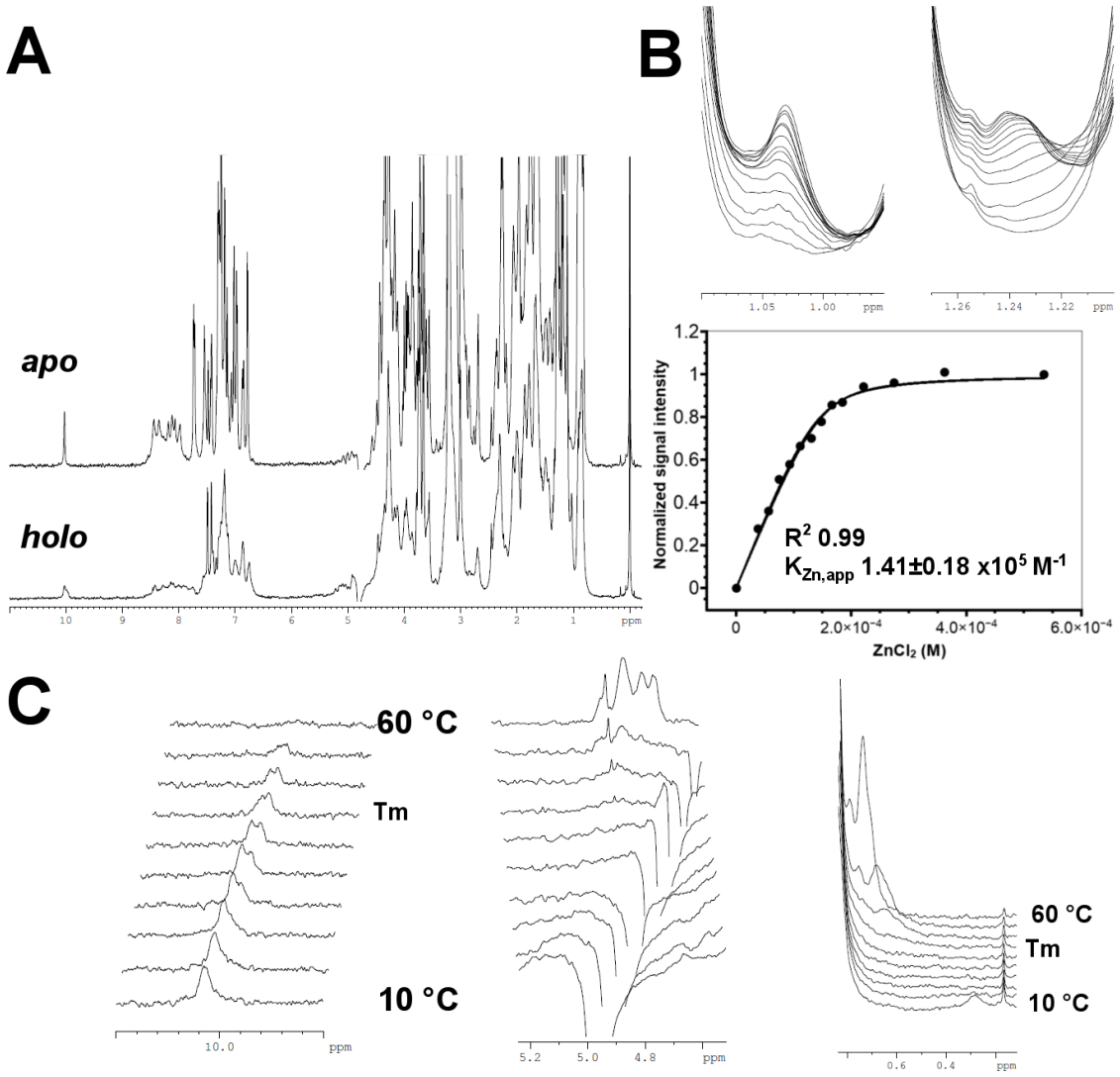



**Figure 6.5 –** $^{1}$H-NMR of 150 µM RD01v2-Zn(II) complex in 50 mM NaCl, pH 7.5. A – overlap of spectra obtained in the before (*apo*) and after (*holo*) addition of 0-534 µM of $ZnCl_2$ at 25 °C. B – Details of aliphatic region signal changes upon metal additions (top). Data from signal at 1.05 ppm (bottom) used for fitting to the 1:1 complex formation model (solid line, details of model derivation in Chapter 4). C – temperature effects in 10 to 60 °C range for W7 (left), Hα region (centre) and aliphatic region (right).

For RD02, W23 signal changes were monitored upon Zn(II) addition and results shown in Figure 6.6. A characteristic sharp signal at 10.25 ppm was observed corresponding to the *apo* form of the peptide in solution (denoted A). At sub-equimolar Zn(II) additions this signal is shifted and broadened, corresponding to either a transient species or a dimer (denoted B). Dimer formation was not considered in previous chapters since the data obtained for micromolar peptide and Zn(II) concentrations could be well fitted to a monomer model.[51] However, at mM concentrations at which the NMR titration were made its formation could not be ruled out. Concomitant with

[51] Data from Zi competition assays from Chapter 4 was fitted to a dimer model ($R^2$ 0.97) but quality was lower than the one obtained for the monomer model ($R^2$ 0.99).

the increase of the signal B was the appearance of a new broad signal (denoted C) which did not reach full intensity up to 1.75 Zn(II) molar excess. This signal C should correspond to W23 in the *holo* form of the peptide, which despite being relatively constrained upon α3 formation, it can sample multiple side chain conformations in MD simulation trajectories.

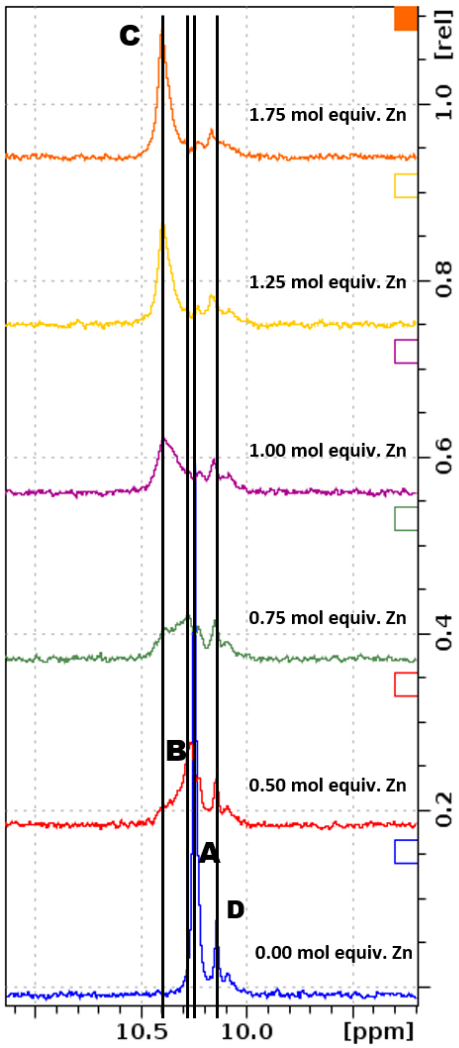



**Figure 6.6 –** $^{1}$H-NMR of 1 mM RD02-Zn(II) complex in 50 mM NaCl at 25 °C, pH 7.5. Overlap of W23 signals obtained after 0-1.75 mM additions of $ZnCl_2$. (A) corresponds to the *apo* form of the peptide, (B) to a transient species at sub-equimolar Zn(II) concentrations, (C) to the metal-complex form and (D) to a small impurity or negligible *apo* state.

Further peak assignment would be useful to probe the His-Zn(II) interactions and relative mobility of the $Glu_{cat}$ residue. However, this would require extensive analysis of the data and combination with other homonuclear and/or heteronuclear techniques for their identification. Structural elucidation of inactive designs has been a common issue in enzyme design projects, which limits the understanding on the molecular basis for catalytic proficiency. In most cases the designs do not fold properly and exhibit either misconfigurations of designed sequence changes or enhanced fold flexibility and/or reduced thermodynamic stability. In the case of Sp1f2, structural elucidation

was achieved since tight binding of the metal ion reduces interference from exchange phenomena and the peptide assumes a well-defined fold topology amenable for signal assignment by heteronuclear $^{1}H/^{15}N$-NMR methods. In the case of HP35, the peptide assumes also a well-defined fold and its structural elucidation was possible through bi-dimensional $^{1}H$-NMR methods.

## 6.4 Conclusion

Structural characterization of RD-Zn(II) complexes was approached in the current chapter. Following secondary-structure characterization of the complexes in Chapter 4, attempts to elucidate their tertiary structures was approached by molecular dynamics simulations and nuclear magnetic resonance spectroscopy. Structural features of substrate-free peptides in the microsecond time scales probed by simulation were in reasonable agreement with experimental findings, pointing to the presence of mobile secondary structure elements in all RD peptides. This was attributed to the introduced sequence changes which led to reconfiguration of residue interaction networks and consequent drift from native fold topologies. Conservation of active site geometrical features designed in Chapter 2 was also addressed, with RD01 and RD01v2 presenting considerable disruption of Zn(II) coordination interactions and RD02 presenting more native-like features, similar to those found in simulations of astacin. Disruption of catalytic interactions was also observed, thus providing a rational for the low hydrolytic activities of peptides described in Chapter 5. Attempts to elucidate experimentally the tertiary structure by nuclear magnetic resonance spectroscopy of the designs failed due to incompatibility of physicochemical/dynamical properties and probed nuclear relaxation phenomena.

While RD01 and RD01v2 represent cases of considerable native scaffold destabilization, RD02 retained more native-like features in line with scaffold rankings at the computational design stage. Molecular dynamics simulations in longer time-scales under explicit solvent conditions thus proved to be a useful tool to further explore the flexible nature of peptide designs, which are not accessible by static treatments employed in the Rosetta. The application of the cationic dummy atom approach for Zn(II) treatment was instrumental to probe conservation of active site geometrical features. The integration of long simulations trajectories in the design process stems as an ideal bridge between modelling and experiments particularly suited to address stability and optimization of tested RD designs and eventually in evaluation of new RD candidates.

# Final Conclusions

The development of metalloprotease activity in alternative scaffolds suitable for bioengineering applications is an interesting and promising area in biotechnology, and it has been the main subject of this thesis. The analysis of sequence-structure-dynamics relationships between metalloproteases identified conserved first and second coordination sphere interactions with the metal ion at the active sites. These observations were used to screen with the Rosetta enzyme design software a set of 43 peptides and small protein scaffolds (20-64 residues) for accommodation of a general active site model derived from the MA(M) subclan of metalloproteases. The use of NMR structures allowed the inherent flexibility of small scaffolds to be also accounted for. One of such scaffolds, the *zinc finger 2 of human Sp1 transcription factor* – Sp1f2 was computationally redesigned in two rounds, RD01 and RD01v2 scaffolds, the latter being guided by experimental results in order to include sequence modifications for increased scaffold stability. The multivariate analysis of Rosetta scoring parameters identified the best candidate with native-like features from the remaining 42 scaffolds, corresponding to the *human villin headpiece C-terminal subdomain*, HP35. Its sequence was extensively redesigned into the RD02 scaffold to develop affinity for binding the metal ion, to accommodate the metalloprotease active site model and to increase scaffold stability (Chapter 2).

After production of the RD peptides through chemical synthesis (Chapter 3), the physicochemical properties of the corresponding peptide-Zn(II) complexes were addressed (Chapter 4). The RD01 design presented a fold similar to the native metallopeptide when coordinated to Zn(II) through the three histidine residues of the active site model, although with reduced thermal stability. Sequence modifications introduced in the second round of design resulted in stabilization of the corresponding RD01v2-Zn(II) complex despite deviations from native fold topology and stability remained unchanged. The RD02 design adopted a fold similar to the native scaffold upon coordination to the Zn(II) metal through the introduced histidine residues. Zn(II) binding affinities and thermal stability of the RD02-Zn(II) complex were similar to previous designs (affinity constants in the $10^5$ $M^{-1}$ range, melting temperatures between 37-50 °C), despite the sequence modifications made specifically to address these issues. Folding was dependent on metal coordination in all RD designs. The computational approach thus proved to be successful in the redesign of structural metal sites or in the *de novo* design of Zn(II) binding sites in small scaffolds.

The designed metallopeptides acted as modest catalysts of ester hydrolysis but failed to present target metalloprotease activity towards the modelled diAla peptide substrate (Chapter 5). RD01 and RD01v2 designs presented hydrolytic activities in range with other designs of the native scaffold where only fist coordination sphere modifications were done. Therefore, the second coordination sphere interactions included in the active site model did not result in increased catalytic

proficiency of the designed scaffolds. On the other hand, the hydrolytic activity of RD02 is the result of successful design of a catalytic metal site into the HP35 scaffold, although within the range of RD01 and RD01v2 designs. The catalytic rate enhancements of ester hydrolysis obtained for RD metallopeptides ($k_2$ values in the order of $10^{-1}$ $M^{-1}s^{-1}$) are within range of other small metal-dependent designs, although being 2 to 4 orders of magnitude below those of other designs with more complex folds and native metalloenzymes, respectively. This points to possible limitations in developing efficient biocatalysts based on scaffolds with reduced size and minimal fold topology.

Structural characterization of the RD metallopeptides was attempted to establish correlations between structural features of the designs and the observed stability and catalytic activities (Chapter 6). The dynamics of scaffolds in solution were probed by simulation, revealing high backbone flexibility and partial disruption/mobility of secondary structure elements. As a result, the active site residue positions drifted away from the idealized geometries of the corresponding model developed in Chapter 2. The disruption of the Zn(II) first coordination sphere was more pronounced in RD01 and RD01v2 designs than in the RD02 design, which could be correlated with the metal-induced folding and lower thermal stabilities of the RD01 scaffolds (Chapter 4), and also to the design features described in Chapter 2. Second coordination sphere interactions were also disrupted to a similar extent in all the RD metallopeptides, which provided a rational for the low catalytic activities described in Chapter 5. The catalytic glutamate residue is not preorganized for transition-state stabilization in any of the RD peptides given its high solvent-exposure and the lack of stabilizing interactions. This contrasts with the tight structural conservation found in native metalloproteases in both the substrate bound/unbound forms. Nuclear magnetic resonance spectroscopy results further revealed the dynamical features of the scaffolds.

Given that target functionality could not be achieved, immobilization of the RD peptides on solid support was not approached. Nonetheless, the insights obtained from both simulation and experiments provide valuable clues to improve further designs. A dynamical and explicit treatment of the peptide, metal and solvent interactions in microsecond time-scales proved to be useful in the identification of structural design flaws and native scaffold limitations. Its implementation in the final stages of the computational design can therefore be of great help to filter out candidates with unstable scaffolds or mechanistically-irrelevant active site geometries.

As an outlook, additional analysis of the RD metallopeptides developed in this work can also be envisioned to address the chemical and structural determinants of the low catalytic activities observed. The role played by the Zn(II) metal ion and designed glutamate residue can be probed with additional catalytic studies (e.g. at variable pH values) and compared with more detailed structural characterizations. The relatively high number of sequence modifications introduced in the 31 to 35 residue-long RD scaffolds rendered metallopeptides too flexible and unstable to hold the active site preorganization and attain target functionality. Additional sequence modifications

could be explored for increasing activity/stability, or alternatively the set of selected peptides/small-proteins could be re-screened to find candidates with suitable catalytic and dynamic properties to act as biocatalysts.

Finally, the work pipeline developed in this thesis could be readily adapted for other enzyme design projects where large sets of protein structures need to be screened and a description of the target catalytic mechanism is available. Overall, this project contributed to further improve computational and experimental approaches for screening the potential of alternative scaffolds as enzymes for bioengineering applications.

# References


1. Polaina J, MacCabe AP. Industrial Enzymes. Polaina J, MacCabe AP, editors. Industrial Enzymes: Structure, Function and Applications. Dordrecht: Springer Netherlands; 2007.
2. Liese A, Seelbach K, Wandrey C, editors. Industrial Biotransformations. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA; 2006.
3. Lu Y, Yeung N, Sieracki N, Marshall NM. Design of functional metalloproteins. Nature. 2009;460: 855–862.
4. Lin YW. Rational design of metalloenzymes: From single to multiple active sites. Coord Chem Rev. Elsevier B.V.; 2017;336: 1–27.
5. Nanda V, Koder RL. Designing artificial enzymes by intuition and computation. Nat Chem. 2010;2: 15–24.
6. Golynskiy M V., Seelig B. De novo enzymes: from computational design to mRNA display. Trends Biotechnol. Elsevier Ltd; 2010;28: 340–345.
7. Kiss G, Celebi-Ölçüm N, Moretti R, Baker D, Houk KN. Computational enzyme design. Angew Chem Int Ed Engl. 2013;52: 5700–25.
8. Zanghellini A. De novo computational enzyme design. Curr Opin Biotechnol. Elsevier Ltd; 2014;29: 132–138.
9. Muñoz Robles V, Ortega-Carrasco E, Alonso-Cotchico L, Rodriguez-Guerra J, Lledós A, Maréchal JD. Toward the computational design of artificial metalloenzymes: From protein-ligand docking to multiscale approaches. ACS Catal. 2015;5: 2469–2480.
10. Heinisch T, Pellizzoni M, Dürrenberger M, Tinberg CE, Köhler V, Klehr J, et al. Improving the catalytic performance of an artificial metalloenzyme by computational design. J Am Chem Soc. 2015;137: 10414–10419.
11. Nanda V, Senn S, Pike DH, Rodriguez-Granillo A, Hansen WA, Khare SD, et al. Structural principles for computational and de novo design of 4Fe–4S metalloproteins. Biochim Biophys Acta - Bioenerg. Elsevier B.V.; 2016;1857: 531–538.
12. Yu F, Cangelosi VM, Zastrow ML, Tegoni M, Plegaria JS, Tebo AG, et al. Protein design: toward functional metalloenzymes. Chem Rev. 2014;114: 3495–578.
13. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487: 545–74.
14. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, et al. Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. 2013. pp. 109–143.
15. O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, Dimaio F, et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput. 2015;11: 609–622.
16. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput. 2017;13: 3031–3048.
17. Hellinga HW, Richards FM. Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. J Mol Biol. 1991;222: 763–785.
18. Bolon DN, Mayo SL. Enzyme-like proteins by computational design. Proc Natl Acad Sci. The National Academy of Sciences; 2001;98: 14274–14279.
19. Gainza P, Roberts KE, Georgiev I, Lilien RH, Keedy DA, Chen C, et al. osprey. Methods in enzymology. 2013. pp. 87–107.
20. Pierce NA, Winfree E. Protein Design is NP-hard. Protein Eng Des Sel. 2002;15: 779–782.
21. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, et al. De Novo Computational Design of Retro-Aldol Enzymes. Science. 2008;319: 1387–1391.
22. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St Clair JL, et al. Computational

design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science. 2010;329: 309–13.
23. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, et al. Kemp elimination catalysts by computational enzyme design. Nature. 2008;453: 190–195.
24. Khersonsky O, Röthlisberger D, Wollacott AM, Murphy P, Dym O, Albeck S, et al. Optimization of the In-Silico-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution. J Mol Biol. 2011;407: 391–412.
25. Rawlings ND, Barrett AJ, Finn R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2016;44: D343–D350.
26. Rawlings ND, Barrett AJ. Introduction. Handbook of Proteolytic Enzymes. 3rd ed. Elsevier; 2013. pp. 325–370.
27. Messerschmidt A, Huber R, Poulos T, Wieghardt K. Handbook of Metalloproteins. Messerschmidt A, Huber R, Poulas T, Wieghardt K, Cygler M, Bode W, editors. Chichester: John Wiley & Sons, Ltd; 2006.
28. MORIHARA K. PRODUCTION OF ELASTASE AND PROTEINASE BY PSEUDOMONAS AERUGINOSA. J Bacteriol. 1964;88: 745–57.
29. Smith H. Discovery of the anthrax toxin: the beginning of studies of virulence determinants regulated in vivo. Int J Med Microbiol. Urban & Fischer; 2001;291: 411–417.
30. Fukushima J, Takeuchi H, Tanaka E, Hamajima K, Sato Y, Kawamoto S, et al. Molecular cloning and partial DNA sequencing of the collagenase gene of Vibrio alginolyticus. Microbiol Immunol. 1990;34: 977–84.
31. Bond MD, Van Wart HE. Characterization of the individual collagenases from Clostridium histolyticum. Biochemistry. 1984;23: 3085–3091.
32. Kessler E, Safrin M, Olson JC, Ohman DE. Secreted LasA of Pseudomonas aeruginosa is a staphylolytic protease. J Biol Chem. 1993;268: 7503–8.
33. Zhou M-Y, Chen X-L, Zhao H-L, Dang H-Y, Luan X-W, Zhang X-Y, et al. Diversity of Both the Cultivable Protease-Producing Bacteria and Their Extracellular Proteases in the Sediments of the South China Sea. Microb Ecol. 2009;58: 582–590.
34. López-Otín C, Overall CM. Protease degradomics: A new challenge for proteomics. Nat Rev Mol Cell Biol. 2002;3: 509–519.
35. Grasso G, Bonnet S. Metal complexes and metalloproteases: targeting conformational diseases. Metallomics. 2014;6: 1346.
36. Siegbahn PEM, Himo F. The quantum chemical cluster approach for modeling enzyme reactions. Wiley Interdiscip Rev Comput Mol Sci. 2011;1: 323–336.
37. Senn HM, Thiel W. QM/MM Methods for Biomolecular Systems. Angew Chemie Int Ed. 2009;48: 1198–1229.
38. Blomberg MRA, Borowski T, Himo F, Liao R, Siegbahn PEM. Quantum Chemical Studies of Mechanisms for Metalloenzymes. Chem Rev. 2014;114: 3601–3658.
39. Amata O, Marino T. Human insulin-degrading enzyme working mechanism. J Am Chem Soc. 2009;131: 14804–14811.
40. Blumberger J, Lamoureux G, Klein ML. Peptide Hydrolysis in Thermolysin: Ab Initio QM/MM Investigation of the Glu143-Assisted Water Addition Mechanism. J Chem Theory Comput. 2007;3: 1837–1850.
41. Chen S-L, Li Z-S, Fang W-H. Theoretical investigation of astacin proteolysis. J Inorg Biochem. Elsevier Inc.; 2012;111: 70–79.
42. Szeto MWY, Mujika JI, Zurek J, Mulholland AJ, Harvey JN. QM/MM study on the mechanism of peptide hydrolysis by carboxypeptidase A. J Mol Struct THEOCHEM. Elsevier B.V.; 2009;898: 106–114.
43. Pelmenschikov V, Blomberg MR a, Siegbahn PEM. A theoretical study of the mechanism for peptide hydrolysis by thermolysin. J Biol Inorg Chem. 2002;7: 284–98.
44. Bora RP, Barman A, Zhu X, Ozbil M, Prabhakar R. Which one among aspartyl protease, metallopeptidase, and artificial metallopeptidase is the most efficient catalyst in peptide hydrolysis? J Phys Chem B. 2010;114: 10860–75.
45. Díaz N, Suárez D. Peptide hydrolysis catalyzed by matrix metalloproteinase 2: a computational study. J Phys Chem B. 2008;112: 8412–24.
46. Vasilevskaya T, Khrenova MG, Nemukhin A V., Thiel W. Methodological aspects of QM/MM calculations: A case study on matrix metalloproteinase-2. J Comput Chem. 2016; 1801–1809.

47. Navrátil V, Klusák V, Rulíšek L. Theoretical Aspects of Hydrolysis of Peptide Bonds by Zinc Metalloenzymes. Chem - A Eur J. 2013;19: 16634–16645.
48. Yiallouros I, Große Berkhoff E, Stöcker W. The roles of Glu93 and Tyr149 in astacin-like zinc peptidases. FEBS Lett. 2000;484: 224–228.
49. Matthews B. Structural basis of the action of thermolysin and related zinc peptidases. Acc Chem Res. 1988; 333–340.
50. Pelmenschikov V, Siegbahn PEM. Catalytic Mechanism of Matrix Metalloproteinases: Two-Layered ONIOM Study. Inorg Chem. 2002;41: 5659–5666.
51. Gomis-Rüth FX. Structural aspects of the metzincin clan of metalloendopeptidases. Mol Biotechnol. 2003;24: 157–202.
52. Smith CR, Smith GK, Yang Z, Xu D, Guo H. Quantum mechanical/molecular mechanical study of anthrax lethal factor catalysis. Theor Chem Acc. 2011;128: 83–90.
53. Brás NF, Fernandes PA, Ramos MJ. QM/MM Study and MD Simulations on the Hypertension Regulator Angiotensin-Converting Enzyme. ACS Catal. 2014;4: 2587–2597.
54. Adekoya OA, Sylte I. The thermolysin family (M4) of enzymes: Therapeutic and biotechnological potential. Chem Biol Drug Des. 2009;73: 7–16.
55. Wu J-W, Chen X-L. Extracellular metalloproteases from bacteria. Appl Microbiol Biotechnol. 2011;92: 253–262.
56. Priest F. Enzymes extracellular. In: Lederberg J, editor. Encyclopedia of microbiology. Academic Press; 1992. pp. 451–460.
57. Patil U, Chaudhari A. Purification and characterization of solvent-tolerant, thermostable, alkaline metalloprotease from alkalophilic *Pseudomonas aeruginosa* MTCC 7926. J Chem Technol Biotechnol. John Wiley & Sons, Ltd.; 2009;84: 1255–1262.
58. Qian Z-J, Jung W-K, Kim S-K. Free radical scavenging activity of a novel antioxidative peptide purified from hydrolysate of bullfrog skin, Rana catesbeiana Shaw. Bioresour Technol. 2008;99: 1690–1698.
59. Song L, Li T, Yu R, Yan C, Ren S, Zhao Y. Antioxidant Activities of Hydrolysates of Arca Subcrenata Prepared with Three Proteases. Mar Drugs. 2008;6: 607–619.
60. Shen S, Chahal B, Majumder K, You S-J, Wu J. Identification of Novel Antioxidative Peptides Derived from a Thermolytic Hydrolysate of Ovotransferrin by LC-MS/MS. J Agric Food Chem. 2010;58: 7664–7672.
61. Fontana A, de Laureto PP, Spolaore B, Frare E, Picotti P, Zambonin M. Probing protein structure by limited proteolysis. Acta Biochim Pol. 2004;51: 299–321.
62. Gao X, Bain K, Bonanno JB, Buchanan M, Henderson D, Lorimer D, et al. High-throughput Limited Proteolysis/Mass Spectrometry for Protein Domain Elucidation. J Struct Funct Genomics. 2005;6: 129–134.
63. Suggs W, Van Wart H, Sharefkin JB. Enzymatic harvesting of adult human saphenous vein endothelial cells: Use of a chemically defined combination of two purified enzymes to attain viable cell yields equal to those attained by crude bacterial collagenase preparations. J Vasc Surg. Elsevier; 1992;15: 205–213.
64. Durham DR, Fortney DZ, Nanney LB. Preliminary evaluation of vibriolysin, a novel proteolytic enzyme composition suitable for the debridement of burn wound eschar. J Burn Care Rehabil. 1993;14: 544–51.
65. Lousa D, Baptista AM, Soares CM. Analyzing the Molecular Basis of Enzyme Stability in Ethanol/Water Mixtures Using Molecular Dynamics Simulations. J Chem Inf Model. 2012;52: 465–473.
66. Kim J, Grate JW, Wang P. Nanostructures for enzyme stabilization. Chem Eng Sci. 2006;61: 1017–1026.
67. Küchler A, Yoshimoto M, Luginbühl S, Mavelli F, Walde P. Enzymatic reactions in confined environments. Nat Nanotechnol. Nature Publishing Group; 2016;11: 409–420.
68. Santana SDF, Pina AS, Roque ACA. Immobilization of enterokinase on magnetic supports for the cleavage of fusion proteins. J Biotechnol. Elsevier B.V.; 2012;161: 378–382.
69. Wang F, Xiao J, Pan L, Yang M, Zhang G, Jin S, et al. A Systematic Survey of Mini-Proteins in Bacteria and Archaea. Bongard J, editor. PLoS One. 2008;3: e4027.
70. Hobbs EC, Fontaine F, Yin X, Storz G. An expanding universe of small proteins. Curr Opin Microbiol. Elsevier Ltd; 2011;14: 167–173.
71. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. 2014;15: 193–204.

72. Couso J-P, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol. Nature Publishing Group; 2017;
73. Tiessen A, Pérez-Rodríguez P, Delaye-Arredondo LJ. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. BMC Res Notes. 2012;5: 85.
74. Chen LH, Kenyon GL, Curtin F, Harayama S, Bembenek ME, Hajipour G, et al. 4-Oxalocrotonate tautomerase, an enzyme composed of 62 amino acid residues per monomer. J Biol Chem. 1992;267: 17716–21.
75. Whitman CP. The 4-oxalocrotonate tautomerase family of enzymes: how nature makes new enzymes using a β–α–β structural motif. Arch Biochem Biophys. 2002;402: 1–13.
76. Zhang X, Houk KN. Why enzymes are proficient catalysts: beyond the Pauling paradigm. Acc Chem Res. 2005;38: 379–85.
77. Garcia-Viloca M, Gao J, Karplus M, Truhlar DG. How enzymes work: analysis by modern rate theory and computer simulations. Science. 2004;303: 186–95.
78. Tantillo DJ, Houk N. Theozymes and compuzymes : biological catalysis theoretical models for biological catalysis. Curr Opin Chem Biol. 1998; 743–750.
79. Zhang X, DeChancie J, Gunaydin H, Chowdry AB, Clemente FR, Smith, et al. Quantum Mechanical Design of Enzyme Active Sites. J Org Chem. 2008;73: 889–899.
80. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–42.
81. Potestio R, Aleksiev T, Pontiggia F, Cozzini S, Micheletti C. ALADYN: a web server for aligning proteins by matching their large-scale motion. Nucleic Acids Res. 2010;38: W41-5.
82. Bakan A, Meireles LM, Bahar I. ProDy: protein dynamics inferred from theory and experiments. Bioinformatics. 2011;27: 1575–7.
83. Tronrud DE, Monzingo a F, Matthews BW. Crystallographic structural analysis of phosphoramidates as inhibitors and transition-state analogs of thermolysin. Eur J Biochem. 1986;157: 261–8.
84. Englert L, Silber K, Steuber H, Brass S, Over B, Gerber H-D, et al. Fragment-based lead discovery: screening and optimizing fragments for thermolysin inhibition. ChemMedChem. 2010;5: 930–40.
85. Guffy SL, Der BS, Kuhlman B. Probing the minimal determinants of zinc binding with computational protein design. Protein Eng Des Sel. 2016;29: 327–338.
86. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, et al. New algorithms and an in silico benchmark for computational enzyme design. Protein Sci. 2006;15: 2785–2794.
87. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D. De Novo Enzyme Design Using Rosetta3. Uversky VN, editor. PLoS One. 2011;6: e19230.
88. Oka S, Shiraishi Y, Yoshida T, Ohkubo T, Sugiura Y, Kobayashi Y. NMR structure of transcription factor Sp1 DNA binding domain. Biochemistry. 2004;43: 16027–35.
89. Nomura A, Sugiura Y. Hydrolytic reaction by zinc finger mutant peptides: successful redesign of structural zinc sites into catalytic zinc sites. Inorg Chem. 2004;43: 1708–13.
90. Besold AN, Widger LR, Namuswe F, Michalek JL, Michel SLJ, Goldberg DP. Revisiting and re-engineering the classical zinc finger peptide: consensus peptide-1 (CP-1). Mol BioSyst. Royal Society of Chemistry; 2016;12: 1183–1193.
91. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014;42: D304-9.
92. Polticelli F. Structural determinants of mini-protein stability. Biochem Mol Biol Educ. 2001;29: 16–20.
93. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23: 2947–2948.
94. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, et al. A new bioinformatics analysis tools framework at EMBL–EBI. Nucleic Acids Res. 2010;38: W695–W699.
95. Shannon P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003;13: 2498–2504.
96. Schrödinger L. The PyMOL Molecular Graphics System, Version1.3r1. 2010.

97. Humphrey W, Dalke a, Schulten K. VMD: visual molecular dynamics. J Mol Graph. 1996;14: 33–8, 27–8.
98. Stone J. An efficient library for parallel ray tracing and animation. Masters Theses. University of Missouri--Rolla. 1998.
99. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminform. 2011;3: 33.
100. Crooks GE. WebLogo: A Sequence Logo Generator. Genome Res. 2004;14: 1188–1190.
101. Sousa SF, Fernandes PA, Ramos MJ. The Carboxylate Shift in Zinc Enzymes: A Computational Study. J Am Chem Soc. 2007;129: 1378–1385.
102. Khare SD, Kipnis Y, Greisen P, Takeuchi R, Ashani Y, Goldsmith M, et al. Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. Nat Chem Biol. Nature Publishing Group; 2012;8: 294–300.
103. Struthers MD, Cheng RP, Imperiali B. Design of a monomeric 23-residue polypeptide with defined tertiary structure. Science. 1996;271: 342–5.
104. Struthers M, Ottesen JJ, Imperiali B. Design and NMR analyses of compact, independently folded BBA motifs. Fold Des. 1998;
105. Struthers MD, Cheng RP, Imperiali B. Economy in Protein Design: Evolution of a Metal-Independent ββα Motif Based on the Zinc Finger Domains. J Am Chem Soc. 1996;118: 3073–3081.
106. Jolliffe I. Introduction. Principal Component Analysis. New York: Springer-Verlag; 2002. pp. 2–9.
107. Jolliffe IT. Choosing a Subset of Principal Components or Variables. Princ Compon Anal. New York: Springer-Verlag; 2002; 112–132.
108. Jolliffe IT. Principal Components as a Small Number of Interpretable Variables: Some Examples. Principal Component Analysis. New York: Springer-Verlag; 2002. pp. 63–77.
109. Vermeulen W, Vanhaesebrouck P, Van Troys M, Verschueren M, Fant F, Goethals M, et al. Solution structures of the C-terminal headpiece subdomains of human villin and advillin, evaluation of headpiece F-actin-binding requirements. Protein Sci. 2004;13: 1276–1287.
110. Lei H, Duan Y. Two-stage Folding of HP-35 from Ab Initio Simulations. J Mol Biol. 2007;370: 196–206.
111. Bi Y, Cho J-H, Kim E-Y, Shan B, Schindelin H, Raleigh DP. Rational Design, Structural and Thermodynamic Characterization of a Hyperstable Variant of the Villin Headpiece Helical Subdomain. Biochemistry. 2007;46: 7497–7505.
112. Ensign DL, Kasson PM, Pande VS. Heterogeneity Even at the Speed Limit of Folding: Large-scale Molecular Dynamics Study of a Fast-folding Variant of the Villin Headpiece. J Mol Biol. 2007;374: 806–816.
113. Lei H, Su Y, Jin L, Duan Y. Folding network of villin headpiece subdomain. Biophys J. Biophysical Society; 2010;99: 3374–3384.
114. Jani V, Sonavane UB, Joshi R. Microsecond scale replica exchange molecular dynamic simulation of villin headpiece: an insight into the folding landscape. J Biomol Struct Dyn. 2011;28: 845–60.
115. Piana S, Lindorff-Larsen K, Shaw DE. Protein folding kinetics and thermodynamics from atomistic simulation. Proc Natl Acad Sci. 2012;109: 17845–17850.
116. Xiao S, Patsalo V, Shan B, Bi Y, Green DF, Raleigh DP. Rational modification of protein stability by targeting surface sites leads to complicated results. Proc Natl Acad Sci. 2013;110: 11337–11342.
117. McKnight JC, Doering DS, Matsudaira PT, Kim PS. A Thermostable 35-Residue Subdomain within Villin Headpiece. J Mol Biol. 1996;260: 126–134.
118. Frank BS, Vardar D, Buckley D a, McKnight CJ. The role of aromatic residues in the hydrophobic core of the villin headpiece subdomain. Protein Sci. 2002;11: 680–687.
119. Godoy-Ruiz R, Henry ER, Kubelka J, Hofrichter J, Muñoz V, Sanchez-Ruiz JM, et al. Estimating free-energy barrier heights for an ultrafast folding protein from calorimetric and kinetic data. J Phys Chem B. 2008;112: 5938–5949.
120. Xiao S, Raleigh DP. A critical assessment of putative gatekeeper interactions in the villin headpiece helical subdomain. J Mol Biol. Elsevier B.V.; 2010;401: 274–285.
121. Hsu W-L, Shih T-C, Horng J-C. Folding stability modulation of the villin headpiece helical subdomain by 4-fluorophenylalanine and 4-methylphenylalanine. Case D, editor. Biopolymers. 2015;103: 627–637.

122. McKnight CJ, Matsudaira PT, Kim PS. NMR structure of the 35-residue villin headpiece subdomain. Nat Struct Biol. 1997;4: 180–184.
123. Obexer R, Godina A, Garrabou X, Mittl PRE, Baker D, Griffiths AD, et al. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. Nat Chem. Nature Publishing Group; 2017;9: 50–56.
124. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 2017;357: 168–175.
125. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. Nature. 2003;423: 185–190.
126. Check Hayden E. Chemistry: Designer debacle. Nature. Nature Publishing Group; 2008;453: 275–278.
127. Merrifield RB. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. J Am Chem Soc. 1963;85: 2149–2154.
128. Carpino LA, Han GY. 9-Fluorenylmethoxycarbonyl amino-protecting group. J Org Chem. 1972;37: 3404–3409.
129. Angeletti RH, Bonewald LF, Fields GB. [32] Six-year study of peptide synthesis. Methods in Enzymology. Academic Press; 1997. pp. 697–717.
130. Fields GB. Introduction to Peptide Synthesis. Current Protocols in Protein Science. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2001. p. 18.1.1-18.1.9.
131. Chan WC, White PD. Fmoc solid phase peptide synthesis : a practical approach. Oxford University Press; 2000.
132. Pedersen SL, Tofteng a P, Malik L, Jensen KJ. Microwave heating in solid-phase peptide synthesis. Chem Soc Rev. 2012;41: 1826–1844.
133. Isidro-Llobet A, Álvarez M, Albericio F. Amino Acid-Protecting Groups. Chem Rev. 2009;109: 2455–2504.
134. Dias AMGC, Santos R dos, Iranzo O, Roque ACA. Affinity adsorbents for proline-rich peptide sequences: a new role for WW domains. RSC Adv. 2016;6: 68979–68988.
135. Aguilar M-I. HPLC of Peptides and Proteins. New Jersey: Humana Press; 2003.
136. Chapman JR. Practical organic mass spectrometry : a guide for chemical and biochemical analysis. J. Wiley; 1995.
137. Nomura A, Sugiura Y. Contribution of Individual Zinc Ligands to Metal Binding and Peptide Folding of Zinc Finger Peptides. Inorg Chem. 2002;41: 3693–3698.
138. Nomura A, Sugiura Y. Sequence-selective and hydrolytic cleavage of DNA by zinc finger mutants. J Am Chem Soc. 2004;126: 15374–5.
139. Kiefer LL, Paterno SA, Fierke CA. Hydrogen bond network in the metal binding site of carbonic anhydrase enhances zinc affinity and catalytic efficiency. J Am Chem Soc. 1995;117: 6831–6837.
140. Marino SF, Regan L. Secondary ligands enhance affinity at a designed metal-binding site. Chem Biol. Cell Press; 1999;6: 649–655.
141. Vita C, Roumestand C, Toma F, Menez A. Scorpion toxins as natural scaffolds for protein engineering. Proc Natl Acad Sci. 1995;92: 6404–6408.
142. Kiyokawa T, Kanaori K, Tajima K, Koike M, Mizuno T, Oku J-I, et al. Binding of Cu(II) or Zn(II) in a de novo designed triple-stranded alpha-helical coiled-coil toward a prototype for a metalloenzyme. J Pept Res. 2004;63: 347–353.
143. Zastrow ML, Pecoraro VL. Influence of active site location on catalytic activity in de novo -designed zinc metalloenzymes. J Am Chem Soc. 2013;135: 5895–5903.
144. Cangelosi VM, Deb A, Penner-Hahn JE, Pecoraro VL. A De Novo Designed Metalloenzyme for the Hydration of CO2. Angew Chemie Int Ed. 2014;53: 7900–7903.
145. Regan L, Clarke ND. A tetrahedral zinc(II)-binding site introduced into a designed protein. Biochemistry. 1990;29: 10878–10883.
146. Müller HN, Skerra a. Grafting of a high-affinity Zn(II)-binding site on the beta-barrel of retinol-binding protein results in enhanced folding stability and enables simplified purification. Biochemistry. 1994;33: 14126–14135.
147. Wade WS, Koh JS, Han N, Hoekstra DM, Lerner R a. Engineering metal coordination sites into the antibody light chain. J Am Chem Soc. 1993;115: 4449–4456.
148. Adams JT, Deweese JA. Creation of a Novel Biosensor for Zn(II). J Am Chem Soc. 1994;53: 1745–1747.

149. Pessi a, Bianchi E, Crameri a, Venturini S, Tramontano a, Sollazzo M. A designed metal-binding protein with a novel fold. Nature. 1993;362: 367–369.
150. Hunt JA, Fierke CA. Selection of Carbonic Anhydrase Variants Displayed on Phage. J Biol Chem. American Society for Biochemistry and Molecular Biology; 1997;272: 20364–20372.
151. Hunt JA, Ahmed M, Fierke CA. Metal Binding Specificity in Carbonic Anhydrase Is Influenced by Conserved Hydrophobic Core Residues. Biochemistry. 1999;38: 9054–9062.
152. Salgado EN, Radford RJ, Tezcan FA. Metal-Directed Protein Self-Assembly. Acc Chem Res. 2010;43: 661–672.
153. Mills JH, Khare SD, Bolduc JM, Forouhar F, Mulligan VK, Lew S, et al. Computational Design of an Unnatural Amino Acid Dependent Metalloprotein with Atomic Level Accuracy. J Am Chem Soc. 2013;135: 13393–13399.
154. Der BS, Edwards DR, Kuhlman B. Catalysis by a de novo zinc-mediated protein interface: implications for natural enzyme evolution and rational enzyme engineering. Biochemistry. 2012;51: 3933–40.
155. Lippard SJ, Berg JM. Principles of Bioinorganic Chemistry. Principles of Bioinorganic Chemistry. University Science Books; 1995.
156. Rush RM, Yoe JH. Colorimetric Determination of Zinc and Copper with 2-Carboxy-2´-hydroxy-5´-sulfoformazylbenzene. Anal Chem. 1954;26: 1345–1347.
157. Visual egta titration of calcium in the,presence of magnesium. Talanta. Elsevier; 1959;2: 38–51.
158. Shaw CF, Laib JE, Savas MM, Petering DH. Biphasic kinetics of aurothionein formation from gold sodium thiomalate: a novel metallochromic technique to probe zinc(2+) and cadmium(2+) displacement from metallothionein. Inorg Chem. 1990;29: 403–408.
159. Frankel a D, Berg JM, Pabo CO. Metal-dependent folding of a single zinc finger from transcription factor IIIA. Proc Natl Acad Sci. 1987;84: 4841–4845.
160. Siedlecka M, Goch G, Ejchart a, Sticht H, Bierzynski a. Alpha-helix nucleation by a calcium-binding peptide loop. Proc Natl Acad Sci U S A. 1999;96: 903–8.
161. Suzuki K, Hiroaki H, Kohda D, Nakamura H, Tanaka T. Metal ion induced self-assembly of a designed peptide into a triple-stranded alpha-helical bundle: A novel metal binding site in the hydrophobic core. J Am Chem Soc. 1998;120: 13008–13015.
162. Farrer BT, Harris NP, Balchus KE, Pecoraro VL. Thermodynamic model for the stabilization of trigonal thiolato mercury(II) in designed three-stranded coiled coils. Biochemistry. 2001;40: 14696–14705.
163. Kim CA, Berg JM. Thermodynamic β -sheet propensities measured using a zinc-finger host peptide. Nature. 1993;362: 267–270.
164. Bianchi E, Folgori A, Wallace A, Nicotra M, Acali S, Phalipon A, et al. A Conformationally Homogeneous Combinatorial Peptide Library. J Mol Biol. Academic Press; 1995;247: 154–160.
165. Reddi a R, Guzman TR, Breece RM, Tierney DL, Gibney BR. Deducing the Energetic Cost of Protein Folding in Zinc Finger Proteins Using Designed Metallopeptide. J Am Chem Soc. 2007;129: 12815–12827.
166. Sénèque O, Bonnet E, Joumas FL, Latour J-M. Cooperative Metal Binding and Helical Folding in Model Peptides of Treble-Clef Zinc Fingers. Chem - A Eur J. 2009;15: 4798–4810.
167. Chen YH. Determination of the helix and β form of proteins in aqueous solution by circular dichroism. Biochemistry. 1974;13: 3350–3359.
168. Kelly SM, Jess TJ, Price NC. How to study proteins by circular dichroism. Biochim Biophys Acta. 2005;1751: 119–39.
169. Greenfield NJ. Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. Nat Protoc. 2007;1: 2527–2535.
170. Gill SC, von Hippel PH. Calculation of protein extinction coefficients from amino acid sequence data. Anal Biochem. 1989;182: 319–26.
171. Edelhoch H. Spectroscopic determination of tryptophan and tyrosine in proteins. Biochemistry. 1967;6: 1948–54.
172. Ellman GL. Tissue sulfhydryl groups. Arch Biochem Biophys. 1959;82: 70–77.

173. Thordarson P. Determining association constants from titration experiments in supramolecular chemistry. Chem Soc Rev. 2011;40: 1305–1323.
174. Weber G, Anderson SR. Multiplicity of Binding. Range of Validity and Practical Test of Adair's Equation *. Biochemistry. 1965;4: 1942–1947.
175. Deranleau DA. Theory of the measurement of weak molecular complexes. I. General considerations. J Am Chem Soc. 1969;91: 4044–4049.
176. Mekmouche Y, Coppel Y, Hochgräfe K, Guilloreau L, Talmard C, Mazarguil H, et al. Characterization of the ZnII binding to the peptide amyloid-beta1-16 linked to Alzheimer's disease. Chembiochem. 2005;6: 1663–71.
177. Säbel CE, Neureuther JM, Siemann S. A spectrophotometric method for the determination of zinc, copper, and cobalt ions in metalloproteins using Zincon. Anal Biochem. Elsevier Inc.; 2010;397: 218–26.
178. Sénèque O, Latour J-M. Coordination Properties of Zinc Finger Peptides Revisited: Ligand Competition Studies Reveal Higher Affinities for Zinc and Cobalt. J Am Chem Soc. 2010;132: 17760–17774.
179. Namdarghanbari MA. Cellular Zinc Trafficking : the Zinc Proteome and Its Reactions with Cadmium. Theses and Dissertations. University of Wisconsin Milwaukee. 2014.
180. Bombarda E, Grell E, Roques BP, Mély Y. Molecular Mechanism of the Zn2+-Induced Folding of the Distal CCHC Finger Motif of the HIV-1 Nucleocapsid Protein. Biophys J. Elsevier; 2007;93: 208–217.
181. Imanishi M, Hori Y, Nagaoka M, Sugiura Y. DNA-Bending Finger: Artificial Design of 6-Zinc Finger Peptides with Polyglycine Linker and Induction of DNA Bending. Biochemistry. 2000;39: 4383–4390.
182. Kochańczyk T, Drozd A, Krężel A. Relationship between the architecture of zinc coordination and zinc binding affinity in proteins - insights into zinc regulation. Metallomics. 2015;7: 244–257.
183. Reddi AR, Pawlowska M, Gibney BR. Evaluation of the Intrinsic Zn(II) Affinity of a Cys3His1 Site in the Absence of Protein Folding Effects. Inorg Chem. 2015;54: 5942–8.
184. Rich A, Bombarda E. Thermodynamics of Zn2+ Binding to Cys2His2 and Cys2HisCys Zinc Fingers and a Cys4 Transcription Factor Site. J Am Chem Soc. 2012;134: 10405–10418.
185. MacPhee CE, Perugini MA, H. Sawyer W, Howlett GJ. Trifluoroethanol induces the self-association of specific amphipathic peptides. FEBS Lett. Federation of European Biochemical Societies; 1997;416: 265–268.
186. Suárez-Diez M, Pujol AM, Matzapetakis M, Jaramillo A, Iranzo O. Computational protein design with electrostatic focusing: Experimental characterization of a conditionally folded helical domain with a reduced amino acid alphabet. Biotechnol J. 2013;8: 855–864.
187. Gatti-Lafranconi P, Hollfelder F. Flexibility and reactivity in promiscuous enzymes. Chembiochem. 2013;14: 285–92.
188. Dellus-Gur E, Toth-Petroczy A, Elias M, Tawfik DS. What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. J Mol Biol. Elsevier Ltd; 2013;425: 2609–21.
189. Tokuriki N, Tawfik DS. Protein Dynamism and Evolvability. Science. 2009;324: 203–207.
190. Münz M, Hein J, Biggin PC. The Role of Flexibility and Conformational Selection in the Binding Promiscuity of PDZ Domains. Zhou H-X, editor. PLoS Comput Biol. Public Library of Science; 2012;8: e1002749.
191. Piazzetta P, Marino T, Russo N, Salahub DR. The role of metal substitution in the promiscuity of natural and artificial carbonic anhydrases. Coord Chem Rev. Elsevier B.V.; 2017;345: 73–85.
192. Schulenburg C, Hilvert D. Protein conformational disorder and enzyme catalysis. Top Curr Chem. Springer, Berlin, Heidelberg; 2013;337: 41–68.
193. Moroz YS, Dunston TT, Makhlynets O V., Moroz O V., Wu Y, Yoon JH, et al. New Tricks for Old Proteins: Single Mutations in a Nonenzymatic Protein Give Rise to Various Enzymatic Activities. J Am Chem Soc. 2015;137: 14905–14911.
194. Burton AJ, Thomson AR, Dawson WM, Brady RL, Woolfson DN. Installing hydrolytic activity into a completely de novo protein framework. Nat Chem. Nature Publishing Group; 2016;8: 837–844.
195. Srivastava KR, Durani S. Design of a zinc-finger hydrolase with a synthetic αββ protein.

PLoS One. 2014;9: e96234.
196. Rufo CM, Moroz YS, Moroz O V., Stöhr J, Smith T a., Hu X, et al. Short peptides self-assemble to produce catalytic amyloids. Nat Chem. 2014;6: 303–309.
197. Broo KS, Brive L, Ahlberg P, Baltzer L. Catalysis of hydrolysis and transesterification reactions of p- Nitrophenyl esters by a designed helix-loop-helix dimer. J Am Chem Soc. 1997;119: 11362–11372.
198. Zastrow ML, Peacock AF a, Stuckey J a, Pecoraro VL. Hydrolytic catalysis and structural stabilization in a designed metalloprotein. Nat Chem. Nature Publishing Group; 2012;4: 118–23.
199. Song WJ, Tezcan FA. A designed supramolecular protein assembly with in vivo enzymatic activity. Science. 2014;346: 1525–1528.
200. Bai Y, Ling Y, Shi W, Cai L, Jia Q, Jiang S, et al. Heteromeric assembled polypeptidic artificial hydrolases with a six-helical bundle scaffold. Chembiochem. 2011;12: 2647–58.
201. Árus D, Nagy NV, Dancs Á, Jancsó A, Berkecz R, Gajda T. A minimalist chemical model of matrix metalloproteinases--can small peptides mimic the more rigid metal binding sites of proteins? J Inorg Biochem. 2013;126: 61–9.
202. Gordon SR, Stanley EJ, Wolf S, Toland A, Wu SJ, Hadidi D, et al. Computational Design of an α-Gliadin Peptidase. J Am Chem Soc. 2012;134: 20513–20520.
203. Kim MG, Yoo SH, Chei WS, Lee TY, Kim HM, Suh J. Soluble artificial metalloproteases with broad substrate selectivity, high reactivity, and high thermal and chemical stabilities. J Biol Inorg Chem. 2010;15: 1023–31.
204. Yenjai S, Malaikaew P, Liwporncharoenvong T, Buranaprapuk A. Selective cleavage of pepsin by molybdenum metallopeptidase. Biochem Biophys Res Commun. Elsevier Inc.; 2012;419: 126–9.
205. Kim MG, Kim HM, Suh J. Artificial metalloprotease based on Co(III)oxacyclen-aldehyde conjugate. Bull Korean Chem Soc. 2011;32: 3113–3116.
206. Zhang T, Ozbil M, Barman A, Paul TJ, Bora RP, Prabhakar R. Theoretical insights into the functioning of metallopeptidases and their synthetic analogues. Acc Chem Res. 2015;48: 192–200.
207. Innocenti A, Scozzafava A, Parkkila S, Puccetti L, De Simone G, Supuran CT. Investigations of the esterase, phosphatase, and sulfatase activities of the cytosolic mammalian carbonic anhydrase isoforms I, II, and XIII with 4-nitrophenyl esters as substrates. Bioorganic Med Chem Lett. 2008;18: 2267–2271.
208. Kezdy FJ, Bender ML. The Kinetics of the α-Chymotrypsin-Catalyzed Hydrolysis of p-Nitrophenyl Acetate *. Biochemistry. 1962;1: 1097–1106.
209. Jencks WP, Gilchrist M. Nonlinear structure-reactivity correlations. The reactivity of nucleophilic reagents toward esters. J Am Chem Soc. 1968;90: 2622–2637.
210. Risso VA, Martinez-Rodriguez S, Candel AM, Krüger DM, Pantoja-Uceda D, Ortega-Muñoz M, et al. De novo active sites for resurrected Precambrian enzymes. Nat Commun. 2017;8: 16113.
211. STÖCKER W, SAUER B, ZWILLING R. Kinetics of Nitroanilide Cleavage by Astacin. Biol Chem Hoppe Seyler. 1991;372: 385–392.
212. Wolz RL. A Kinetic Comparison of the Homologous Proteases Astacin and Meprin A. Arch Biochem Biophys. 1994;310: 144–151.
213. Orning L, Gierse JK, Fitzpatrick FA. The bifunctional enzyme leukotriene-A4 hydrolase is an arginine aminopeptidase of high efficiency and specificity. J Biol Chem. 1994;269: 11269–73.
214. Blomberg R, Kries H, Pinkas DM, Mittl PRE, Grütter MG, Privett HK, et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. Nature. 2013;503: 418–21.
215. Renata H, Wang ZJ, Arnold FH. Expanding the enzyme universe: Accessing non-natural reactions by mechanism-guided directed evolution. Angew Chemie - Int Ed. 2015;54: 3351–3367.
216. Kiss G, Röthlisberger D, Baker D, Houk KN. Evaluation and ranking of enzyme designs. Protein Sci. 2010;19: 1760–1773.
217. Alexandrova AN, Röthlisberger D, Baker D, Jorgensen WL. Catalytic Mechanism and Performance of Computationally Designed Enzymes for Kemp Elimination. J Am Chem Soc. American Chemical Society; 2008;130: 15907–15915.
218. Privett HK, Kiss G, Lee TM, Blomberg R, Chica R a, Thomas LM, et al. Iterative approach

to computational enzyme design. Proc Natl Acad Sci. 2012;109: 3790–3795.
219. Hünenberger PH, Gunsteren WF. Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications. Computer Simulation of Biomolecular Systems. Dordrecht: Springer Netherlands; 1997. pp. 3–82.
220. Kiss G, Pande VS, Houk KN. Molecular Dynamics Simulations for the Ranking, Evaluation, and Refinement of Computationally Designed Proteins. Methods in Enzymology. 1st ed. Elsevier Inc.; 2013. pp. 145–170.
221. Dodani SC, Kiss G, Cahn JKB, Su Y, Pande VS, Arnold FH. Discovery of a regioselectivity switch in nitrating P450s guided by molecular dynamics simulations and Markov models. Nat Chem. 2016;8: 419–425.
222. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: A message-passing parallel molecular dynamics implementation. Comput Phys Commun. 1995;91: 43–56.
223. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1–2: 19–25.
224. Li W, Wang J, Zhang J, Wang W. Molecular simulations of metal-coupled protein folding. Curr Opin Struct Biol. Elsevier Ltd; 2015;30: 25–31.
225. Lemkul JA, Roux B, Van Der Spoel D, Mackerell AD. Implementation of extended Lagrangian dynamics in GROMACS for polarizable simulations using the classical Drude oscillator model. J Comput Chem. 2015;36: 1473–1479.
226. Lemkul JA, Huang J, Roux B, Mackerell AD. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. Chem Rev. 2016;116: 4983–5013.
227. Pang YP, Xu K, Yazal JE, Prendergas FG. Successful molecular dynamics simulation of the zinc-bound farnesyltransferase using the cationic dummy atom approach. Protein Sci. Wiley-Blackwell; 2000;9: 1857–65.
228. Pang Y-P. Novel Zinc Protein Molecular Dynamics Simulations: Steps Toward Antiangiogenesis for Cancer Treatment. J Mol Model. 1999;5: 196–202.
229. Oelschlaeger P, Schmid RD, Pleiss J. Insight into the mechanism of the IMP-1 metallo-beta-lactamase by molecular dynamics simulations. Protein Eng. 2003;16: 341–350.
230. Oelschlaeger P, Schmid RD, Pleiss J. Modeling Domino Effects in Enzymes: Molecular Basis of the Substrate Specificity of the Bacterial Metallo-β-lactamases IMP-1 and IMP-6. Biochemistry. American Chemical Society; 2003;42: 8945–8956.
231. Park JG, Sill PC, Makiyi EF, Garcia-Sosa AT, Millard CB, Schmidt JJ, et al. Serotype-selective, small-molecule inhibitors of the zinc endopeptidase of botulinum neurotoxin serotype A. Bioorganic Med Chem. 2006;14: 395–408.
232. Carvalho HF, Barbosa AJM, Roque ACA, Iranzo O, Branco RJF. Integration of Molecular Dynamics Based Predictions into the Optimization of De Novo Protein Designs: Limitations and Benefits. In: Samish I, editor. Computational Protein Design. New York, NY: Springer New York; 2017. pp. 181–201.
233. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins Struct Funct Bioinforma. 2010;78: 1950–1958.
234. Aliev AE, Kulke M, Khaneja HS, Chudasama V, Sheppard TD, Lanigan RM. Motional timescale predictions by molecular dynamics simulations: Case study using proline and hydroxyproline sidechain dynamics. Proteins Struct Funct Bioinforma. 2014;82: 195–215.
235. Best RB, Hummer G. Optimized Molecular Dynamics Force Fields Applied to the Helix−Coil Transition of Polypeptides. J Phys Chem B. American Chemical Society; 2009;113: 9004–9015.
236. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983;79: 926.
237. Nosé S. A molecular dynamics method for simulations in the canonical ensemble. Mol Phys. 1984;52: 255–268.
238. Hoover WG. Canonical dynamics: Equilibrium phase-space distributions. Phys Rev A. 1985;31: 1695–1697.
239. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys. 1981;52: 7182–7190.
240. Nosé S, Klein ML. Constant pressure molecular dynamics for molecular systems. Mol

Phys. 1983;50: 1055–1076.
241. Darden T, York D, Pedersen L. Particle mesh Ewald: An N ·log( N ) method for Ewald sums in large systems. J Chem Phys. 1993;98: 10089–10092.
242. Daura X, Gademann K, Jaun B, Seebach D, Van Gunsteren WF, Mark AE. Peptide Folding: When Simulation Meets Experiment. Angew Chem Int Ed. 1999;38: 38: 236-240.
243. Kabsch W, Sander C. Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers. 1983;22: 2577–2637.
244. Touw WG, Baakman C, Black J, Te Beek TAH, Krieger E, Joosten RP, et al. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 2015;43: D364–D368.
245. Hocking HG, Häse F, Madl T, Zacharias M, Rief M, Žoldák G. A Compact Native 24-Residue Supersecondary Structure Derived from the Villin Headpiece Subdomain. Biophys J. 2015;108: 678–686.
246. Okazaki K -i., Takada S. Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms. Proc Natl Acad Sci. National Academy of Sciences; 2008;105: 11182–11187.

# Annex 1

## Comparison of the Internal Dynamics of Metalloproteases Provides New Insights on Their Function and Evolution†

Henrique F. Carvalho[1,2], Ana C. A. Roque[1], Olga Iranzo[3*], Ricardo J. F. Branco[1*]

1- UCIBIO, REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.

2- Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal.

3- Aix Marseille Université, Centrale Marseille, CNRS, iSm2 UMR 7313, 13397, Marseille, France.

†Author submitted manuscript. Originally published in colour version in Carvalho, HF, Roque ACA, Iranzo O, Branco RJF., *Comparison of the Internal Dynamics of Metalloproteases Provides New Insights on Their Function and Evolution,* PLoS ONE, 2015. 10(9): e0138118. September 23, 2015

*Corresponding authors email: OI: olga.iranzo@univ-amu.fr; RJFB: ricardo.branco@fct.unl.pt

## Abstract

Metalloproteases have evolved in a vast number of biological systems, being one of the most diverse types of proteases and presenting a wide range of folds and catalytic metal ions. Given the increasing understanding of protein internal dynamics and its role in enzyme function, we are interested in assessing how the structural heterogeneity of metalloproteases translates into their dynamics. Therefore, the dynamical profile of the clan MA type protein thermolysin, derived from an Elastic Network Model of protein structure, was evaluated against those obtained from a set of experimental structures and molecular dynamics simulation trajectories. A close correspondence was obtained between modes derived from the coarse-grained model and the subspace of functionally-relevant motions observed experimentally, the later being shown to be encoded in the internal dynamics of the protein. This prompted the use of dynamics-based comparison methods that employ such coarse-grained models in a representative set of clan members, allowing for its quantitative description in terms of structural and dynamical variability. Although members show structural similarity, they nonetheless present distinct dynamical profiles, with no apparent correlation between structural and dynamical relatedness. However, previously unnoticed dynamical similarity was found between the relevant members Carboxypeptidase *Pfu*, Leishmanolysin, and *Botulinum* Neurotoxin Type A, despite sharing no structural similarity. Inspection of the respective alignments shows that dynamical similarity has a functional basis, namely the need for maintaining proper intermolecular interactions with the respective substrates. These results suggest that distinct selective pressure mechanisms act on metalloproteases at structural and dynamical levels through the course of their evolution. This work shows how new insights on metalloprotease function and evolution can be assessed with comparison schemes that incorporate

information on protein dynamics. The integration of these newly developed tools, if applied to other protein families, can lead to more accurate and descriptive protein classification systems.

# Introduction

Proteases are a vast class of enzymes found in all kingdoms of life that participate in a wide range of biological processes [1]. They present different catalytic chemistries, structures, specificities, oligomeric states, and are grouped into distinct families and clans according to different classification schemes. Examples are the MEROPS [2], SCOP [3], and CATH databases [4], which use a combination of sequence- and structure-based methods for grouping distinct proteins. The need for a better understanding of their function has also led to the search of other common features shared between a wide range of known class members [5,6]; perhaps the most pervasive similarity was identified by Tyndall *et al.,* where it was observed an almost universal binding of the Aspartic, Serine, Cysteine, and Metallo- proteases to the extended β-strand conformations of substrates, products, and inhibitors of peptidic and non-peptidic origin [5]. Nonetheless, there is still the need for a better characterization of the multiple factors governing protease function and evolution. We therefore tested if the employment of novel protein comparison tools can, when combined with conventional comparison methods, help in the search of additional features shared between distinct protease members.

Protein internal dynamics plays an important role on enzyme function, since it encompasses the space of catalytically-relevant structural changes occurring in a given fold during the reaction path [7–13]. These structural changes span a broad range of time-scales and magnitudes; from harmonic vibrations of bonds and angles occurring at the femtosecond time-scale, to global conformational fluctuations of large domains at the microsecond time-scale, some of them associated with substrate binding and product release [14–18]. Therefore, understanding the relation between protein internal dynamics and its structural and functional features is a challenging task, since it depends not only on the analysis of protein dynamics at different time scales, but also on their functionally-relevant molecular states (*e.g.* bound *vs* unbound state).

There is an ongoing debate on how protein internal dynamics is subject to evolutionary selection due to its functional significance and how it is related to sequence and structure evolution [19,20]. It has been shown that internal dynamics and backbone flexibility are conserved in homologous proteins [6,21–27]. Specifically, it has been observed that: *i)* low-frequency, large-amplitude normal modes tend to be evolutionarily conserved [24,28] and; *ii)* there is a significant correspondence between low-energy modes determined for superfamily structural cores and the modes of structural variance observed within protein superfamilies [29,30]. These findings support the notion that conservation of protein dynamics is subject to evolutionary selection and were based on the observation that ligand binding can be described in many cases by few low-energy normal modes [31,32]. However, similarities of low-energy modes observed between non-homologous proteins with the same architecture and even between unrelated proteins [26,33], together with the observation that low-energy modes are more robust to random mutations suggests that protein dynamics may not always be subject to evolutionary selection [34–36].

Additional insights have been provided by phylogenetic studies addressing the evolution of normal modes. It was shown that changes in protein dynamics are associated with functional divergence in enzymatic families and that non-homologous enzymes that perform similar functions also share similar motions related to catalysis [37,38]. Protein dynamics has also been found to diverge between structurally related proteins at functionally important sites [39,40], and this divergence has been argued to be dependent on other functional requirements, such as intracellular localization [41]. Finally, protein dynamics has also been associated with the evolution of new enzymatic functions and on the promiscuity of enzymes [42–45]. These findings suggest that protein dynamics plays an important role in the function and evolution of enzymes, although the extent to which evolution has selected for this particular trait still remains unclear.

The development of dynamics-based comparison methods has been crucial to the above mentioned studies and has provided insights that may not be detectable from sequence or structure comparison methods alone [27,40,46–52]. These methods rely on Essential Dynamics Analysis (EDA), to retrieve the collective motions of protein structures, which are typically obtained from Molecular Dynamics (MD) simulations or alternatively by simplified, coarse-grained representations of protein structure, such as Elastic Network Models (ENM). The similarity between modes

of two distinct structures can be compared based on different quantitative schemes and therefore new relations between proteins can be sought based on their dynamical properties [50].

In the search for common features shared between different types of proteases, Carnevale *et al.* carried out pairwise structure- and dynamics-based alignment of 17 representative protease structures with minimal mutual sequence identity and distinct folds [6]. In most cases the division into distinct folds was consistent with the division in clans of the MEROPS classification scheme, indicating that structures with different evolutionary origins adopt distinct folds. Nonetheless, significant structural similarities among proteases of different clans were identified, thus suggesting a convergent evolutionary process. Indeed, in pairs of structures showing higher structural similarity, the aligned segments in both structures consisted on regions close to the active-site, even for pairs from distinct clans (*i.e.* different catalytic chemistries). The authors proposed that a criterion for catalytic activity not dependent on chemical determinants could be at play, namely the dependence on specific and concerted protein motions related to function. A close correspondence of the internal dynamics between some pairs was identified, which consisted in rearrangements of active-site surroundings that lead to distortions of the substrate-accommodating pockets. Therefore, it was suggested that a "dynamical selection" process, operated by the necessity to interact with substrates in well-defined geometrical arrangements, may lead to convergence or conservation of the internal dynamics in proteases. Indeed, recent work reported by Micheletti further identified significant dynamical similarities between proteases from different clans with no detectable structural similarity [50]. It was therefore suggested that dynamics-based comparison methods could be useful in detecting functionally-related features shared by proteases that otherwise would remain elusive using only sequence- and structure-based methods.

Metalloproteases (MPs) are one of the most diverse types of proteases, presenting a wide range of folds and catalytic metal ions. They are divided in more than 40 families identified among all kingdoms of life. In the case of the MEROPS MA clan, where most of the known MPs are grouped by common ancestry, its members are characterized by a single catalytic zinc ion, a consensus HEXXH sequence motif and a common fold architecture. However, this structural conservation is not observed at the domain level since members from different families have distinct domain composition and topology. MPs are therefore attractive candidates to study the relationship between structure and dynamics within a protein clan. In this work, the suitability of employing coarse-grained methods to the study of MP internal dynamics was first made by comparison of ENM-derived internal dynamics profile of thermolysin to those obtained by Principal Component Analysis (PCA) of crystal structures and EDA of MD simulation trajectories. Subsequently, an analysis of pairwise structural- and dynamics-based alignments of a representative set of MPs from 13 families of the MA clan was performed. It was found for members of this clan that dynamical similarity does not appear to correlate with structural similarity. Interestingly, pairs having high dynamical similarity despite having no structural similarity were identified. Inspection of the produced alignments indicates that in these cases conservation of internal dynamics has a functional basis, namely to dictate proper interactions with the substrate. Our data show the suitability of using simple comparison schemes that incorporate information on dynamics to provide new insights on MPs function and evolution, unveiling their potential as tools to study the role of internal dynamics in protein evolution.

# Methods

### Internal Dynamics of Thermolysin

PCA of the structural set was made to obtain the respective Principal Components (PC), using the ProDy software [53,54]. The 3*N* x 3*N* covariance matrix was calculated over *n*, where *N*=316 is the number of residues in thermolysin (represented by the respective $C^{\alpha}$ atom coordinates), and *n*=112 is the number of thermolysin crystal structures (Uniprot ID: P00800) retrieved from the PDB [55], with corresponding IDs (Table A in S1 File). Structures were initially superposed to the unbound crystal structure (PDB ID: 1L3F) to obtain mean coordinates, then iteratively superposed until convergence to eliminate rigid-body translational and rotational differences.

MD simulation of thermolysin in the unbound form (PDB ID: 1L3F) was performed using GROMACS 4.6.1 simulation package with GPU acceleration [56–58], with the AMBER99SB-

ILDN force field [59] (Sim1). The system was solvated with explicit Simple Point Charge (Extended) water model (SPC/E, [60]) and placed in a dodecahedral box, each edge at least 12 Å from the protein surface. The system was charge-neutralized by addition of two sodium counterions and minimized in two steps to remove atom clashes and bond contacts: first by a steepest descent minimization algorithm (2000 steps), followed by a conjugated gradient algorithm (1000 steps). The energy-minimized model was coupled to the V-rescale thermostat (300 K, coupling time constant 0.1 ps [61]) and Berendsen barostat (1 bar, coupling time-constant 0.6 ps [62]) and then equilibrated, where the force-constant of positional restraints for heavy-atoms was decreased from 1000 kJ/mol, 100 kJ/mol to 10 kJ/mol in three consecutive steps (100 ps). A production phase was finally run for 20 ns, with an integration step of 2 fs. Long-range electrostatics were treated with the Particle-Mesh Ewald algorithm and distance constraints between all H-bonds was enforced by the LINCS algorithm [63]. Although the employed force field does not appropriately represent the interactions between the catalytic zinc metal ion and the coordinated residues [64], this metal ion is not considered to have a structural role [65]. An additional non-biologically relevant zinc ion found at the active site of the crystallographic structure was removed. The protonation states of active site residues from the conserved HEXXH sequence motif were manually attributed, with H142 and H146 monoprotonated at the $N^{\delta}$ position and E143 and E146 not protonated, taking into account its specific pKa. A replicate of the MD simulation was carried out (Sim2), where all abovementioned simulation set-up parameters were kept unchanged. After removal of the global rotation and translation of the protein, simulation trajectories show a RMSD convergence after 1 ns of the production phase (Figure A in S2 File). Therefore, only the 1-20 ns interval of full trajectories were used for EDA with the ProDy software, with the 1 ns frame being used as the reference structure. EDA is based on PCA, with the difference that the respective PCs (termed here as ED modes) are calculated based on $n$=9500 trajectory frames taken at intervals of 2 ps, considering only the $N$=316 $C^{\alpha}$ atom coordinates.

The Anisotropic Network Model (ANM, [66–68]) of thermolysin in the unbound form (PDB ID: 1L3F) was calculated using the ProDy software. The $C^{\alpha}$ atom coordinates were used as node representations of each residue of the protein ($N$=316) to build the respective 3$N$x3$N$ Hessian matrix. Variations of the model included additional nodes, matching the coordinates of the catalytic zinc ion alone ($N$=317) or the catalytic zinc ion and the four calcium ions ($N$=321). The uniform force constant γ=1 was used to calculate the overall potential of the system and the interaction cutoff distance $r_c$=15 Å was used to generate the respective Kirchhoff matrix of inter-residue contacts.

The collectivity degree, $\kappa$, was used as a measure of the number of atoms significantly affected by a given PC, ED or NM mode [69]. This value varies from $\kappa$=1 for modes describing global translations of the protein to $\kappa = N^{-1}$ if only one atom is affected ($N$=316). The overlap, or correlation cosine, between two modes is given by the dot product of the respective eigenvectors after normalization, being equal to one if two modes are identical. The subspace overlap between two sets of modes is given by corresponding Root Mean Square Inner Product (RMSIP) value [70]. The respective overlap between the covariance matrices was also calculated, with a value of 1 if the two matrices are identical and of 0 when the respective subspaces are orthogonal [71].

## Selection of representative structures

A representative set of MP structures from distinct families belonging to the MA clan of the peptidase database MEROPS (release 9.9) was selected [2]. In MEROPS, members are grouped in families based on their sequence similarity. Families are further grouped in clans when there is detectable structural homology, implying common ancestry. For each of the 23 (out of 42) families that contain members with resolved structures in the PDB, a representative structure was selected based on the following criteria: *i)* unbound structure containing no inhibitor molecule, substrate or substrate analog molecule and; *ii)* the structure with highest resolution or with no mutations. Only unbound crystal structures were selected since they are assumed to represent the native conformation of the corresponding protein. Also, the degree and mode of conformational change upon substrate or inhibitor binding is not well characterized for most clan members as it is for thermolysin, and the effects associated with ligand heterogeneity can be ruled out. The resulting set is comprised of 13 structures belonging to distinct families. For each structure, the respective information from the SCOP and CATH databases was retrieved [3,4]. In MEROPS,

MA clan members are grouped into the MA(E) and MA(M) subclans, commonly termed as gluzincins and metzincins, respectively. These two subclans are divided based on the nature of the third zinc ligand: in MA(E) a glutamate, 18-72 residues apart from the conserved HEXXH motif towards the C-terminal; in MA(M) an histidine or aspartate in the extended HEXXHXXGXXH/D motif [1]. All structures correspond to the monomeric form of the proteins and the majority is characterized by a two-domain peptidase unit, with a conserved N-terminal domain and a more variable C-terminal domain. The active site is generally located between these two domains. Despite these general features, the set is structurally heterogeneous, with proteins containing domains that differ at the class, architecture and topology level of CATH classification criteria. In most cases, MEROPS classification is coincident with SCOP classification at the family level, except for the SCOP family 55505 “Neurolysin-like” that combines MEROPS families M2, M3 and M32; and also family 55487 “Zinc Protease” that combines M7 and M35 families. In the case of families M10 and M12, which are divided into subfamilies due to deep sequence divergence between their members, a representative was selected for each subfamily. Representatives of the MA(E) subclan families M3, M4 and M27 are endopeptidases and M1, M2 and M32 are exopeptidases, while in the MA(M) subclan all representatives are endopeptidases.

### Structure-based alignments

Pairwise structural alignment of structures was performed with the DaliLite web-server for all 78 distinct pairs [72]. The embedded DALI algorithm identifies blocks of residues between two distinct structures that have similar inter-residue distances. Matching regions are evaluated based on a knowledge-based score and the produced alignment is the one maximizing this value for a variable number of distinct blocks. The statistical significance of the alignment is quantified in terms of a Z-Score that compares the obtained score with the one expected for a pair of structurally unrelated structures of the same size. A Z-Score greater than 2.0 is considered significant and was used as threshold value for a pair of structures to be considered as structurally similar [72]. In the MEROPS, a structure is grouped in a predefined clan if a Z-Score greater than 6.0 is obtained between the structure and at least one member of that clan. Therefore, not all pairs of representatives are expected to be structurally similar.

### Dynamics-based alignments

Dynamics-based alignment of all pairs used for structural alignments were performed using the ALADYN web server [51]. First, the implemented algorithm calculates the low-energy modes for each structure and then it detects regions of both proteins with similar dynamic profiles. Calculation of modes is based on the coarse-grained β-Gaussian ENM, where amino acids are represented by a two-centroid representation: $C^{\alpha}$ atom for the main chain and; $C^{\beta}$ for the orientation of the side chains (except for Gly residues) [73,74]. This ENM has been shown to describe protein motions similarly to the employed ANM [75]. The dynamics-based alignment is made by rewarding superpositions of proteins regions that exhibit high overlap between the 10 first modes for each amino acid pair within the cutoff distance of 7 Å. This allows for the alignment of proteins with different sequences and size. Following the optimization of scoring function, the statistical significance of the resulting alignment is evaluated against a reference probability distribution of scores obtained from alignments of unrelated protein pairs, being expressed in terms of a P-value. Dynamics-based alignments of two structures with P-value smaller than 0.02 are considered statistically significant and was used as the cut-off value to consider two structures as dynamically similar [51].

## Results and Discussion

### Internal Dynamics of Thermolysin

Thermolysin, the MA clan type peptidase, is a 316 residue-long thermostable neutral MP from *Bacillus thermoproteolyticus* [1]. It presents endopeptidase activity towards peptide substrates,

cleaving peptidic bonds preferentially close to aromatic residues. The active site contains a catalytic zinc ion bound to two histidines (H142 and H146) and one glutamate (E166) residue, and an additional catalytic glutamate (E143) residue [76,77]. It is located at the bottom of a pocket formed by the two protein domains: the N-terminal domain composed mostly of β-sheets containing the conserved HEXXH sequence motif with the corresponding H142, E143 and H146; the C-terminal domain composed mostly of α-helices, where E166 is located. Given our interest in analyzing and comparing the dynamical properties of MPs, we first characterized the dynamical features of thermolysin, particularly those that are functionally relevant. In order to do this, we used the ProDy software, and the type of results produced can be found in Fig. 1 and Fig. 2.

Several studies of thermolysin have been made in the last decades, with multiple crystal structures obtained in different conditions available in the PDB. These structures provide snapshots of the motions undergone by the protein upon interaction with different molecules, including substrate-analogs and inhibitors, thus giving the opportunity to describe the conformational subspace related with its function [78–80]. For this purpose, we applied PCA to a set of 112 thermolysin crystal structures in order to characterize the collective modes of atomic displacements (see Methods section for details on PCA calculations) [78,81]. As shown in Table 1, a small set of PCs describes the majority of atomic positional variations occurring in the structural set, with the first six PCs (PC1-PC6) describing 80% of the total variance. PC1 alone accounts for 35% of the variance, with a cumulative displacement of < 5 Å. The respective structural variation for each residue along PC1 is shown in Fig. 1A and 1B. Although these structural differences are small (0.60±0.08 Å), they are nonetheless highly collective, affecting approximately half of the atoms in the protein ($0.43 \leq \kappa \leq 0.69$). Higher fluctuations are seen for the N-terminal region (including the surface residues 42-62). Regions 105-117 and 210-230, which form the top of the active site pocket and contain residues involved in substrate binding [82], also exhibit high, anti-correlated variations (Fig 1A inset). Conversely, the pocket bottom region where the zinc binding residues are located shows relatively lower variation, with the exception of the catalytic residue E143. The structural fluctuations seen for this catalytic residue towards the pocket bottom reflects local accommodations of the structure to the presence of different ligand molecules in the active site [65, 83]. Projection of each structure onto the subspace spanned by PC1-PC2 is shown in Fig. 1C. Two distinct clusters corresponding to the subsets of bound and unbound forms are obtained, as seen by the distribution of structures along the PC1 axis (P-value in Table 1). This indicates that the presence of molecules in the active site is associated with conformational changes in the protein that are effectively described by PC1.The variations described by PC1 point to an opening-closure movement of the active site pocket, *i.e.* an hinge-bending motion of N- and C-terminal domains with the vertex at the pocket bottom. Large scale, hinge-bending displacements were initially described for thermolysin and related neutral proteases [84]. The correspondence with hinge-bending motions was made by analysis of short MD trajectories and later confirmed by the reported unbound crystallographic structure of thermolysin [83,85]. Therein, the detected hinge-bending motions were related to transitions between “open” and “closed” conformations of the unbound and bound forms of the protein, respectively, indicating their functional role. Therefore, the current results show that PC1 alone can describe to a reasonable extent the functionally-relevant conformational changes of thermolysin, with motions in the positive direction along the PC1 axis describing an “opening” of the active site pocket and in the negative direction to its “closure”.

**Table 1: Fraction of variance and collectivity of the first 6 PC, ED and NM modes obtained for thermolysin.** The collectivity degree is expressed as κ and reflects the portion of atoms in the structure affected by a given mode.

*Anderson-Darling normality test for the projection of structures along PC modes.

| PC modes | | | ED modes Sim1/Sim2 | | | NM modes | | |
|---|---|---|---|---|---|---|---|---|
| | Fraction of Variance (P-value)* | κ | | Fraction of Variance | κ | | Fraction of Variance | κ |
| PC1 | 0.35 ($2.2x10^{-16}$) | 0.69 | ED1 | 0.23 /0.18 | 0.46/0.52 | NM1 | 0.18 | 0.64 |
| PC2 | 0.23 ($5.9x10^{-11}$) | 0.55 | ED2 | 0.1/0.08 | 0.62/0.33 | NM2 | 0.14 | 0.63 |
| PC3 | 0.13 ($1.8x10^{-4}$) | 0.61 | ED3 | 0.07/0.04 | 0.65/0.61 | NM3 | 0.09 | 0.64 |
| PC4 | 0.04 ($8.0x10^{-2}$) | 0.43 | ED4 | 0.05/0.04 | 0.63/0.26 | NM4 | 0.06 | 0.13 |
| PC5 | 0.03 ($1.0x10^{-2}$) | 0.54 | ED5 | 0.04/0.03 | 0.35/0.61 | NM5 | 0.05 | 0.09 |
| PC6 | 0.02 ($4.1x10^{-3}$) | 0.56 | ED6 | 0.03/0.03 | 0.71/0.4 | NM6 | 0.05 | 0.38 |

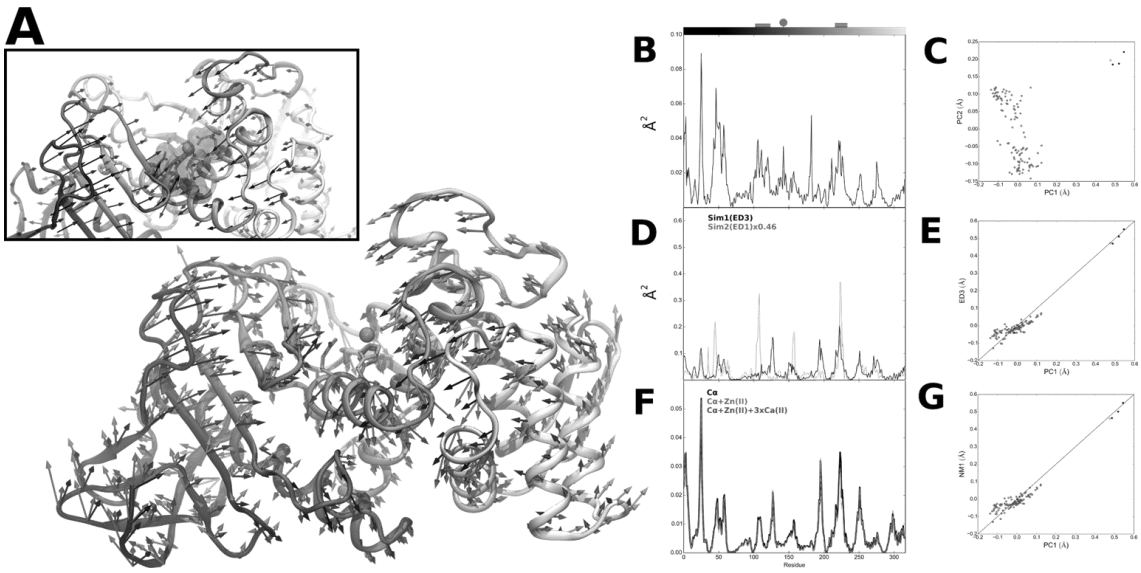



**Fig. 1 – Thermolysin PC, ED and NM modes.** (A) Visual representation of residue-level PC1 mode vectors (black, scale 2.05), NM1 mode vectors (green, scale 1.47) and ED3 mode vectors from Sim1 (red, scale 3.17). Inset: Details of active site region. Colored ribbons corresponding to the top of the active site pocket (residues 105-177 in blue and 210-230 in green). Active site residues H142, H146 and E166 in green and catalytic E143 in cyan sticks representation. (B) Square fluctuations as a function of residue index (1.316) obtained for PC1. Top bars correspond to residue coloring as presented in (A), with catalytic E143 represented as cyan circle. (C) Projection of thermolysin crystal structures along PC1 and PC2. Structures were grouped into bound (red) and unbound (black) groups; unbound reference structure (PDB ID: 1L3F) used in PCA, EDA and NM calculations (green). (D) Square fluctuations as a function of residue index obtained for ED3 from Sim1 and ED1 from Sim2. (E) Cross-projection of crystal structures along PC1 and ED3 from Sim1 (r=0.94). (F) Square fluctuations as a function of residue index obtained for NM1 from calculated thermolysin ANM and respective variants. (G) Cross-projection of crystal structures along PC1 and NM1 (r=0.95).

In order to test if the conformational changes described by the PCA are intrinsic, *i.e.* if they are encoded in the internal dynamics of thermolysin, we performed EDA on snapshot conformations from two replicates (Sim1 and Sim2) of a 20 ns MD simulation trajectory of the protein in its unbound state (see Methods section for details) [86,87]. A similar analysis has been previously reported for other proteases of different clans (with distinct catalytic chemistry) [79,88]. EDA is focused on the subspace of PCs, typically the top-ranking modes, that describe the majority of collective atomic motions along a simulation trajectory, and which are typically related to protein

function. As indicated in Table 1, the first six EDA-derived PCs (ED1-ED6) have lower variance values than the corresponding structurally-derived PCs (PC1-PC6), with ED1-ED6 of Sim1 accounting for only 52% of the total variance (40% for Sim2). However, in terms of collectivity degrees, EDs and PCs are similar ($0.35 \leq \kappa \leq 0.71$ for Sim1 and $0.33 \leq \kappa \leq 0.61$ for Sim2), reflecting the collective nature of motions described by ED1-ED6 modes. Although conformational changes related to ligand-binding typically occur at longer time-scales, EDA of MD simulations of a few ns generally provides a reasonable description of the full conformational space explored by the protein [70, 71]. During the total simulation time analyzed (19 ns after RMSD convergence), the sampled subspaces are convergent for each simulation. This is shown by comparing the covariance of residue fluctuations between two time intervals, the interval of the initial 11.4 ns and the full time interval. The overlap of covariance matrices obtained is 0.63 for Sim1 and 0.65 for Sim2. Moreover, the overlap between subspaces explored during the two time intervals yields a RMSIP of 0.93 for both Sim1 and Sim2. When comparing Sim1 against Sim2, the sampled conformational subspaces defined by the respective ED1-ED6 modes are similar, with a RMSIP of 0.79. However, the overlap between the respective covariance matrices is relatively low (0.38). These results indicate that the two simulations exhibit similar dynamical behavior during the analyzed time window, although the sampled conformational space explored are distinct, as it has been reported for other proteins [89,90].

While it remains uncertain if the protein explores a distinct conformational space on longer time scales, which would require simulation times from μs to ms or other methods more suitable to characterize the protein potential energy surface, the remaining analysis is focused on the correspondence between the results obtained from the current EDA with the experimentally-derived PCA of thermolysin. Therefore, in order to compare the obtained EDs with structurally-derived PCs, the overlap between each of the respective first six modes was calculated. As shown is Fig. 2A, the highest value was obtained between modes ED3 and PC1 from Sim1, with an overlap of 0.72 (for Sim2 an overlap of 0.71 was obtained between ED1 and PC1). The large overlap between ED3 and PC1 from Sim1 translates in similar directions of residue fluctuations shown in Fig. 1A and 1D, particularly in the region comprising the active site pocket. Further confirmation of the high similarity between these modes is obtained in Fig. 1E, where cross-projection of crystal structures along PC1 and ED3 from Sim1 yields a distribution along the diagonal with a clear separation between bound and unbound subsets. Indeed, the large cumulative overlap of 0.85 between ED1-ED3 from Sim1 and PC1 (0.80 for ED1-ED3 and PC1 of Sim2) show that functionally-relevant conformational changes are effectively captured by the first three ED modes and therefore can be said to be encoded in the internal dynamics of the protein. However, there is low correspondence between subspaces defined by the respective modes, as given by the low RMSIP value of 0.49 for Sim1 and 0.46 for Sim2 between ED1-ED3 and PC1 (for ED1-ED10 and PC1-PC10 the RMSIP is 0.56 for Sim1 and 0.53 for Sim2) [70,91].The obtained results show that a high conformational space is sampled during MD simulations, which includes a functionally-relevant subspace (particularly the ones described by ED3 from Sim1 or ED1 from Sim2) that is only explored upon interaction with the ligand (as described by PC1). This is in line with conformational selection models of protein function described elsewhere [80,92].

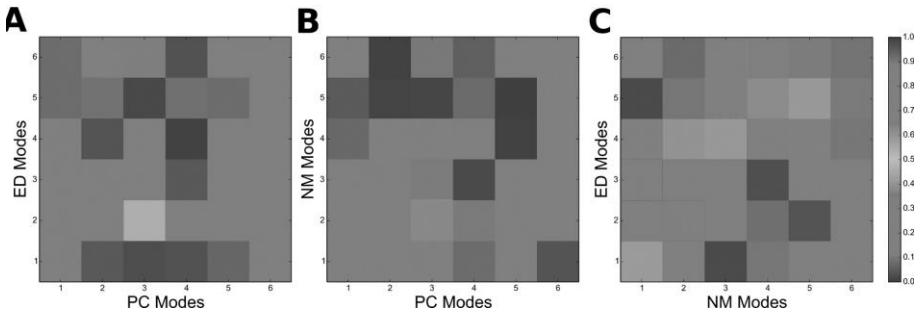



**Fig. 2 – Comparison of the first six PC, ED and NM modes.** (A) Overlap between PC and ED modes from Sim1; (B) Overlap between PC and NM modes and; (C) Overlap between NM and ED modes from Sim1.

ENMs have been extensively evaluated against experimental and computational benchmarks [73,93–101]. They have also been used previously to describe functional aspects of MP internal

dynamics, although the relation between collective motions of MPs and their function was only indirectly established [50]. Specifically, it was shown that Atrolysin E and other non-metallo proteases have similar dynamical profiles in the respective active site regions. However, regions close to metal-centers and active sites tend to exhibit similar and relatively restrained, dynamical profiles [102,103]. Since MPs contain a metal-center in their active site, the functional relevance of the respective ENM-predicted dynamical profiles should be more thoroughly addressed. For this purpose, we first tested the suitability of using coarse-grained ENMs in reproducing the internal dynamics of thermolysin. The ANM implemented in the ProDy software was employed to the unbound structure of thermolysin and the space of collective motions was characterized by Normal Mode Analysis (NMA) [54,93,104–106].

The ANM employs a $C^{\alpha}$-based node representation of protein structure. Given the presence of metal ions in clan MA members, the addition of node representations for metal ions was also evaluated for thermolysin, using a similar approach to the one made for another ENM [103]. Variations of thermolysin ANM were generated to represent the zinc ion found at the active site and also four additional calcium ions found at the N-terminal domain. While the later are crucial for the thermostability of the protein, the zinc ion is not considered to play a crucial structural role, since its metal-substituted forms present very similar tertiary structures [65]. Fig. 1F shows that the respective residue square fluctuations profiles of the model variants produced are almost identical to the $C^{\alpha}$-only ANM ($r > 0.995$ for both variants), with only slight variations on fluctuation amplitudes in the 220-226 region, at the top of the active site pocket. The close similarity between the two variants indicates that the introduction of calcium nodes does not produce significant changes in the global dynamical profile of the protein, but the inclusion of the zinc node alone produces small local changes in the dynamic profile of the active site pocket. Since the calcium nodes are introduced at a highly clustered region of the network of inter-residue contacts, its topology is mostly unchanged. The introduction of the zinc node, on the other hand, leads to new network connections in the active site region, producing slight variations on its topology that nonetheless result in very similar dynamical profiles. Therefore, the inclusion of additional nodes for metal ion representation produces only local changes on the ANM-derived dynamical profile of thermolysin, the clan MA type protein.

The results above are in line with recent studies where the employment of coarse-grained models was evaluated for globular folds [21, 36], such as the one of thermolysin. It was found that the respective ENMs can effectively capture their essential dynamics [21] and that these models are robust to local perturbations [36]. Given the conserved tertiary structure of clan MA members, the accuracy of the respective ENMs in describing their internal dynamics is not expected to greatly increase with the inclusion of additional metal ion nodes. We therefore focused the remaining analysis on the results obtained for the $C^{\alpha}$-based ANM of thermolysin ($N$=316), particularly on the subset of NMA-derived low-frequency NM modes (NM1-NM6).

As shown in Table 1, the variance of NM1-NM6 is significantly lower than the corresponding PCs and similar to ED modes, accounting for 57% of the total variance. In particular, NM1 accounts for only 18% of the total variance, similar to ED1 from Sim1 and Sim2. In terms of collectivity degrees, the three lowest-frequency modes (NM1-NM3) are highly collective ($0.63 \leq \kappa \leq 0.64$), in the range of PCs and EDs, and the remaining modes (NM4-NM6) have significantly lower values ($0.09 \leq \kappa \leq 0.38$). The overlap between NM modes and PCs is shown in Fig. 2B. Remarkably, a large overlap of 0.71 is found between NM1 and PC1. Fig. 1A shows the close correspondence between NM1 mode directions and PC1. Again, a high similarity of motions is observed in the region comprising the active site pocket. This similarity is also reflected in terms of the corresponding profiles of residue fluctuations in Fig. 1F. Cross-projection of the structural set along NM1 and PC1 shown in Fig. 1G further confirms the close correspondence between these modes. This indicates that the lowest-frequency mode (NM1) predicted by the ANM can effectively reproduce the functionally-relevant conformational change described by PC1 [68]. Indeed, structural variations described by the first PCs are well covered by the low-frequency modes, with a large cumulative overlap of 0.81 between NM1-NM3 and PC1 and a RMSIP of 0.66 between subspaces defined by NM1-NM3 and PC1-PC3 (for NM1-NM10 and PC1-PC10 the RMSIP is 0.58).

Comparison between NMs and EDs was also made by calculating the overlap between corresponding modes. As it can be seen in Fig. 2C, there is significant overlap between NM modes and EDs from Sim1, with a particularly large overlap of 0.82 between NM1 and ED3 (0.69 between

ED1 and NM1 from Sim2). Remarkably, a high RMSIP of 0.72 is also obtained between subspaces defined by NM1-NM3 and ED1-ED3 for both Sim1 and Sim2 (for NM1-NM10 and ED1-ED10 the RMSIP is 0.76 for Sim1 and 0.74 for Sim2). Therefore, it can be said that the employed coarse-grained ANM can reproduce fairly well the conformational space explored by thermolysin during the two independent MD simulations.

In conclusion, the results show that both MD simulations and ANM provide a reasonable description of thermolysin internal dynamics, particularly the subspace of collective motions with functional relevance. This prompted the use of ENM-based methods to study other evolutionarily-related MPs and to quantitatively compare their internal dynamics. The obtained results will be discussed in following section.

## Structural and Dynamical Alignments

Structure- and dynamics-based alignments of a representative set of 13 MA clan proteins (Table 2) was made using the DaliLite and ALADYN web-servers, respectively, as described in Methods section. The employed algorithm in ALADYN is based on the β-Gaussian ENM, where a two-nodes per residue representation is used. Using this approach, the inclusion of additional metal ion nodes could not be evaluated as in the previous section for thermolysin ANM. However, given the structural heterogeneity of clan representatives, the difference between their respective dynamical profiles is expected to surpass the local changes produced in each protein by the inclusion of metal nodes.

The scores obtained for each of the resulting 78 aligned pairs are shown in Fig. 3 and Table B in S1 File, being classified into three distinct groups: 1) MA(E), with both proteins belonging to the MA(E) subclan; 2) MA(M), for proteins belonging to the MA(M) subclan and; 3) Mixed, with each protein belonging to different subclans. Given that each representative protein is not a homologue of the remaining representative proteins [2], the simplified mapping of scores shown in Fig. 3A can be seen as an additional layer to the typical representation of protein relatedness in terms of sequence similarity, providing a characterization of the clan in terms of structural and dynamical diversity. As it will be discussed below, it allows for the identification of unnoticed functional similarities between distinct proteins and provides a description of how structure and internal dynamics of proteins are related within a given clan. It can be useful, *e.g.* in the field of structural genomics, where protein function assignment could be made based not only on sequence and structural similarity, but also on information obtained from dynamics-based alignments with a set of proteins with known catalytic function.

The threshold values to consider a pair either structurally or dynamically similar are based on the employed methods. Overall, the majority of pairs analyzed are structurally similar; with 87% of pairs having Z-scores > 2.0 [72]. However, 69% of pairs are not dynamically similar, since they have P-values > 0.02 [51]. Regarding the structural conservation of the three types of pairs considered, all 21 MA(M) pairs are structurally similar while for MA(E) pairs this is the case for 87% of the 15 pairs. In the case of mixed pairs, 81% of the 42 are structurally similar, although with Z-scores clustered near the threshold of structural similarity. Regarding the internal dynamics of subclan members, remarkably, only 14% of MA(M) pairs have dynamical similarity, while 53% of MA(E) pairs also present it. In the case of mixed pairs, only 17% have dynamical similarity. Since the internal dynamics of proteins are ultimately dependent on their structure, structural similarity is expected to be related with dynamical similarity, *i.e.* pairs with higher structural similarity scores would tend to have higher dynamical similarity scores. However, no correlation is apparent between Z-scores and P-values for all analyzed pairs, although it is noted that pairs with Z-scores > 19.0 are associated with a high dynamical similarity scores (P-value < 0.001), corresponding to MA(E) pairs M3-M2, M3-M32 and M32-M2. As it will be discussed below, MA(E) pairs M32-M8 and M32-M27 are particularly relevant since their respective alignments have Z-Scores < 2.0 and P-values < 0.001. In the case of MA(M) pairs, there is an apparent inverted relation between dynamical and structural similarities, since pairs with lower structural similarity (lower Z-scores) show higher dynamical similarity (lower P-values), *e.g.* pair members M35-M12(B), M35-M12(A), M8-M10(A).

**Table 2 - Set of 13 representative metalloproteases of the MA Clan.**

| MEROPS Sub-clan | MEROPS Family | Protein | PDB ID | CATH Superfamily | SCOP Family |
|---|---|---|---|---|---|
| MA(E) | M4 | Thermolysin | 1L3F | 1.10.390.10; 3.10.170.10 | 55490 |
| MA(E) | M1 | Leukotriene A4 hydrolase | 1H19 | 1.10.390.10 | 64338 |
| MA(E) | M3 | Neurolysin | 1I1I | 1.10.1370.10; 3.40.390.10; 1.20.1050.40 | 55505 |
| MA(E) | M32 | Carboxypeptidase *Pfu* | 1KA4 | n.a. | 55505 |
| MA(E) | M2 | Angiotensin Converting Enzyme | 1O8A | n.a. | 55505 |
| MA(E) | M27 | Botulinum neurotoxin type A | 3BON | n.a. | 55512 |
| MA(M) | M7 | Extracellular small neutral protease | 1C7K | 3.40.390.10 | 55487 |
| MA(M) | M35 | Peptidyl-Lys metalloendopeptidase | 1G12 | 3.40.390.10 | 55487 |
| MA(M) | M8 | Leishmanolysin | 1LML | 3.90.132.10; 3.10.170.20; 2.30.34.10; 2.10.55.10 | 55499 |
| MA(M) | M10(A) | Interstitial collagenase | 1CGE | 3.40.390.10 | 55528 |
| MA(M) | M10(B) | Serralysin | 1AKL | n.a.; 2.150.10.10 | 55508 |
| MA(M) | M12(A) | Astacin | 1AST | 3.40.390.10 | 55516 |
| MA(M) | M12(B) | Snake venom metalloproteinase adamalysin-2 | 1IAG | 3.40.390.10 | 55519 |

A similar analysis was made by Carnevale *et al.*, where no general trend between structural and dynamical similarity scores for pairs of proteases from different clans and catalytic chemistries was found [6]. In that case, no specific correlation was expected since the analyzed representatives were considered to have different evolutionary origin and therefore share minimal structural homology. Nonetheless, structural and dynamical similarity was identified for some pairs, with the authors arguing for convergent evolutionary pressure to be at play in such cases. It was therefore suggested that in cases where structurally variability is observed, a compensatory mechanism for dynamical conservation could maintain the catalytic capacity of the proteins. Conversely, since in this study MP representatives are from the same clan and therefore have assumed common ancestry, no structural or dynamical similarity between two representatives is expected to reflect a divergent evolutionary process. Given that structural and dynamical similarities are apparently not correlated in this case, it is suggested that selective pressure is acting independently on structure and internal dynamics.

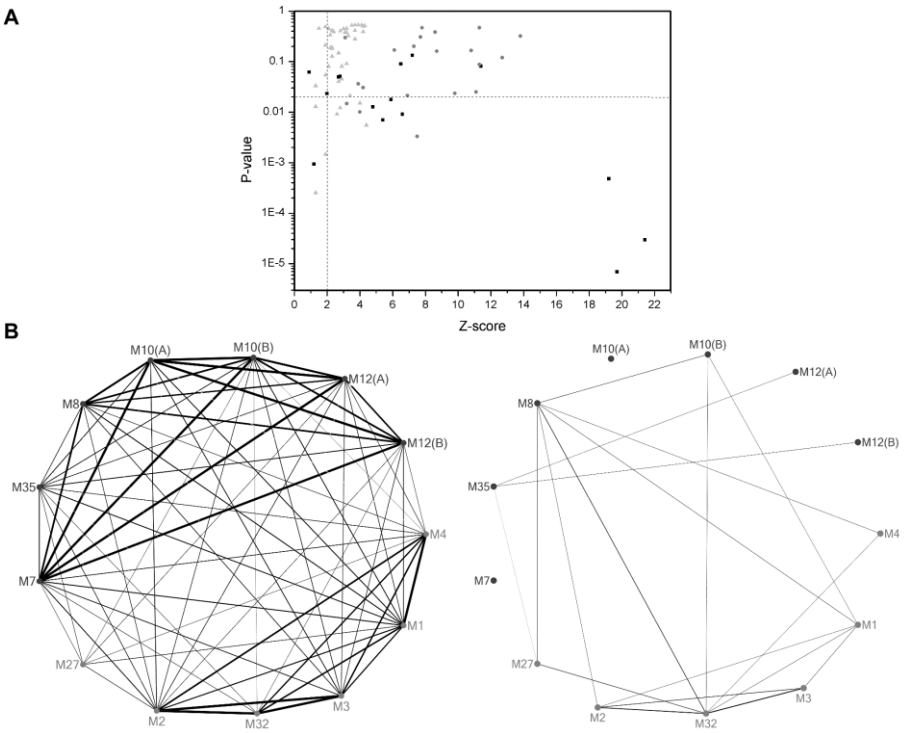



**Fig. 3 – Dynamical and structural variability of MA clan representative structures.** (A) Mapping of structural (Z-score) and dynamical (P-values) alignment scores obtained for the pairs of MP structures. Red circles: Metzincin pairs; Black squares: Gluzincin pairs; Gray triangles: Mixed pairs. Dashed lines indicate threshold values for structural (vertical) and dynamical (horizontal) similarity. Labeled pairs were selected for further inspection. (B) Graph representation of the structural (left) and dynamical (right) similarity between MP representatives (Blue: Gluzincins; Green: Metzincins). Edge width is proportional to the corresponding Z-score and P-value and only edges with width values above the corresponding thresholds are represented.

Fig. 3B provides an overview of how individual representatives are related in terms of structural and dynamical similarity. Only pairs considered to be structurally- or dynamically- similar are linked in the weighed graphs, where edge width is related to Z-score and P-value. In terms of structural similarity, members are more connected to representatives belonging to the same subclan, but overall there are also inter-subclan connections, reflecting the structural conservation across the entire clan. There is one exception for representative M27, *Botulinum* Neurotoxin Type A Light Chain, which shows low connectivity to the remaining representatives.

Regarding dynamical similarity, there is less connectivity between representatives, which may reflect divergence of internal dynamics along the clan. Also, subclan similarity is less pronounced as it can be seen for the case of, *e.g.* representative M8, which has higher connectivity with MA(E) subclan members than with those of MA(M) subclan where it belongs to. Notable exceptions are representatives Extracellular Small Neutral Protease (M7) and Interstitial Collagenase (M10(A)) that share no dynamical similarity with the remaining MP representatives. These two proteins are relatively small in size; Extracellular Small Neutral Protease is a 132-residue long protein and the structure of Interstitial Collagenase corresponds only to the 168-residue long catalytic domain. Although their internal dynamics are also characterized by hinge-bending motions (not shown), the amplitude of motions of their smaller domains has no correspondence with the motions of the larger domains of the remaining representatives.

These results provide a quantitative measure of the structural and dynamical similarity that characterizes MA clan members and provides a “horizontal” view on MP evolution. In order to understand if there is a functional basis for the conservation or divergence of such features, a chosen set of alignments was further inspected. For pairs with high structural and dynamical similarity M32-M2, M3-M32 and M3-M2, the structural homology between the corresponding representatives Angiotensin Converting Enzyme (M2), Neurolysin (M3) and Carboxypeptidase *Pfu* (M32) had been previously reported [107-109]. Indeed, these proteins are grouped in the same

SCOP family and their high dynamical similarity can be related directly to their structural resemblance. Nonetheless, Angiotensin Converting Enzyme and Carboxypeptidase *Pfu* have no attributed CATH codes and Neurolysin presents three domains, including the common 3.40.390.10 domain found in other representatives. The alignments of Neurolysin and Carboxypeptidase *Pfu*, which represent the pair with highest dynamical similarity, were chosen for further inspection. Neurolysin is a 681 residue-long endopeptidase that cleaves the 13-residue peptide neurotensin but it has also activity towards a diverse set of oligopeptide sequences [107]. Carboxypeptidase *Pfu* is a 499-residue long thermostable carboxypeptidase homodimer with broad substrate specificity [108].

Both structures are mainly α-helical in content and characterized by a deep narrow channel that divide the structures into two domains, with a wider opening at one end and with the active site located at the bottom. This prevents activity towards large, folded substrates. Their structure-based alignment is shown in Fig. 4A. The alignment produces a RMSD of 3.7 Å for 449 amino acids used with 15% sequence identity. Aligned regions consist on core regions surrounding the active sites and α-helices that constitute the channel base and walls. The respective dynamics-based alignment, shown in Fig. 4B, produces a lower number of equivalent residues, with RMSD of 2.8 Å for 371 amino acids with 13.5% sequence identity (RMSIP of 0.870). It reveals that these regions undergo very similar deformations resembling hinge motions, most likely corresponding to channel opening for substrate access. The aligned portions in both structure- and dynamics-based alignments have high degree of identity.

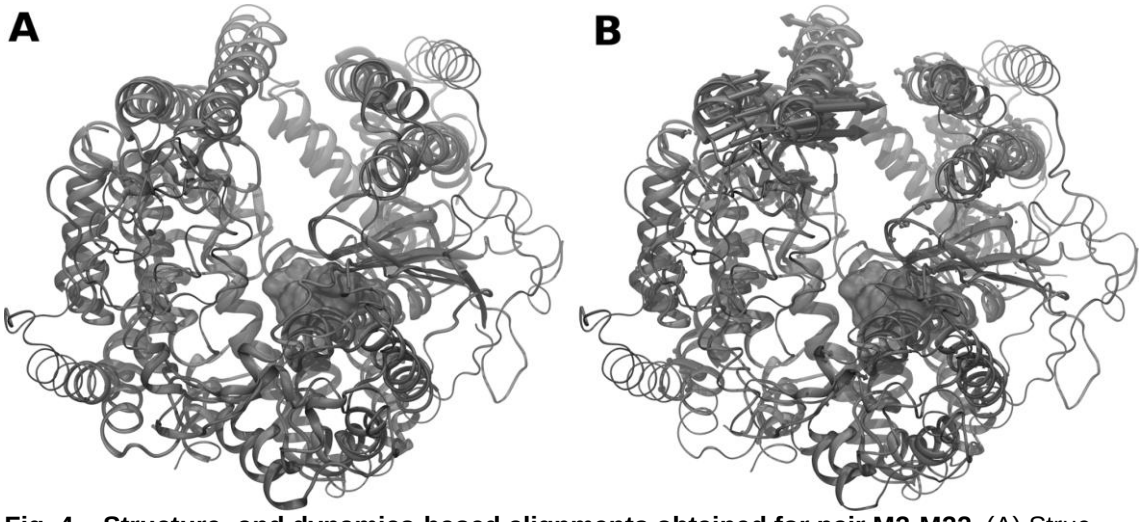



**Fig. 4 – Structure- and dynamics-based alignments obtained for pair M3-M32.** (A) Structure-based and (B) Dynamics-based alignment of M3 representative Neurolysin (blue, PDB ID: 1I1I) and M32 representative Carboxypeptidase *Pfu* (red, PDB ID: 1KA4). Produced alignments were obtained using the DaliLite and ALADYN web-servers (see Methods). Aligned residues colored in cartoon representation, non-aligned residues in colored ribbons and active site residues in surface representations (Neurolysin: H474, E475, H478 and E503; Carboxypeptidase *Pfu* H269, E270, H273 and E299). Colored arrows indicate modes of motion of aligned portions along the first mode.

The motions produced by the employed β-Gaussian ENM are in agreement with findings reported for each MP. Indeed, hinge movements of Carboxypeptidase *Pfu* sub-domains were previously argued to be involved in channel closure and this conformational change was proposed to have a functional role, namely to extend the number of interactions with the substrate [108]. In the case of Neurolysin, it was suggested that both domains are rigid but present some mobility in relation to each other due to looser packing at the base of the channel [107]. In both cases, the amplitude of hinge movements is restricted due to the presence of cross-domain segments that tighten the channel at one end and limit the access of longer substrates to the buried active site. Remarkably, Carboxypeptidase *Pfu* cross-domain α-helix ($\alpha_4$, residues 81-99), that contains the conserved R92 considered to have a crucial role in substrate-length restriction, is almost perfectly

aligned with an equivalent cross-domain α-helix of Neurolysin ($\alpha_5$, residues 137-152) in both structure- and dynamics-based alignments. Moreover, it is noted that Neurolysin residue Lys148 and the conserved Arg92 of Carboxypeptidase *Pfu* are very close positioned in the dynamics-based alignment of the respective structures, that together with sharing the chemical nature suggests their functional equivalence (Figure B in S2 File). Therefore, the conservation of internal dynamics between these two MPs may not be only a direct consequence of their structural similarity, but may also have a functional basis. Specifically, the requirement to have specific channel opening amplitudes in order to restrain substrate-length, while more local variations in structure allow for distinct specificities in terms of substrate sequence recognition and binding.

Further support for the conservation of internal dynamics between these two MPs comes from the findings reported for Angiotensin Converting Enzyme, which shares high structural and dynamical similarity with both Neurolysin and Carboxypeptidase *Pfu* [110,111]. The authors showed that for this MP hinge-bending motions have a functional role, since channel opening allows for substrate access to the active site. Furthermore, it was shown that these motions are conserved between other M2 family members. Finally, the presumed dynamical resemblance between M2 family members and another member of the M3 family, Carboxypeptidase *Dcp*, has been previously noted [112]. Together with these findings, results obtained for M2, M3 and M32 representatives suggest that conservation of internal dynamics is not limited to homologues, but that it can be extended for other families that share some structural similarity.

Pairs M32-M8, M27-M8 and M32-M27 were also analyzed due to their high dynamical similarity with no structural similarity. Both Carboxypeptidase *Pfu* (M32) and *Botulinum* Neurotoxin Type A (M27) belong to the Gluzincin subclan but have no attributed CATH numbering. Leishmanolysin (M8) belongs to the Metzincin subclan and has four CATH domains, with the particular feature of lacking the conserved 3.40.390.10 domain shared by other Metzincin subclan representatives. The pair of Leishmanolysin and Carboxypeptidase *Pfu* was chosen for further inspection, as it corresponds to the only Mixed pair with very high dynamical similarity. Leishmanolysin (also termed GP63) is a 478 residue-long surface glycoprotein from *Leishmania major* that occurs as a dimer and has activity towards a wide variety of peptidic substrates [113]. It adopts a predominantly compact fold composed of mostly β-sheet secondary structure elements, in contrast with the predominant α-helical structure adopted by Carboxypeptidase *Pfu* [114]. The structure differs from Carobxypeptidase *Pfu* and other MPs as it is composed of three domains: the N-terminal domain that contains the active site HEXXH sequence motif; the central domain, that presents an unique 62 residue-long insertion between the conserved Glycine of the HEXXHXXGXXH/D metzincin sequence motif and the third active site residue Histidine; and the C-terminal domain, which contains the membrane anchoring point and is composed mainly of β-strand and random coil elements, being positioned at one end of the active site cleft base. Both N- and central domain form the active site cleft. EDA of Leishmanolysin MD trajectories revealed that large-scale, collective motions dominate the conformational changes explored by the protein. These consist in hinge-bending motions characterized by rigid body movements of the N-terminal domain relative to the central and C-terminal domains, with the hinge axis located at base of the active site cleft [115]. The corresponding structure-based alignment is shown in Fig. 5A, with a RMSD of 4.6 Å for 88 amino acids used with 9% sequence identity. Unlike the pairs with high structural and dynamical similarity with aligned regions spanning almost the entirety of the structures, the aligned portions of Leishmanolysin and Carboxypeptidase *Pfu* are restricted to the N-terminal domain, including the α-helix containing the active site HEXXH sequence motif. Indeed, this was typically observed for other non-structurally similar pairs of MPs [1]. The overall orientation of the structures is kept, with the substrate-binding pockets and active residues being almost identically positioned in both structures and with the positioning of the C-terminal domain of Leishmanolysin at the more open end of Carboxypeptidase *Pfu* channel. Dynamics-based alignment shown in Fig. 5B produces a slightly higher number of equivalent residues, with an RMSD of 3.3 Å for 91 amino acids with 7.7% sequence identity (RMSIP of 0.832). Remarkably, the dynamics-based alignment results in a complete horizontal rotation of Leishmanolysin in relation to Carboxypeptidase *Pfu* when compared to the structure-based alignment. This results in the positioning of its C-terminal domain at the cross-domain segments that constitute the more closed end of Carboxypeptidase *Pfu* channel and the active sites become approximately 10 Å apart from each other. Nonetheless, the substrate binding-pocket of both MPs retain their orientation and regions undergoing hinge-bending motions are almost identically positioned, indicating their dynamical

equivalence. This is also observed in the structure-based and dynamics-based alignments produced for Leishmanolysin and *Botulinum* Neurotoxin Type A (data not shown), thus suggesting that there is conservation of internal dynamics even when remarkable structural differences between members occur. Moreover, this conservation has a functional basis, namely to allow for proper orientation of interactions between the proteins and their substrates. Therefore, the relatedness of Leishmanolysin, *Botulinum* Neurotoxin Type A and Carboxypeptidase *Pfu,* which has so far been only considered at the clan level, becomes evident when their internal dynamics are taken into account.

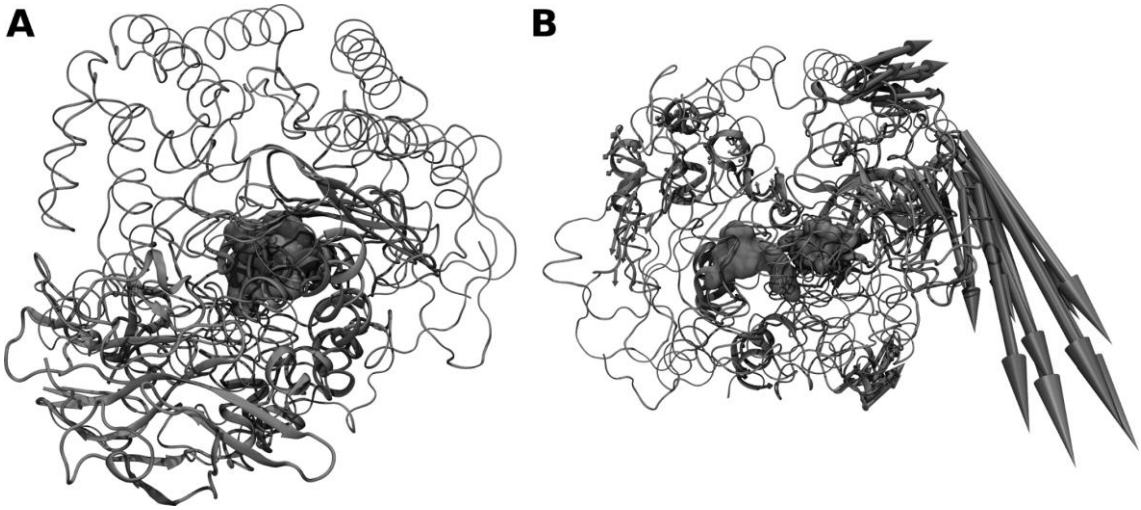



**Fig. 5 – Structure- and dynamics-based alignment obtained for pair M32-M8.** (A) Structure-based and (B) Dynamics-based alignment of M8 representative Leishmanolysin (green, PDB ID: 1LML) and M32 representative Carboxypeptidase *Pfu* (red, PDB ID: 1KA4). Produced alignments were obtained using the DaliLite and ALADYN web-servers (see Methods). Aligned residues colored in cartoon representation, non-aligned residues in colored ribbons and active site residues in surface representations (Leishmanolysin: H264, E265, H268, H334 and; Carboxypeptidase Pfu H269, E270, H273 and E299). Colored arrows indicate modes of motion of aligned portions along the first mode. Leishmanolysin C-terminal domain in cyan colored tube representation.

# Conclusions

This work provides a quantitative characterization of the structural and dynamical diversity occurring within the MEROPS MA clan of MPs. It shows that metalloproteases of this clan have distinct dynamical profiles despite their overall structural similarity. Also, it is shown that in cases where high dynamical similarity is observed, the predominant modes correspond to hinge-bending motions associated with substrate-binding. These motions are functionally relevant and appear to be conserved in the clan even when remarkable structural differences between its members occur. Therefore, besides providing a description of the structural and dynamical features of a set of proteins, this type of analysis can also provide new insights on enzyme function that remained unnoticed so far. For metalloproteases, it is suggested that the need to maintain proper substrate interactions has an important role on the conservation of their internal dynamics. Therefore, the type of interactions between a protein and its ligand and the associated motions should be more carefully considered when comparing the internal dynamics of a diverse set of functionally distinct proteins. Together, this work contributes to the development of simple and effective approaches that incorporate quantitative analysis of dynamical similarity between proteins to study the evolution of metalloprotease internal dynamics and the factors governing them.

# References

1. Barrett AJ. Handbook of Proteolytic Enzymes. 3rd ed. Handbook of Proteolytic Enzymes. Elsevier; 2013. pp. 325–370. doi:10.1016/B978-0-12-382219-2.00077-6

2. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2014;42: D503–9. doi:10.1093/nar/gkt953
3. Murzin A, Brenner S. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247: 536–540. doi:10.1016/S0022-2836(05)80134-2
4. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, et al. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res. 2013;41: D490–8. doi:10.1093/nar/gks1211
5. Tyndall J, Nall T, Fairlie D. Proteases universally recognize beta strands in their active sites. Chem Rev. 2005;105: 973–1000. doi:10.1021/cr040669e
6. Carnevale V, Raugei S, Micheletti C, Carloni P. Convergent Dynamics in the Protease Enzymatic Superfamily. 2006; 9766–9772.
7. Gagné D, Doucet N. Structural and functional importance of local and global conformational fluctuations in the RNase A superfamily. FEBS J. 2013;280: 5596–607. doi:10.1111/febs.12371
8. García-Meseguer R, Martí S, Ruiz-Pernía JJ, Moliner V, Tuñón I. Studying the role of protein dynamics in an SN2 enzyme reaction using free-energy surfaces and solvent coordinates. Nat Chem. Nature Publishing Group; 2013;5: 566–71. doi:10.1038/nchem.1660
9. Hammes-Schiffer S, Benkovic SJ. Relating protein motion to catalysis. Annu Rev Biochem. 2006;75: 519–41. doi:10.1146/annurev.biochem.75.103004.142800
10. McGowan LC, Hamelberg D. Conformational plasticity of an enzyme during catalysis: intricate coupling between cyclophilin A dynamics and substrate turnover. Biophys J. Biophysical Society; 2013;104: 216–26. doi:10.1016/j.bpj.2012.11.3815
11. Luk LYP, Javier Ruiz-Pernía J, Dawson WM, Roca M, Loveridge EJ, Glowacki DR, et al. Unraveling the role of protein dynamics in dihydrofolate reductase catalysis. Proc Natl Acad Sci U S A. 2013;110: 16344–9. doi:10.1073/pnas.1312437110
12. Glowacki D, Harvey J, Mulholland A. Taking Ockham's razor to enzyme dynamics and catalysis. Nat Chem. 2012;4: 169–176. doi:10.1038/NCHEM.1244
13. Hammes-Schiffer S. Catalytic efficiency of enzymes: a theoretical analysis. Biochemistry. 2013;52: 2012–20. doi:10.1021/bi301515j
14. Henzler-Wildman K a, Lei M, Thai V, Kerns SJ, Karplus M, Kern D. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature. 2007;450: 913–6. doi:10.1038/nature06407
15. Jones AR, Levy C, Hay S, Scrutton NS. Relating localized protein motions to the reaction coordinate in coenzyme $B_{II}$-dependent enzymes. FEBS J. 2013;280: 2997–3008. doi:10.1111/febs.12223
16. Hay S, Scrutton NS. Good vibrations in enzyme-catalysed reactions. Nat Chem. Nature Publishing Group; 2012;4: 161–8. doi:10.1038/nchem.1223
17. Schwartz SD, Schramm VL. Enzymatic transition states and dynamic motion in barrier crossing. Nat Chem Biol. 2009;5: 551–8. doi:10.1038/nchembio.202
18. Ma B, Nussinov R. Enzyme dynamics point to stepwise conformational selection in catalysis. Curr Opin Chem Biol. Elsevier Ltd; 2010;14: 652–9. doi:10.1016/j.cbpa.2010.08.012
19. Liu Y, Bahar I. Sequence evolution correlates with structural dynamics. Mol Biol Evol. 2012;29: 2253–63. doi:10.1093/molbev/mss097
20. Marsh J a, Teichmann S a. Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. Bioessays. 2014;36: 209–18. doi:10.1002/bies.201300134
21. Maguid S, Fernandez-Alberti S, Ferrelli L, Echave J. Exploring the common dynamics of homologous proteins. Application to the globin family. Biophys J. Elsevier; 2005;89: 3–13. doi:10.1529/biophysj.104.053041
22. Maguid S, Fernández-Alberti S, Parisi G, Echave J. Evolutionary conservation of protein backbone flexibility. J Mol Evol. 2006;63: 448–57. doi:10.1007/s00239-005-0209-x
23. Raimondi F, Orozco M, Fanelli F. Deciphering the deformation modes associated with function retention and specialization in members of the Ras superfamily. Structure. Elsevier Ltd; 2010;18: 402–14. doi:10.1016/j.str.2009.12.015
24. Marcos E, Crehuet R, Bahar I. On the conservation of the slow conformational dynamics within the amino acid kinase family: NAGK the paradigm. PLoS Comput Biol. 2010;6: e1000738. doi:10.1371/journal.pcbi.1000738
25. Luebbering EK, Mick J, Singh RK, Tanner JJ, Mehra-Chaudhary R, Beamer LJ. Conservation of functionally important global motions in an enzyme superfamily across varying quaternary structures. J Mol Biol. Elsevier Ltd; 2012;423: 831–46. doi:10.1016/j.jmb.2012.08.013
26. Maguid S, Fernandez-Alberti S, Echave J. Evolutionary conservation of protein vibrational dynamics.

Gene. 2008;422: 7–13. doi:10.1016/j.gene.2008.06.002
27. Zen A, Carnevale V, Lesk A, Micheletti C. Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families. Protein Sci. 2008;17: 918–929. doi:10.1110/ps.073390208
28. Pang A, Arinaminpathy Y, Sansom MSP, Biggin PC. Comparative molecular dynamics--similar folds and similar motions? Proteins. 2005;61: 809–22. doi:10.1002/prot.20672
29. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. An analysis of core deformations in protein superfamilies. Biophys J. 2005;88: 1291–9. doi:10.1529/biophysj.104.052449
30. Velázquez-Muriel J a, Rueda M, Cuesta I, Pascual-Montano A, Orozco M, Carazo J-M. Comparison of molecular dynamics and superfamily spaces of protein domain deformation. BMC Struct Biol. 2009;9: 6. doi:10.1186/1472-6807-9-6
31. Tobi D, Bahar I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc Natl Acad Sci U S A. 2005;102: 18908–18913.
32. Bahar I, Lezon TR, Yang L-W, Eyal E. Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys. 2010;39: 23–42. doi:10.1146/annurev.biophys.093008.131258
33. Hollup SM, Fuglebakk E, Taylor WR, Reuter N. Exploring the factors determining the dynamics of different protein folds. Protein Sci. 2011;20: 197–209. doi:10.1002/pro.558
34. Echave J. Why are the low-energy protein normal modes evolutionarily conserved? Pure Appl Chem. 2012;84: 1931–1937. doi:10.1351/PAC-CON-12-02-15
35. Liberles D a, Teichmann S a, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, et al. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 2012;21: 769–85. doi:10.1002/pro.2071
36. Echave J, Fernández FM. A perturbative view of protein structural variation. Proteins. 2010;78: 173–80. doi:10.1002/prot.22553
37. Lai J, Jin J, Kubelka J, Liberles D a. A phylogenetic analysis of normal modes evolution in enzymes and its relationship to enzyme function. J Mol Biol. Elsevier Ltd; 2012;422: 442–59. doi:10.1016/j.jmb.2012.05.028
38. Ramanathan A, Agarwal PK. Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. PLoS Biol. 2011;9: e1001193. doi:10.1371/journal.pbio.1001193
39. Keskin O, Jernigan R, Bahar I. Proteins with similar architecture exhibit similar large-scale dynamic behavior. Biophys J. 2000;78: 2093–2106.
40. Münz M, Lyngsø R, Hein J, Biggin PC. Dynamics based alignment of proteins: an alternative approach to quantify dynamic similarity. BMC Bioinformatics. 2010;11: 188. doi:10.1186/1471-2105-11-188
41. Bhabha G, Ekiert DC, Jennewein M, Zmasek CM, Tuttle LM, Kroon G, et al. Divergent evolution of protein conformational dynamics in dihydrofolate reductase. Nat Struct Mol Biol. Nature Publishing Group; 2013;20: 1243–9. doi:10.1038/nsmb.2676
42. Dellus-Gur E, Toth-Petroczy A, Elias M, Tawfik DS. What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. J Mol Biol. Elsevier Ltd; 2013;425: 2609–21. doi:10.1016/j.jmb.2013.03.033
43. Gatti-Lafranconi P, Hollfelder F. Flexibility and reactivity in promiscuous enzymes. Chembiochem. 2013;14: 285–92. doi:10.1002/cbic.201200628
44. Münz M, Hein J, Biggin PC. The role of flexibility and conformational selection in the binding promiscuity of PDZ domains. PLoS Comput Biol. 2012;8: e1002749. doi:10.1371/journal.pcbi.1002749
45. Tokuriki N, Tawfik D. Protein dynamism and evolvability. Science. 2009;324: 203–207. doi:10.1126/science.1169375
46. Pandini A, Mauri G, Bordogna A, Bonati L. Detecting similarities among distant homologous proteins by comparison of domain flexibilities. Protein Eng Des Sel. 2007;20: 285–99. doi:10.1093/protein/gzm021
47. Gherardini PF, Helmer-Citterich M. Structure-based function prediction: approaches and applications. Brief Funct Genomic Proteomic. 2008;7: 291–302. doi:10.1093/bfgp/eln030
48. Hensen U, Meyer T, Haas J, Rex R, Vriend G, Grubmüller H. Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. PLoS One. 2012;7: e33931. doi:10.1371/journal.pone.0033931
49. Tobi D. Dynamics alignment: comparison of protein dynamics in the SCOP database. Proteins. 2012;80: 1167–76. doi:10.1002/prot.24017
50. Micheletti C. Comparing proteins by their internal dynamics: exploring structure-function relationships beyond static structural alignments. Phys Life Rev. Elsevier B.V.; 2013;10: 1–26. doi:10.1016/j.plrev.2012.10.009
51. Potestio R, Aleksiev T, Pontiggia F, Cozzini S, Micheletti C. ALADYN: a web server for aligning

proteins by matching their large-scale motion. Nucleic Acids Res. 2010;38: W41–5. doi:10.1093/nar/gkq293
52. Tobi D. Normal Mode Dynamics Comparison of Proteins. 2014;40700: 1118–1125. doi:10.1002/ijch.201300142
53. Bakan A, Meireles LM, Bahar I. ProDy: protein dynamics inferred from theory and experiments. Bioinformatics. 2011;27: 1575–7. doi:10.1093/bioinformatics/btr168
54. Bakan A, Dutta A, Mao W, Liu Y, Chennubhotla C, Lezon TR, et al. Evol and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics. 2014;30: 2681–3. doi:10.1093/bioinformatics/btu336
55. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–42.
56. Hess B, Kutzner C, Van Der Spoel D, Lindahl E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput. 2008;4: 435–447. doi:10.1021/ct700301q
57. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: A message-passing parallel molecular dynamics implementation. Comput Phys Commun. 1995;91: 43–56. doi:10.1016/0010-4655(95)00042-E
58. Lindahl E, Hess B, Spoel D Van Der. GROMACS 3 . 0 : a package for molecular simulation and trajectory analysis. J Mol Model. 2001;43: 306–317. doi:10.1007/s008940100045
59. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins Struct Funct Bioinforma. 2010;78: 1950–1958. doi:10.1002/prot.22711
60. Berendsen HJC, Grigera JR, Straatsma TP. The Missing Term in Effective Pair Potentials. J Phys Chem. 1987;91: 6269–6271. doi:10.1021/j100308a038
61. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys. 2007;126: 014101. doi:10.1063/1.2408420
62. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys. 1984;81: 3684. doi:10.1063/1.448118
63. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. J Comput Chem. 1997;18: 1463–1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H
64. Xu D, Cui Q, Guo H. Quantum mechanical/molecular mechanical studies of zinc hydrolases. Int Rev Phys Chem. 2014;33: 1–41. doi:10.1080/0144235X.2014.889378
65.Holland DR, Hausrath AC, Juers D, Matthews BW. Structural analysis of zinc substitutions in the active site of thermolysin. Protein Sci. 1995;4: 1955–1965. doi:10.1002/pro.5560041001
66. Doruker P, Atilgan AR, Bahar I. Dynamics of proteins predicted by molecular simulations and analytical approaches: Application to alpha-amylase inhibitor. Proteins Struct Funct Genet. 2000;40: 512–524. doi:10.1002/1097-0134(20000815)40:3<512::AID-PROT180>3.0.CO;2-M
67. Atilgan a R, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J. 2001;80: 505–515. doi:10.1016/S0006-3495(01)76033-X
68. Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode calculations. Protein Eng. 2001;14: 1–6. doi:10.1093/protein/14.1.1
69. Brüschweiler R. Collective protein dynamics and nuclear spin relaxation. J Chem Phys. 1995;102: 3396. doi:10.1063/1.469213
70. Amadei A, Ceruso MA, Di Nola A. On the convergence of the conformational coordinates basis set obtained by the Essential Dynamics analysis of proteins' molecular dynamics simulations. Proteins Struct Funct Genet. 1999;36: 419–424. doi:10.1002/(SICI)1097-0134(19990901)36:4<419::AID-PROT5>3.0.CO;2-U
71. Hess B. Convergence of sampling in protein simulations. Phys Rev E - Stat Nonlinear, Soft Matter Phys. 2002;65: 1–10. doi:10.1103/PhysRevE.65.031910
72. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics. 2000;16: 566–567. doi:10.1093/bioinformatics/16.6.566
73. Micheletti C, Carloni P, Maritan A. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. Proteins. 2004;55: 635–45. doi:10.1002/prot.20049
74. Zen A, Carnevale V, Lesk AM, Micheletti C. Correspondences between low-energy modes in enzymes: dynamics-based alignment of enzymatic functional families. Protein Sci. 2008;17: 918–929. doi:10.1110/ps.073390208
75. Fuglebakk E, Reuter N, Hinsen K. Evaluation of protein elastic network models based on an analysis

of collective motions. J Chem Theory Comput. 2013;9: 5618–5628. doi:10.1021/ct400399x
76. Pelmenschikov V, Blomberg MR a, Siegbahn PEM. A theoretical study of the mechanism for peptide hydrolysis by thermolysin. J Biol Inorg Chem. 2002;7: 284–98. doi:10.1007/s007750100295
77. Blumberger J, Lamoureux G, Klein ML. Peptide Hydrolysis in Thermolysin: Ab Initio QM/MM Investigation of the Glu143-Assisted Water Addition Mechanism. J Chem Theory Comput. 2007;3: 1837–1850. doi:10.1021/ct7000792
78. Bakan A, Bahar I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. Proc Natl Acad Sci U S A. 2009;106: 14349–54. doi:10.1073/pnas.0904214106
79. Yang L, Song G, Carriquiry A, Jernigan RL. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. Structure. 2008;16: 321–30. doi:10.1016/j.str.2007.12.011
80. Meireles L, Gur M, Bakan A, Bahar I. Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins. Protein Sci. 2011;20: 1645–58. doi:10.1002/pro.711
81. Ichiye T, Karplus M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. Proteins. 1991;11: 205–217. doi:10.1002/prot.340110305
82. Kester WR, Matthews BW. Crystallographic study of the binding of dipeptide inhibitors to thermolysin: implications for the mechanism of catalysis. Biochemistry. 1977;16: 2506–2516.
83. Hausrath AC, Matthews BW. Thermolysin in the absence of substrate has an open conformation. Acta Crystallogr Sect D Biol Crystallogr. International Union of Crystallography; 2002;58: 1002–1007. doi:10.1107/S090744490200584X
84. Holland DR, Tronrud DE, Pley HW, Flaherty KM, Stark W, Jansonius JN, et al. Structural comparison suggests that thermolysin and related neutral proteases undergo hinge-bending motion during catalysis. Biochemistry. 1992;31: 11310–6.
85. Aalten D Van, Amadei A, Linssen a B, Eijsink V, Vriend G, Berendsen HJC. The essential dynamics of thermolysin: Confirmation of the hinge-bending motion and comparison of simulations in vacuum and water. Proteins Struct Funct Genet. 1995;22: 45–54. doi:10.1002/prot.340220107
86. Amadei a, Linssen a B, Berendsen HJ. Essential dynamics of proteins. Proteins. 1993;17: 412–25. doi:10.1002/prot.340170408
87. Daidone I, Amadei A. Essential dynamics: foundation and applications. Wiley Interdiscip Rev Comput Mol Sci. 2012;2: 762–770. doi:10.1002/wcms.1099
88. Ermakova E, Kurbanov R. Effect of ligand binding on the dynamics of trypsin. Comparison of different approaches. J Mol Graph Model. Elsevier Inc.; 2014;49: 99–109. doi:10.1016/j.jmgm.2014.02.001
89. Skjaerven L, Martinez A, Reuter N. Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. Proteins Struct Funct Bioinforma. 2011;79: 232–243. doi:10.1002/prot.22875.
90. Bakan A, Bahar I. Computational Generation inhibitor-Bound Conformers of P38 Map Kinase and Comparison with Experiments. Pacific Symp Biocomput. 2011; 181–192.
91. Fuglebakk E, Echave J, Reuter N. Measuring and comparing structural fluctuation patterns in large protein datasets. Bioinformatics. 2012;28: 2431–40. doi:10.1093/bioinformatics/bts445.
92. Bahar I, Chennubhotla C, Tobi D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. Curr Opin Struct Biol. 2007;17: 633–640. doi:10.1016/j.sbi.2007.09.011
93. Eyal E, Yang L-W, Bahar I. Anisotropic network model: systematic evaluation and a new web interface. Bioinformatics. 2006;22: 2619–27. doi:10.1093/bioinformatics/btl448
94. Fuglebakk E, Reuter N, Hinsen K. Evaluation of Protein Elastic Network Models Based on an Analysis of Collective Motions. J Chem Theory Comput. 2013;9: 5618–5628. doi:10.1021/ct400399x
95. Rueda M, Chacón P, Orozco M. Thorough Validation of Protein Normal Mode Analysis: A Comparative Study with Essential Dynamics. Structure. 2007;15: 565–575. doi:10.1016/j.str.2007.03.013
96. Riccardi D, Cui Q, Phillips GN. Evaluating elastic network models of crystalline biological molecules with temperature factors, correlated motions, and diffuse X-ray scattering. Biophys J. Biophysical Society; 2010;99: 2616–2625. doi:10.1016/j.bpj.2010.08.013
97. Romo TD, Grossfield A. Validating and improving elastic network models with molecular dynamics simulations. Proteins Struct Funct Bioinforma. 2011;79: 23–34. doi:10.1002/prot.22855
98. Leioatts N, Romo TD, Grossfield A. Elastic network models are robust to variations in formalism. J Chem Theory Comput. 2012;8: 2424–2434. doi:10.1021/ct3000316
99. Kundu S, Melton JS, Sorensen DC, Phillips GN. Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys J. Elsevier; 2002;83: 723–732. doi:10.1016/S0006-

3495(02)75203-X
100. Yang LW, Eyal E, Chennubhotla C, Jee J, Gronenborn AM, Bahar I. Insights into Equilibrium Dynamics of Proteins from Comparison of NMR and X-Ray Data with Computational Predictions. Structure. 2007;15: 741–749. doi:10.1016/j.str.2007.04.014
101. Gur M, Zomot E, Bahar I. Global motions exhibited by proteins in micro- to milliseconds simulations concur with anisotropic network model predictions. J Chem Phys. 2013;139: 121912. doi:10.1063/1.4816375
102. Yang L-W, Bahar I. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure. 2005;13: 893–904. doi:10.1016/j.str.2005.03.015
103. Dutta A, Bahar I. Metal-binding sites are designed to achieve optimal mechanical and signaling properties. Structure. Elsevier Ltd; 2010;18: 1140–8. doi:10.1016/j.str.2010.06.013
104. Bakan A, Meireles LM, Bahar I. ProDy: protein dynamics inferred from theory and experiments. Bioinformatics. 2011;27: 1575–7. doi:10.1093/bioinformatics/btr168
105. Bahar I, Lezon TR, Bakan A, Shrivastava IH. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. Chem Rev. 2010;110: 1463–97. doi:10.1021/cr900095e
106. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. 2001;80.
107. Brown CK, Madauss K, Lian W, Beck MR, Tolbert WD, Rodgers DW. Structure of neurolysin reveals a deep channel that limits substrate access. Proc Natl Acad Sci U S A. 2001;98: 3127–32. doi:10.1073/pnas.051633198
108. Arndt JW, Hao B, Ramakrishnan V, Cheng T, Chan SI, Chan MK. Crystal structure of a novel carboxypeptidase from the hyperthermophilic archaeon Pyrococcus furiosus. Structure. 2002;10: 215–24.
109. Natesh R, Schwager S, Sturrock E, Acharya K. Crystal structure of the human angiotensin-converting enzyme–lisinopril complex. Nature. 2003;421: 551–554.
110. Lee MM, Isaza CE, White JD, Chen RP-Y, Liang GF-C, He HT-F, et al. Insight into the substrate length restriction of M32 carboxypeptidases: characterization of two distinct subfamilies. Proteins. 2009;77: 647–57. doi:10.1002/prot.22478
111. Watermeyer JM, Sewell BT, Schwager SL, Natesh R, Corradi HR, Acharya KR, et al. Structure of testis ACE glycosylation mutants and evidence for conserved domain movement. Biochemistry. 2006;45: 12654–63. doi:10.1021/bi061146z
112. Comellas-Bigler M, Lang R, Bode W, Maskos K. Crystal structure of the E. coli dipeptidyl carboxypeptidase Dcp: further indication of a ligand-dependent hinge movement mechanism. J Mol Biol. 2005;349: 99–112. doi:10.1016/j.jmb.2005.03.016
113. Etges R, Bouvier J, Bordier C. The major surface protein of Leishmania promastigotes is a protease. J Biol Chem. 1986;261: 9098–101.
114. Schlagenhauf E, Etges R, Metcalf P. The crystal structure of the Leishmania major surface proteinase leishmanolysin (gp63). Structure. 1998;6: 1035–46.
115. Bianchini G, Bocedi A, Ascenzi P, Gavuzzo E, Mazza F, Aschi M. Molecular dynamics simulation of Leishmania major surface metalloprotease GP63 (leishmanolysin). Proteins Struct Funct Bioinforma. 2006;64: 385–390. doi:10.1002/prot.21009

# Supporting Information

## S1 File

**SI Table 1.** Thermolysin structures (Uniprot ID: P00800) retrieved from the PDB.

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1FJ3 | 1FJO | 1FJQ | 1FJT | 1FJU | 1FJV | 1FJW | 1GXW | 1HYT | 1KEI | 1KJO | 1KJP | 1KKK | 1KL6 | 1KR6 | 1KRO | 1KS7 | 1KTO | **1L3F** |
| 1LNA | 1LNB | 1LNC | 1LND | 1LNE | 1LNF | 1OS0 | 1PE5 | 1PE7 | 1PE8 | 1QF0 | 1QF1 | 1QF2 | 1THL | 1TLI | 1TLP | 1TLX | 1TMN | 1Y3G |
| 1Z9G | 1ZDP | 2A7G | 2G4Z | 2TLI | 2TLX | 2TMN | 2WHZ | 2WI0 | 3DNZ | 3DO0 | 3DO1 | 3DO2 | 3EIM | 3F28 | 3F2P | 3FB0 | 3FBO | 3FCQ |
| 3FGD | 3FLF | 3FOR | 3FV4 | 3FVP | 3FXP | 3FXS | 3LS7 | 3MS3 | 3MSA | 3MSF | 3MSN | 3N21 | 3NN7 | 3P7P | 3P7Q | 3P7R | 3P7S | 3P7T |
| 3P7U | 3P7V | 3P7W | 3QGO | 3QH1 | 3QH5 | 3SSB | **3T2H** | **3T2I** | **3T2J** | 3T73 | 3T74 | 3T87 | 3T8C | 3T8D | 3T8F | 3T8G | 3T8H | 3TLI |
| 3TMN | 3ZI6 | 4D91 | 4D9W | 4H57 | 4TLI | 4TLN | 4TMN | 5TLI | 5TLN | 5TMN | 6TLI | 6TMN | 7TLI | 7TLN | 8TLI | 8TLN | | |

Structures were obtained using the Prody software (as of 09/2013). Bold: unbound crystal structure used for ANM generation. Underlined: unbound crystal structures obtained in the presence of cryoprotectors.

**SI Table 2.** List of Z-scores and P-values obtained for the alignments of MP representative structures.

| Pair | Z-Score | P-value | Gluzincin | Mixed | Metzincin |
|---|---|---|---|---|---|
| M3-M27 | 0.9 | 6.18E-02 | x | | |
| M32-M27 | 1.2 | 9.37E-04 | x | | |
| M32-M10(B) | 1.3 | 1.29E-02 | | x | |
| M32-M8 | 1.3 | 2.55E-04 | | x | |
| M4-M10(B) | 1.3 | 3.35E-02 | | x | |
| M27-M10(A) | 1.5 | 4.55E-01 | | x | |
| M32-M12(B) | 1.9 | 2.12E-01 | | x | |
| M3-M35 | 1.9 | 4.87E-01 | | x | |
| M27-M10(B) | 1.9 | 5.43E-02 | | x | |
| M27-M8 | 1.9 | 1.47E-03 | | x | |
| M4-M27 | 2 | 2.32E-02 | x | | |
| M32-M10(A) | 2.1 | 4.54E-01 | | x | |
| M27-M7 | 2.1 | 4.46E-01 | | x | |
| M4-M12(B) | 2.1 | 8.04E-02 | | x | |
| M27-M12(B) | 2.2 | 1.89E-01 | | x | |
| M32-M35 | 2.2 | 3.39E-01 | | x | |
| M32-M12(A) | 2.3 | 1.27E-01 | | x | |
| M27-M35 | 2.3 | 1.91E-01 | | x | |
| M2-M12(B) | 2.3 | 4.14E-01 | | x | |
| M27-M12(A) | 2.3 | 1.81E-01 | | x | |
| M32-M7 | 2.4 | 3.90E-01 | | x | |
| M4-M8 | 2.6 | 9.08E-03 | | x | |
| M2-M27 | 2.7 | 4.98E-02 | x | | |
| M4-M10(A) | 2.7 | 1.48E-01 | | x | |
| M4-M12(A) | 2.7 | 4.06E-02 | | x | |
| M1-M27 | 2.8 | 5.12E-02 | x | | |
| M2-M8 | 2.8 | 1.22E-02 | | x | |
| M3-M12(A) | 2.8 | 4.03E-01 | | x | |
| M2-M10(B) | 2.9 | 8.12E-02 | | x | |
| M4-M35 | 2.9 | 4.48E-02 | | x | |
| M3-M12(B) | 3 | 4.46E-01 | | x | |
| M1-M12(A) | 3.1 | 3.79E-01 | | x | |
| M35-M8 | 3.1 | 2.97E-01 | | | x |
| M35-M12(A) | 3.2 | 1.48E-02 | | | x |
| M3-M10(B) | 3.2 | 9.09E-02 | | x | |
| M2-M35 | 3.2 | 4.57E-01 | | x | |
| M2-M12(A) | 3.2 | 3.77E-01 | | x | |
| M4-M7 | 3.3 | 3.24E-01 | | x | |
| M3-M8 | 3.4 | 2.12E-02 | | x | |
| M2-M10(A) | 3.5 | 5.27E-01 | | x | |

| | | | | | |
|---|---|---|---|---|---|
| M35-M10(B) | 3.7 | 2.80E-01 | | | x |
| M1-M12(B) | 3.7 | 4.11E-01 | | x | |
| M3-M7 | 3.7 | 5.29E-01 | | x | |
| M35-M10(A) | 3.9 | 3.62E-02 | | | x |
| M3-M10(A) | 3.9 | 5.32E-01 | | x | |
| M1-M10(B) | 4 | 1.52E-02 | | x | |
| M35-M12(B) | 4 | 1.00E-02 | | | x |
| M2-M7 | 4.1 | 5.29E-01 | | x | |
| M1-M35 | 4.2 | 3.91E-01 | | x | |
| M7-M35 | 4.2 | 3.06E-02 | | | x |
| M1-M10(A) | 4.2 | 4.98E-01 | | x | |
| M1-M7 | 4.3 | 5.17E-01 | | x | |
| M1-M8 | 4.4 | 5.52E-03 | | x | |
| M1-M2 | 4.8 | 1.27E-02 | x | | |
| M1-M3 | 5.4 | 7.00E-03 | x | | |
| M4-M32 | 5.9 | 1.77E-02 | x | | |
| M8-M12(A) | 6.1 | 1.68E-01 | | | x |
| M4-M3 | 6.5 | 9.02E-02 | x | | |
| M1-M32 | 6.6 | 9.06E-03 | x | | |
| M12(A)-M12(B) | 6.9 | 2.12E-02 | | | x |
| M4-M2 | 7.2 | 1.33E-01 | x | | |
| M8-M12(B) | 7.3 | 2.01E-01 | | | x |
| M8-M10(B) | 7.5 | 3.30E-03 | | | x |
| M10(B)-M12(B) | 7.7 | 3.07E-01 | | | x |
| M7-M8 | 7.8 | 4.67E-01 | | | x |
| M8-M10(A) | 8.6 | 3.84E-01 | | | x |
| M10(B)-M12(A) | 8.7 | 1.61E-01 | | | x |
| M7-M12(B) | 9.8 | 2.34E-02 | | | x |
| M12(A)-M10(A) | 10.8 | 1.65E-01 | | | x |
| M12(B)-M10(A) | 11.1 | 2.50E-02 | | | x |
| M7-M12(A) | 11.3 | 8.75E-02 | | | x |
| M7-M10(B) | 11.3 | 4.70E-01 | | | x |
| M4-M1 | 11.4 | 8.13E-02 | x | | |
| M7-M10(A) | 12.7 | 1.20E-01 | | | x |
| M10(B)-M10(A) | 13.8 | 3.18E-01 | | | x |
| M3-M2 | 19.2 | 4.78E-04 | x | | |
| M3-M32 | 19.7 | 6.94E-06 | x | | |
| M32-M2 | 21.4 | 2.97E-05 | x | | |

## S2 File

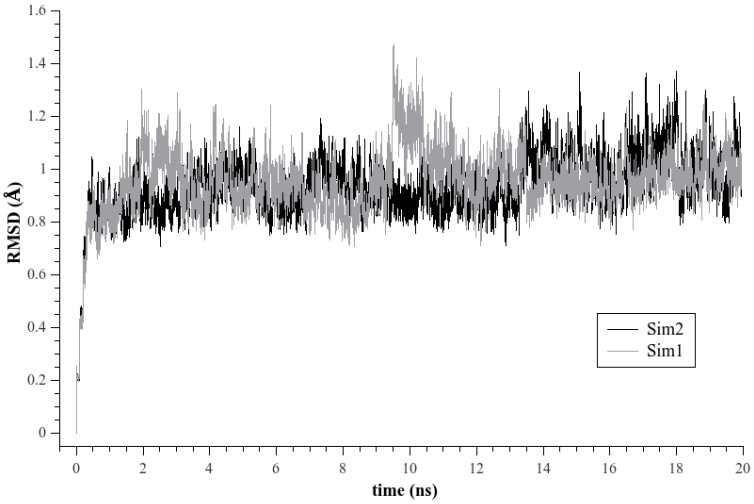



**Figure A.** Root Mean Square Deviations of residue $C^{\alpha}$ obtained for Sim1 and Sim2.

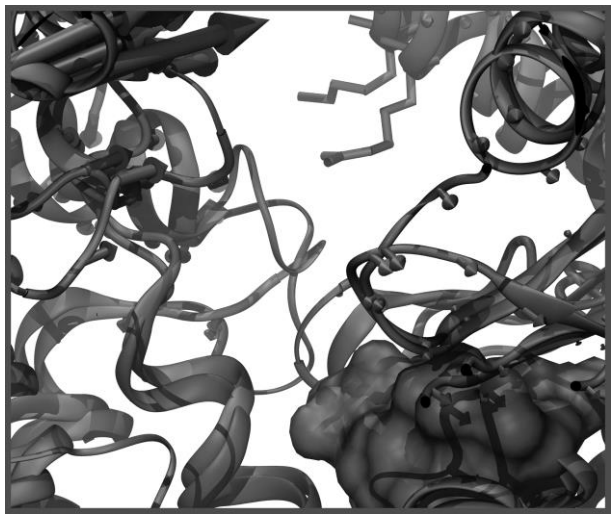


**Figure B: Dynamics-based alignment of Neurolysin (Blue, PDB ID: 1I1I) and Carboxypeptidase Pfu (red, PDB ID: 1KA4).** Neurolysin K148 and Carboxypeptidase Pfu R92 in bond representation. Active site residues in colored surface representations (Neurolysin: H474, E 475, H478 and E503; Carboxypeptidase Pfu H269, E270, H273 and E299).

# Annex 2

**Scheme A1.** Ligand .params files.

*1QJI_diala_conf.params*

```
NAME dA1
IO_STRING dA1 Z
TYPE LIGAND
AA UNK
ATOM ZN1  Zn2p  X   1.96
ATOM  O2  OH    X   -0.70
ATOM  C1  CH1   X   -0.13
ATOM  O3  OOC   X   -0.80
ATOM  N3  Ntrp  X   -0.65
ATOM  C8  CH1   X   -0.13
ATOM  C2  CH3   X   -0.31
ATOM  H12 Hapo  X   0.06
ATOM  H13 Hapo  X   0.06
ATOM  H14 Hapo  X   0.06
ATOM  C5  COO   X   0.58
ATOM  N2  NH2O  X   -0.51
ATOM  H2  Hpol  X   0.39
ATOM  H3  Hpol  X   0.39
ATOM  O4  ONH2  X   -0.59
ATOM  H1  Hapo  X   0.06
ATOM  H7  Hpol  X   0.39
ATOM  C6  CH1   X   -0.13
ATOM  N1  Ntrp  X   -0.65
ATOM  C4  COO   X   0.58
ATOM  O1  ONH2  X   -0.59
ATOM  C3  CH3   X   -0.31
ATOM  H8  Hapo  X   0.06
ATOM  H10 Hapo  X   0.06
ATOM  H11 Hapo  X   0.06
ATOM  H16 Hpol  X   0.39
ATOM  C7  CH3   X   -0.31
ATOM  H4  Hapo  X   0.06
ATOM  H5  Hapo  X   0.06
ATOM  H6  Hapo  X   0.06
ATOM  H9  Hapo  X   0.06
ATOM  H15 Hpol  X   0.39
BOND  C1   O2
BOND  C1   O3
BOND  C1   N3
BOND  C1   C6
BOND  C2   C8
BOND  C2   H12
BOND  C2   H13
BOND  C2   H14
BOND  N1   C4
BOND  N1   C6
BOND  N1   H16
BOND  O1   C4
BOND  N2   C5
BOND  N2   H2
BOND  N2   H3
BOND  O2   H15
BOND  O2  ZN1
BOND  O4   C5
BOND  C3   C4
BOND  C3   H8
BOND  C3   H10
```

```
BOND  C3   H11
BOND  N3   C8
BOND  N3   H7
BOND  C5   C8
BOND  C6   C7
BOND  C6   H9
BOND  C7   H4
BOND  C7   H5
BOND  C7   H6
BOND  C8   H1
CHI 1 ZN1   O2   C1   O3
CHI 2  O2   C1   N3   C8
CHI 3  O2   C1   C6   N1
CHI 4  C6   N1   C4   O1
CHI 5  C1   C6   N1   C4
CHI 6  C1   N3   C8   C2
CHI 7  N3   C8   C5   N2
NBR_ATOM  O2
NBR_RADIUS 8.363764
ICOOR_INTERNAL   ZN1     0.000000    0.000000    0.000000  ZN1    O2    C1
ICOOR_INTERNAL    O2     0.000000  180.000000    1.849189  ZN1    O2    C1
ICOOR_INTERNAL    C1     0.000001   56.729696    1.595460   O2   ZN1    C1
ICOOR_INTERNAL    O3     5.414501   78.336182    1.568874   C1    O2   ZN1
ICOOR_INTERNAL    N3   119.370020   65.773055    1.655588   C1    O2    O3
ICOOR_INTERNAL    C8   -76.489396   61.101004    1.535813   N3    C1    O2
ICOOR_INTERNAL    C2    77.096851   69.172846    1.539788   C8    N3    C1
ICOOR_INTERNAL    H12 -116.166899   70.478464    1.089474   C2    C8    N3
ICOOR_INTERNAL    H13 -120.051541   70.526968    1.090604   C2    C8    H12
ICOOR_INTERNAL    H14 -119.923854   70.546616    1.090087   C2    C8    H13
ICOOR_INTERNAL    C5   121.054923   70.835105    1.524968   C8    N3    C2
ICOOR_INTERNAL    N2   132.297922   63.393354    1.341865   C5    C8    N3
ICOOR_INTERNAL    H2    -3.246935   59.118818    1.009561   N2    C5    C8
ICOOR_INTERNAL    H3  -179.985254   60.445997    1.009692   N2    C5    H2
ICOOR_INTERNAL    O4   179.726837   58.151996    1.223927   C5    C8    N2
ICOOR_INTERNAL    H1   119.995376   71.118126    1.089578   C8    N3    C5
ICOOR_INTERNAL    H7  -121.397122   72.908834    1.009422   N3    C1    C8
ICOOR_INTERNAL    C6   126.918397   69.610456    1.707063   C1    O2    N3
ICOOR_INTERNAL    N1    53.919212   64.020083    1.446771   C6    C1    O2
ICOOR_INTERNAL    C4  -101.124052   59.540397    1.324118   N1    C6    C1
ICOOR_INTERNAL    O1    -0.031823   56.993050    1.232007   C4    N1    C6
ICOOR_INTERNAL    C3  -178.207807   63.265011    1.513557   C4    N1    O1
ICOOR_INTERNAL    H8   -97.218306   69.813070    1.090471   C3    C4    N1
ICOOR_INTERNAL    H10  119.199821   68.909778    1.089870   C3    C4    H8
ICOOR_INTERNAL    H11  120.098134   72.723062    1.089461   C3    C4    H10
ICOOR_INTERNAL    H16 -179.110947   63.140534    0.980142   N1    C6    C4
ICOOR_INTERNAL    C7   127.125575   66.295147    1.525871   C6    C1    N1
ICOOR_INTERNAL    H4  -108.636956   70.497439    1.090028   C7    C6    C1
ICOOR_INTERNAL    H5  -120.078698   70.505253    1.090535   C7    C6    H4
ICOOR_INTERNAL    H6  -119.955610   70.572417    1.090338   C7    C6    H5
ICOOR_INTERNAL    H9   117.493922   78.793685    1.089756   C6    C1    C7
ICOOR_INTERNAL    H15  125.262060   73.810349    0.956392   O2   ZN1    C1
PDB_ROTAMERS 1QJI_diAla_conf_Zn_con.pdb
```

*4AIG_diala_min.params*

```
NAME min
IO_STRING min Z
TYPE LIGAND
AA UNK
ATOM  C1  COO   X   0.55
ATOM  N1  Nhis  X   -0.60
ATOM  C3  CH3   X   -0.34
ATOM  H2  Hapo  X   0.03
ATOM  H4  Hapo  X   0.03
ATOM  H6  Hapo  X   0.03
ATOM  O1  ONH2  X   -0.62
```

```
ATOM  O2  OH    X   -0.73
ATOM ZN1  Zn2p  X   1.93
ATOM  C2  CH3   X   -0.34
ATOM  H1  Hapo  X   0.03
ATOM  H3  Hapo  X   0.03
ATOM  H5  Hapo  X   0.03
BOND  C1   N1
BOND  C1   O1
BOND  C1   O2
BOND  C1   C2
BOND  N1   C3
BOND  O2  ZN1
BOND  C2   H1
BOND  C2   H3
BOND  C2   H5
BOND  C3   H2
BOND  C3   H4
BOND  C3   H6
CHI 1  O1   C1   N1   C3
CHI 2  N1   C1   O2  ZN1
NBR_ATOM  C1
NBR_RADIUS 4.030455
ICOOR_INTERNAL    C1     0.000000    0.000000    0.000000   C1    N1    C3
ICOOR_INTERNAL    N1     0.000000  180.000000    1.585183   C1    N1    C3
ICOOR_INTERNAL    C3    -0.000000   26.151042    1.644308   N1    C1    C3
ICOOR_INTERNAL    H2    75.318211   72.925954    1.089850   C3    N1    C1
ICOOR_INTERNAL    H4   108.657041   54.402937    1.090239   C3    N1    H2
ICOOR_INTERNAL    H6   111.273643   76.386196    1.049497   C3    N1    H4
ICOOR_INTERNAL    O1  -107.053303   68.766849    1.522513   C1    N1    C3
ICOOR_INTERNAL    O2   116.731225   70.741775    1.513449   C1    N1    O1
ICOOR_INTERNAL   ZN1  -108.750832   65.672835    2.002586   O2    C1    N1
ICOOR_INTERNAL    C2   121.612944   64.720447    1.469585   C1    N1    O2
ICOOR_INTERNAL    H1  -174.733641   65.184448    1.090371   C2    C1    N1
ICOOR_INTERNAL    H3  -110.290703   57.783360    1.089917   C2    C1    H1
ICOOR_INTERNAL    H5  -122.351165   87.064207    1.089580   C2    C1    H3
```

**Scheme A2**. AS .cst files
*MAM_diAla.cst*

```
# NOTES
# MAM subclan
# O2 corresponds to Ow atom. C1 to the tetrahedral carbon bound to Ow and Op.
# diAla has 3-letter code dA1.
# 6th column of distanceAB value set to 0 for non-bonded interaction, set to 1
for pseudocovalent interaction.
# When secondary algorithm is used, angle_A angle_B torsion_A torsion_AB tor-
sion_B commented out.

#Glu_cat - catalytic interaction

CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name: ZN1 O2 C1
  TEMPLATE::   ATOM_MAP: 1 residue3: dA1

  TEMPLATE::   ATOM_MAP: 2 atom_type: OOC ,    #either OE1 or OE2
  TEMPLATE::   ATOM_MAP: 2 residue1:  E

  CONSTRAINT:: distanceAB:   5.0  0.3  100  0    1  #2
  CONSTRAINT::    angle_A:  35.1  9.0   30  360  1  #2
  CONSTRAINT::    angle_B:  91.8  4.8   30  360  1  #1
  CONSTRAINT::  torsion_A: 181.4 18.4   30  360  1  #4
  CONSTRAINT:: torsion_AB:  75.4  9.1   30  360  1  #2
  CONSTRAINT::  torsion_B: 145.1  4.7   30  360  1  #1
CST::END
```

```
# His3 - Zn(II) coordination

CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name: ZN1 O2 C1
  TEMPLATE::   ATOM_MAP: 1 residue3: dA1

  TEMPLATE::   ATOM_MAP: 2 atom_type: Ntrp ,  #either ND1 or NE2
  TEMPLATE::   ATOM_MAP: 2 residue1:  H

  CONSTRAINT:: distanceAB:   2.1  0.2  100  1     1  #2
  CONSTRAINT::    angle_A: 144.0  7.8   30  360   1  #2
  CONSTRAINT::    angle_B: 124.3  4.5   30  360   1  #1
  CONSTRAINT::  torsion_A:  30.4 15.5   30  360   1  #3
  CONSTRAINT:: torsion_AB:  16.9 12.0   30  11.25 0  #1
  CONSTRAINT::  torsion_B: 166.7  5.6   30  360   1  #2

#    ALGORITHM_INFO:: match        ;not commented out when secondary algorithm
is used
#      SECONDARY_MATCH: DOWNSTREAM ;not commented out when secondary algorithm
is used
#    ALGORITHM_INFO::END           ;not commented out when secondary algorithm
is used

CST::END



# His1 - Zn(II) coordination

CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN1 O2 C1
  TEMPLATE::   ATOM_MAP: 1 residue3: dA1

  TEMPLATE::   ATOM_MAP: 2 atom_type: Ntrp,   #either ND1 or NE2
  TEMPLATE::   ATOM_MAP: 2 residue1:  H

  CONSTRAINT:: distanceAB:   2.1  0.1 100  1     1  #1
  CONSTRAINT::    angle_A:  96.8  7.1  30  360   1  #2
  CONSTRAINT::    angle_B: 128.4  4.1  30  360   1  #1
  CONSTRAINT::  torsion_A: 250.8 12.6  30  360   1  #3
  CONSTRAINT:: torsion_AB: 158.2 12.9  30  11.25 0  #1
  CONSTRAINT::  torsion_B: 167.1  9.2  30  360   1  #2

#    ALGORITHM_INFO:: match        ;not commented out when secondary algorithm
is used
#      SECONDARY_MATCH: DOWNSTREAM ;not commented out when secondary algorithm
is used
#    ALGORITHM_INFO::END           ;not commented out when secondary algorithm
is used

CST::END

# His2 - Zn(II) coordination

CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN1 O2 C1
  TEMPLATE::   ATOM_MAP: 1 residue3: dA1

  TEMPLATE::   ATOM_MAP: 2 atom_type: Ntrp  ,  #either ND1 or NE2
  TEMPLATE::   ATOM_MAP: 2 residue1:  H

  CONSTRAINT:: distanceAB:   2.1  0.1 100  1     1  #1
  CONSTRAINT::    angle_A:  94.2  8.1  30  360   1  #2
  CONSTRAINT::    angle_B: 123.8  4.5  30  360   1  #1
```

```
  CONSTRAINT::  torsion_A: 146.0 12.4  30  360   1  #3
  CONSTRAINT:: torsion_AB: 198.2  4.9  30  11.25 0  #1
  CONSTRAINT::  torsion_B: 190.4  4.8  30  360   1  #1

#    ALGORITHM_INFO:: match         ;not commented out when secondary algo-
rithm is used
#      SECONDARY_MATCH: DOWNSTREAM  ;not commented out when secondary algo-
rithm is used
#    ALGORITHM_INFO::END            ;not commented out when secondary algo-
rithm is used

CST::END
```

*4AIG_diala_min.cst*

```
# NOTE

#Glu_cat-Ow

CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name: ZN1 O2 C1
  TEMPLATE::   ATOM_MAP: 1 residue3: dA1

  TEMPLATE::   ATOM_MAP: 2 atom_type: OOC ,
  TEMPLATE::   ATOM_MAP: 2 residue1:  E

  CONSTRAINT:: distanceAB:   5.2  0.26   100  0    2
  CONSTRAINT::    angle_A:   31.5  8.97    0  360  2
  CONSTRAINT::    angle_B:   94.00 4.84    0  360  2
  CONSTRAINT::  torsion_A:  186.00 18.38   0  360  3
  CONSTRAINT:: torsion_AB:  82.00  9.07    0  360  2
  CONSTRAINT::  torsion_B:  148.00 4.69    0  360  2
CST::END

# His 3-ZN

CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name: ZN1 O2 C1
  TEMPLATE::   ATOM_MAP: 1 residue3: dA1

  TEMPLATE::   ATOM_MAP: 2 atom_type: Ntrp ,
  TEMPLATE::   ATOM_MAP: 2 residue1:  H

  CONSTRAINT:: distanceAB:  2.30   0.16  100   1     1
  CONSTRAINT::    angle_A:  144.7  7.80    0   360   2
  CONSTRAINT::    angle_B:  131.4  4.55    0   360   2
  CONSTRAINT::  torsion_A:  37.7   15.54   0   360   3
  CONSTRAINT:: torsion_AB:  4.30   12.01   0   11.25 0
  CONSTRAINT::  torsion_B:  171.3  5.57    0   360   2

#    ALGORITHM_INFO:: match
#      SECONDARY_MATCH: DOWNSTREAM
#    ALGORITHM_INFO::END

CST::END



# His 1-ZN

CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN1 O2 C1
  TEMPLATE::   ATOM_MAP: 1 residue3: dA1

  TEMPLATE::   ATOM_MAP: 2 atom_type: Ntrp,
```

```
  TEMPLATE::   ATOM_MAP: 2 residue1:  H

  CONSTRAINT:: distanceAB:  2.10   0.11  100  1     1
  CONSTRAINT::    angle_A:  96.00  7.12   0   360   2
  CONSTRAINT::    angle_B:  128.80 4.05   0   360   2
  CONSTRAINT::  torsion_A:  246.20 12.60  0   360   3
  CONSTRAINT:: torsion_AB:  167.30 12.92  0   11.25 0
  CONSTRAINT::  torsion_B:  161.90 9.22   0   360   2

#    ALGORITHM_INFO:: match
#      SECONDARY_MATCH: DOWNSTREAM
#    ALGORITHM_INFO::END

CST::END

#His 2-ZN

CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN1 O2 C1
  TEMPLATE::   ATOM_MAP: 1 residue3: dA1

  TEMPLATE::   ATOM_MAP: 2 atom_type: Ntrp  ,
  TEMPLATE::   ATOM_MAP: 2 residue1:  H

  CONSTRAINT:: distanceAB:  2.20   0.11  100    1     1
  CONSTRAINT::    angle_A:  83.90  8.05    0    360   2
  CONSTRAINT::    angle_B:  121    4.49    0    360   2
  CONSTRAINT::  torsion_A:  143.50 12.44   0    360   3
  CONSTRAINT:: torsion_AB:  201.90 4.94    0    11.25 0
  CONSTRAINT::  torsion_B:  193.00 4.78    0    360   1

#    ALGORITHM_INFO:: match
#      SECONDARY_MATCH: DOWNSTREAM
#    ALGORITHM_INFO::END

CST::END
```

**Scheme A3.** *options_matcher.flags*

```
-packing
 -ex1:level 3 #Chi'1 sampling level
 -ex2:level 3 #Chi'2 sampling level
 -ex2aro
 -ex1aro
 -use_input_sc
 -linmem_ig 10
-match
 -bump_tolerance 0.2
 -filter_colliding_upstream_residues
 -filter_upstream_downstream_collisions
 -output_format CloudPDB
 -enumerate_ligand_rotamers      #usefull for ligand with rotamers
 -only_enumerate_non_match_redundant_ligand_rotamers  #to be used with enumer-
ate_ligand_rotamers
-dun10 #Usage of Dunbrack 2010 library of rotamers
-mute protocols.idealize
```

**Scheme A4.** *options_enzdes.flags*

```
-dun10
-packing
 -use_input_sc
 -ex1:level 6 #Chi'1 sampling level
 -ex2:level 6 #Chi'2 sampling level
 -ex1aro:level 6
```

```
 -ex2aro:level 6
 -soft_rep_design
 -linmem_ig 10
-enzdes
# -cst_dock True ;Not used since Zn(II) belongs to substrate
 -parser_read_cloud_pdb true
 -cst_opt
 -chi_min
 -bb_min
 -cst_design
 -cst_min
 -design_min_cycles 4
 -detect_design_interface
 -cut1 6.0 #design of residues within 6A of any ligand heavy atom
 -cut2 8.0 #design of residues within 8A of any ligand heavy atom and Cbeta
atom closer to ligand than Ca
 -cut3 10.0 #repack of residues within 6A of any ligand heavy atom
 -cut4 12.0 #repack of residues within 8A of any ligand heavy atom and Cbeta
atom closer to ligand than Ca
# -compare_native ;not used since gave unrealistic REU values for NMR poses
 -final_repack_without_ligand true #SAME AS IN no_unconstrained_repack
-score
   -weights enzdes.wts
-fix_catalytic_aa
-nstruct 1 #for design or -nstruct 10 for full sequence design
```

**Table A1.** Structures of MP-TSA complexes used in the analysis of MA(E) subclan AS.

| | | | |
|---|---|---|---|
| MA(E) | M4 Thermolysin | Thermolysin | 1KJO, 1KL6, 3FGD, 3FV4, 3FVP, 3QHI, 3T73, 3T74, 3T87, 3T8C, 3T8D, 3T8F, 3T8G, 3T8H, 3D9W, 4H57, ~~3FLF, 1PE5, 1HYT, 1PE8, 3FOR, 3F2P, 3FCQ, 3FXP, 6TMN, 1PE7, 5TMN, 4TMN, 2TMN, 1TMN, 1THL, 1KRO, 3F28, 3SSB, 1KR6, 1KS7, 1KTO, 1OS0, 1Y3G~~ |
| | | Elastase | 1U4G, 3DBK |
| | M13 Neprelysin | Neutral Endopeptidase | 1DMT |
| | | ECE-1 | ~~3DWB~~ |
| | | Neprelysin | 1R1H, 2YB9, ~~2QPJ~~ |
| | M2 ACE | Angiotensin Converting Enzyme | 1O86, 1UZE, 2X91, 2X92, 2X93, 2X94, 2X95, 2X96, 2X97, 2XY9, 2XYD |
| | M3 Thimet Oligopeptidase | Peptydil-Dipeptidase | 1Y79 |
| | | Pz-peptidase A | 3AHN, 3AHO |
| | M1 Aminopeptidase N | Leukotriene A4 Hydrolase | 3FHE, 3FUK, ~~3B7R, 2R59~~ |
| | | Aminopeptidase N | ~~2ZXG, 4FYR~~ |
| | | M1 Alanylaminoppetidase | ~~3EBI~~ |
| | M27 Tentoxilysin | Botulinum neurotoxin A | ~~3BWI~~ |

**Table A2.** Scaffolds screened with the MA(A)$_{AS}$:diAla model.

| **Scaffold** | **Classification** (UniprotKB) | **Fold** | **Chain length (L)** | **PDB** | **#NMR X-ray** | **Secondary** | **Classical (UM)** |
|---|---|---|---|---|---|---|---|
| Trp-cage | De Novo Protein | α | 20 | 1L2Y | 38 | - - | - - |
| VPg | Viral Protein P03300 | α | 22 | 2BBL | 10 | OK | E7H14H2H17 |
| | | | | | | | E20H2H17H14 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| HIV gp120 C5 Domain | Viral Protein P19549 | α | 23 | 1MEQ | X-ray | OK | E9H22H5H3 |
| Poneratoxin | Toxin P41736 | α | 25 | 1G92 | 10 | OK | E21H6H15H17 |
| | | | | | | | E23H14H16H19 |
| FBP28 WW Domain | SH3 Domain Q8CGF7 | β | 27 | 1E0L | 10 | OK | E9H26H21H23 |
| Fibritin C-terminal - Foldon | Viral Protein P10104 | β | 27 | 4NCU | X-ray | - - | - - |
| CCK2E3 | Hormone/Growth Factor Receptor P32239 | α | 30 | 1L4T | 1 | OK | - - |
| pGolemi | De Novo Protein | α | 30 | 2K76 | 10 | OK | E13H6H7H16 |
| | | | | | | | E26H6H23H2 |
| Parathyroid Hormone | Hormone/Growth Factor P01270 | α | 31 | 1FVY | 20 | OK | - - |
| Cholecystokinin A | Hormone/Growth Factor P32238 | α | 31 | 1HZN | X-ray | - - | - - |
| DNA polymerase iota | Ubiquitin binding domain Q9UNA4 | α | 32 | 2L0G | 20 | OK | E685H700H679H696 |
| | | | | | | | E688H680H699H679 |
| | | | | | | | E688H680H703H679 |
| | | | | | | | E688H681H699H683 |
| | | | | | | | E688H703H679H682 |
| | | | | | | | E691H680H679H683 |
| | | | | | | | E700H680H699H703 |
| | | | | | | | E702H682H680H699 |
| | | | | | | | E706H680H703H699 |
| | | | | | | | E706H682H680H699 |
| | | | | | | | E706H682H703H699 |
| P-element Somatic Inhibitor | Nuclear Protein Q7JPS0 | α | 33 | 2BN6 | 29 | OK | - - |
| Human Villin Headpiece | Actin Binding P09327 | α | 34 | 1UNC | 25 | OK | E14H29H6H25 |
| | | | | | | | E14H29H17H25 |
| | | | | | | | E18H26H25H22 |
| | | | | | | | E26H6H25H29 |
| | | | | | | | E26H17H25H29 |
| | | | | | | | E28H6H10H17 |
| | | | | | | | E29H14H17H6 |
| Advillin Human | Actin Binding O75366 | α | 36 | 1UND | 25 | - - | - - |
| pPYY | Hormone/Growth Factor P68005 | α | 36 | 1RU5 | 20 | OK | E3H24H5H20 |
| | | | | | | | E7H24H5H20 |
| | | | | | | | E19H11H6H16 |
| | | | | | | | E27H36H34H24 |
| | | | | | | | E30H6H27H23 |
| | | | | | | | E36H20H24H5 |
| WW Domain Prototype | De Novo Protein | β | 37 | 1E0M | 10 | OK | E5H27H22H33 |
| | | | | | | | E5H27H25H33 |
| | | | | | | | E5H27H24H8 |
| | | | | | | | E5H27H24H22 |
| | | | | | | | E27H5H8H24 |
| E3-binding domain | Glycolysis P0AFG6 | α | 37 | 1BBL | 1 | OK | E17H39H16H20 |
| Alpha-T-alpha | De Novo Protein | α | 38 | 1ABZ | 23 | OK | E1H32H36H4 |
| | | | | | | | E1H33H36H4 |
| | | | | | | | E18H28H24H15 |
| | | | | | | | E28H21H25H15 |
| | | | | | | | E31H14H11H24 |
| | | | | | | | E31H14H11H21 |
| | | | | | | | E31H21H11H28 |
| | | | | | | | E31H21H28H24 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | E31H21H11H24 |
| | | | | | | | E31H14H11H28 |
| | | | | | | | E33H3H32H36 |
| | | | | | | | E33H4H32H36 |
| | | | | | | | E35H11H7H31 |
| | | | | | | | E35H11H32H28 |
| CRE-BP1 | DNA binding regulatory protein P15336 | α+β | 38 | 1BHI | 20 | OK | E5H15H8H10 |
| | | | | | | | E19H6H17H4 |
| | | | | | | | E28H11H9H31 |
| Phage Scaffolding Protein | Viral Protein P26748 | α | 40 | 2GP8 | 1 | OK | E13H28H16H32 |
| MafG | DNA binding protein O54790 | α | 41 | 1K1V | 20 | OK | E38H53H37H41 |
| | | | | | | | E38H53H37H49 |
| | | | | | | | E41H34H38H53 |
| Phosphoprotein XD domain | Viral protein P03422 | α | 44 | 2K9D | 20 | OK | E468H504H501H473 |
| | | | | | | | E468H504H501H474 |
| | | | | | | | E468H504H501H505 |
| | | | | | | | E468H505H474H501 |
| | | | | | | | E500H473H504H501 |
| | | | | | | | E502H463H498H467 |
| | | | | | | | E504H468H501H497 |
| GA module | Albumin Binding Protein Q51911 | α | 45 | 1PRB | 1 | OK | - - |
| HHR23A | C-terminal UBA domain P54725 | α | 45 | 1IFY | 10 | OK | E164H171H177H168 |
| FAF UBA domain | Apoptosis Q9UNN5 | α | 45 | 3E21 | X-ray | - - | - - |
| HIV Vpu | HIV protein P19554 | α | 45 | 1VPU | X-ray | - - | - - |
| UBA Domain | Ligase Q13191 | α | 46 | 2JNH | 15 | OK | E8H25H7H11 |
| | | | | | | | E8H21H25H15 |
| | | | | | | | E8H21H11H15 |
| | | | | | | | E29H5H7H33 |
| Sda antikinase | Signaling Protein Q7WY62 | α | 46 | 1PV0 | 25 | - - | - - |
| ATP synthase ipsilon chain | Hydrolase Q5KUJ4 | α | 46 | 2E5T | 20 | OK | E101H124H97H120 |
| | | | | | | | E104H97H101H120 |
| | | | | | | | E104H124H97H120 |
| | | | | | | | E104H124H101H120 |
| | | | | | | | H119H99H100H116 |
| | | | | | | | E127H101H124H97 |
| | | | | | | | E127H101H120H97 |
| dihydrolipoyllysine acetyl transferase | Transferase P11961 | α | 47 | 1W4E | 20 | OK | - - |
| Swa2p | Protein Binding Q06677 | α | 47 | 1PGY | 20 | OK | E4H29H7H25 |
| | | | | | | | E19H47H27H37 |
| | | | | | | | E19H47H27H41 |
| | | | | | | | E26H47H30H27 |
| | | | | | | | E27H41H37H19 |
| | | | | | | | E27H17H37H19 |
| | | | | | | | E27H17H19H41 |
| | | | | | | | E27H45H41H19 |
| | | | | | | | E33H7H29H10 |
| | | | | | | | E44H27H41H31 |
| | | | | | | | E44H27H41H37 |
| | | | | | | | E46H37H27H41 |
| LysM Domain | Hydrolase | α+β | 48 | 1E0G | 20 | OK | E7H28H34H24 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | P0AEZ7 | | | | | | E7H28H24H13 |
| | | | | | | | E48H30H26H46 |
| Lambda-Inte-grase | Viral Protein P03700 | α+β | 48 | 1KJK | 25 | OK | E15H52H55H14 |
| | | | | | | | E27H19H17H34 |
| | | | | | | | E29H52H55H14 |
| | | | | | | | E31H35H33H28 |
| | | | | | | | E31H58H28H30 |
| | | | | | | | E49H14H13H52 |
| | | | | | | | E51H31H30H28 |
| | | | | | | | E51H31H28H35 |
| | | | | | | | E51H33H28H35 |
| | | | | | | | E54H33H30H28 |
| Translation Initi-atior factor 2 N-ter | Translation P0A705 | α | 49 | 1ND9 | 10 | OK | - - |
| CSTF-64 | Nuclear Protein P33240 | α | 49 | 2J8P | 30 | OK | E530H565H534H561 |
| | | | | | | | E535H565H534H561 |
| | | | | | | | E537H530H561H534 |
| | | | | | | | E544H570H541H567 |
| | | | | | | | E545H555H542H563 |
| | | | | | | | E546H538H551H545 |
| | | | | | | | E546H538H552H545 |
| | | | | | | | E546H563H551H545 |
| | | | | | | | E546H563H552H545 |
| | | | | | | | E551H538H555H552 |
| | | | | | | | E551H563H555H552 |
| | | | | | | | E560H532H536H557 |
| POB | Transferase Q8ZUR6 | α | 51 | 1W4J | 20 | OK | E167H138H164H142 |
| NTL9 | RNA Binding Protein P02417 | α+β | 51 | 2HBB | X-ray | OK | - - |
| Engrailed Homeodomain | DNA Binding Protein P02836 | α | 54 | 1ENH | X-ray | OK | E34H15H38H19 |
| Protein G | Protein Binding P06654 | α+β | 56 | 2LGI | 10 | OK | - - |
| Protein A (Sp. aureus) | Protein Binding P38507 | α | 58 | 4NPE | X-ray | OK | - - |
| Thermolysin C-Ter | Hydrolase P00800 | α | 62 | 1TRL | 8 | OK | E300H265H262H295 |
| | | | | | | | E302H265H262H295 |
| | | | | | | | E309H284H288H267 |
| Protein L | Immune System Q51912 | α+β | 64 | 2JZP | 20 | OK | E6H51H55H58 |
| | | | | | | | E41H34H38H47 |
| | | | | | | | E44H18H12H10 |
| | | | | | | | E52H9H57H59 |

# Annex 3

The following derivation of models is based on references [143,173] and adapted suitably for the case of RD peptides and Zn(II).

*Zi binding model:* Under the tested experimental conditions, the apparent binding constant of Zi towards Zn(II) is given by $K_{ZnZi,app} = \frac{[ZiZn]}{[Zn][Zi]}$. Considering the mass balance: $[Zi]_T = [Zi] + [ZiZn]$ and $[Zn]_T = [Zn] + [ZiZn]$, substitution of the terms yields:

$$K_{ZnZi,app} = \frac{[ZiZn]}{([Zn]_T - [ZiZn])([Zi]_T - [ZiZn])}$$

By rearranging and solving the quadratic expression for [ZiZn] and considering only the real solution, the value of [ZiZn] can be related to $K_{ZnZi,app}$, $[Zn]_T$ and $[Zi]_T$ through:

$$[ZiZn] = \frac{\left(\frac{1}{K_{ZnZi,app}} + [Zn]_T + [Zi]_T\right) - \sqrt{\left(\frac{1}{K_{ZnZi,app}} + [Zn]_T + [Zi]_T\right)^2 - 4[Zi]_T[Zn]_T}}{2}$$

The increase in $A_{620}$ is directly related to amount of Zi-Zn(II) complex by the Beer-Lambert equation:

$$A = \varepsilon l c$$

Where *l* is the cell path-length (cm), ε the molecular extinction coefficient ($M^{-1}.cm^{-1}$) and *c* is [ZiZn] (M). Therefore, by combining the two previous equations yields:

$$A_{620} = \varepsilon l[ZiZn] = \varepsilon l \frac{\left(\frac{1}{K_{ZnZi,app}} + [Zn]_T + [Zi]_T\right) - \sqrt{\left(\frac{1}{K_{ZnZi,app}} + [Zn]_T + [Zi]_T\right)^2 - 4[Zi]_T[Zn]_T}}{2}$$

For each point in the titration, $[Zn]_T$ is increased while $[Zi]_T$ is fixed. Fitting of the data to the last equation can be used to obtain the value of ε, $K_{ZnZi,app}$ and the respective dissociation constant of Zi, $K_{dZnZi,app}=1/K_{ZnZi,app}$.

*Competition model:* The apparent binding constants of peptide (P) and Zi are $K_{ZnP,app} \equiv K_P = \frac{[ZnP]}{[P][Zn]}$ and $K_{ZnZi,app} \equiv K_{Zi} = \frac{[ZiZn]}{[Zi][Zn]}$, respectively. Thus, [ZiZn] and [ZnP] are related through:

$$\frac{[ZnP]}{[P]K_P} = \frac{[ZiZn]}{[Zi]K_{Zi}} \leftrightarrow \frac{K_{Zi}}{K_P} = \frac{[ZiZn][P]}{[Zi][ZnP]}$$

Considering the mass balance: $[ZnP] = [Zn]_T - [ZiZn]$, $[P] = [P]_T - [ZnP] \leftrightarrow [P] = [P]_T - ([Zn]_T - [ZiZn])$ and $[Zi] = [Zi]_T - [ZiZn]$, substitution of terms in the previous expression yields the following quadratic equation:

$$\begin{aligned}
\frac{K_{Zi}}{K_P} = \frac{[ZiZn][P]}{[Zi][ZnP]} = \frac{[ZiZn]\left([P]_T - ([Zn]_T - [ZiZn])\right)}{([Zn]_T - [ZiZn])([Zi]_T - [ZiZn])} &\leftrightarrow \frac{K_{Zi}}{K_P} \\
&= \frac{[ZiZn][P]_T - [ZiZn][Zn]_T + [ZiZn]^2}{[Zn]_T[Zi]_T - [Zn]_T[ZiZn] - [ZiZn][Zi]_T - [ZiZn]^2} \\
&\leftrightarrow ([Zn]_T[Zi]_T - [Zn]_T[ZiZn] - [ZiZn][Zi]_T - [ZiZn]^2)\frac{K_{Zi}}{K_P} \\
&= [ZiZn][P]_T - [ZiZn][Zn]_T + [ZiZn]^2 \\
&\leftrightarrow \frac{K_{Zi}}{K_P}[Zn]_T[Zi]_T - \frac{K_{Zi}}{K_P}[Zn]_T[ZiZn] - \frac{K_{Zi}}{K_P}[ZiZn][Zi]_T - \frac{K_{Zi}}{K_P}[ZiZn]^2 \\
&= [ZiZn][P]_T - [ZiZn][Zn]_T + [ZiZn]^2 \\
&\leftrightarrow [ZiZn]^2\left(\frac{K_{Zi}}{K_P} - 1\right) + [ZiZn]\left(-[Zn]_T\frac{K_{Zi}}{K_P} - [Zi]_T\frac{K_{Zi}}{K_P} - [P]_T + [Zn]_T\right) \\
&+ [Zn]_T[Zi]_T\frac{K_{Zi}}{K_P} = 0
\end{aligned}$$

Solving for [ZiZn] and considering only the real solution:

$$[ZiZn] = \frac{[Zn]_T\frac{K_{Zi}}{K_P} + [Zi]_T\frac{K_{Zi}}{K_P} + [P]_T - [Zn]_T - \sqrt{\left(-[Zn]_T\frac{K_{Zi}}{K_P} - [Zi]_T\frac{K_{Zi}}{K_P} - [P]_T + [Zn]_T\right)^2 - 4\left(\frac{K_{Zi}}{K_P} - 1\right)[Zn]_T[Zi]_T\frac{K_{Zi}}{K_P}}}{2\left(\frac{K_{Zi}}{K_P} - 1\right)}$$

Which combined with the Beer-Lambert equation yields the following expression:

$$A_{620} = \varepsilon l \times \frac{[Zn]_T\frac{K_{Zi}}{K_P} + [Zi]_T\frac{K_{Zi}}{K_P} + [P]_T - [Zn]_T - \sqrt{\left(-[Zn]_T\frac{K_{Zi}}{K_P} - [Zi]_T\frac{K_{Zi}}{K_P} - [P]_T + [Zn]_T\right)^2 - 4\left(\frac{K_{Zi}}{K_P} - 1\right)[Zn]_T[Zi]_T\frac{K_{Zi}}{K_P}}}{2\left(\frac{K_{Zi}}{K_P} - 1\right)}$$

For each point in the titration, $[Zi]_T$ is increased and $[P]_T$, $[Zn]_T$ are fixed. In reverse competition titrations, $[Zi]_T$ and $[Zn]_T$ are kept while $[P]_T$ is increased. Using the determined values of $K_{Zi}$ and

ε, the $K_{ZnP,app}$ of a given peptide can thus be determined by fitting of the data to the previous equation.

*Peptide binding model:* Considering the mass balance: $[P]_T = [P] + [ZnP]$ and $[Zn]_T = [Zn] + [ZnP]$; the apparent binding constant $K_{ZnP,app} = \frac{[ZnP]}{[Zn][P]}$ of peptides can be rearranged as:

$$K_{ZnP,app} = \frac{[ZnP]}{([Zn]_T - [ZnP])([P]_T - [ZnP])} \leftrightarrow [ZnP] = K_{ZnP,app}([Zn]_T - [ZnP])([P]_T - [ZnP])$$

$$\leftrightarrow \frac{[ZnP]}{K_{ZnP,app}} = ([Zn]_T - [ZnP])([P]_T - [ZnP]) \leftrightarrow \frac{[ZnP]}{K_{ZnP,app}}$$

$$= [Zn]_T[P]_T - [Zn]_T[ZnP] - [ZnP][P]_T + [ZnP]^2$$

$$\leftrightarrow [Zn]_T[P]_T - [Zn]_T[ZnP] - [ZnP][P]_T + [ZnP]^2 - \frac{[ZnP]}{K_{ZnP,app}} = 0$$

$$\leftrightarrow [ZnP]^2 - [ZnP]\left(\frac{1}{K_{ZnP,app}} + [Zn]_T + [P]_T\right) + [P]_T[Zn]_T = 0$$

Solving the quadratic expression for the real solution yields:

$$[ZnP] = \frac{\left(\frac{1}{K_{ZnP,app}} + [Zn]_T + [P]_T\right) - \sqrt{\left(\frac{1}{K_{ZnP,app}} + [Zn]_T + [P]_T\right)^2 - 4[P]_T[Zn]_T}}{2}$$

For each point in the titration, $[Zn]_T$ is increased while $[P]_T$ is kept constant. The corresponding fraction of peptide-zinc complex ($f_{ZnP}$) formed is both dependent on $[Zn]_T$ and available peptide $[P]_T$, $f_{ZnP} = \frac{[ZnP]}{[ZnP]_T} = \frac{[ZnP]}{[P]_T}$, where $[ZnP]_T=[P]_T$ in saturated regime since $[P]_T$ is the limiting reactant. Conformational changes occurring upon Zn(II) additions are monitored by changes in ellipticity signal $[\theta]_{obs}$, therefore at each point of the titration the fraction of folding is thus $f_F = \frac{[\theta]_{obs} - [\theta]_U}{[\theta]_F - [\theta]_U}$, where $[\theta]_U$ and $[\theta]_F$ are the ellipticity values in the beginning and end-point of the titration (considering that the *apo* form is unfolded and the *holo* form is fully formed at the end). The increase in α-helical content in both types of peptides is used as a measure of folded content, $[\theta]_{obs}=[\theta]_{222}$. Assuming that all peptide-Zn(II) formed is in a folded state, *i.e.* $f_F \approx f_{ZnP}$, then:

$$f_F \approx f_{ZnP} = \frac{[ZnP]}{[P]_T} = \frac{\left(\frac{1}{K_{ZnP,app}} + [Zn]_T + [P]_T\right) - \sqrt{\left(\frac{1}{K_{ZnP,app}} + [Zn]_T + [P]_T\right)^2 - 4[P]_T[Zn]_T}}{2[P]_T}$$

From the Zn(II) titration is possible to obtain the respective $K_{ZnP,app}$ values for each of the designed peptides by tracking ellipticity changes at 222 nm.

*Two-state model of peptide thermal denaturation:* At a given temperature (Kelvin) the constant of folding $K = \frac{[F]}{[U]}$ is related to the fraction of folded peptide $f_F = \frac{[\theta]_{obs} - [\theta]_U}{[\theta]_F - [\theta]_U}$ through:

$$K = \frac{f_F}{(1 - f_F)} \leftrightarrow \frac{K}{(1 + K)} = f_F$$

Where, $[\theta]_{obs}$ is the observed ellipticity at 222 nm, $[\theta]_F$ and $[\theta]_U$ the corresponding ellipticity values when the peptide is fully folded or unfolded, respectively. The free energy of folding ΔG at a given temperature is:

$$\Delta G = -RT \ln K$$

Where R is the gas constant, 1.98 cal.mol$^{-1}$. Therefore, at the temperature $T_m$ where $f_F$=0.5 the system has a ΔG=0. The corresponding enthalpy of folding $\Delta H_{Tm}$ is given by $\Delta H(T) = \Delta H_{Tm} + \int_{Tm}^{T} \Delta C_p dT$, where $\Delta C_P$ is the heat capacity change of a peptide upon folding (usually considered negligible for peptides). Considering that $\Delta S(T) = \frac{\Delta H_{Tm}}{Tm} + \int_{Tm}^{T} \Delta C_p d \ln T$, though the Gibbs-Helmholtz equation:

$$\Delta G(T) = \Delta H - T\Delta S \leftrightarrow \Delta G(T) = \Delta H_{Tm} + \int_{Tm}^{T} \Delta C_p dT - T\left(\frac{\Delta H_{Tm}}{Tm} + \int_{Tm}^{T} \Delta C_p d \ln T\right) \leftrightarrow \Delta G(T)$$

$$= \Delta H_{Tm}\left(\frac{Tm - T}{Tm}\right) + \int_{Tm}^{T} \Delta C_p dT - T\int_{Tm}^{T} \Delta C_p d \ln T \leftrightarrow \Delta G(T)$$

$$= \Delta H_{Tm}\left(1 - \frac{T}{Tm}\right) + \Delta C_p\left[(Tm - T) + T \ln \frac{T}{Tm}\right]$$

The values of $T_m$ and $\Delta H_{Tm}$ from a given peptide are related with ellipticity changes based on by considering the following relations:

$$K = e^{\left(-\frac{\Delta G}{RT}\right)} \leftrightarrow K = e^{\left(-\frac{\Delta H_{Tm}\left(1 - \frac{T}{Tm}\right) + \Delta C_p\left[(Tm - T) + T \ln \frac{T}{Tm}\right]}{RT}\right)}$$

$$f_F = \frac{[\theta]_{obs} - [\theta]_U}{[\theta]_F - [\theta]_U} = \frac{K}{1 + K}$$

$$[\theta]_{obs} = f_F([\theta]_F - [\theta]_U) + [\theta]_U$$

And with corrections of the data for pre- ($F_M$) and post- ($U_M$) transition linear changes in $[\theta]_{obs}$ as a function of temperature:

$$[\theta]_{obs} = f_F[([\theta]_F + F_M T) - ([\theta]_U + U_M T)] + ([\theta]_U + U_M T)$$

Therefore, by monitoring ellipticity changes of the thermal unfolding curve, the values of $T_m$, $\Delta H_{Tm}$ and ΔG can be obtained.